Information Extraction Efficiency of Business Documents Captured with Smartphones and Tablets

Daniel Esser, Klemens Muthmann, Daniel Schuster
Computer Networks Group
Dresden University of Technology
01062 Dresden, Germany
{daniel.esser,klemens.muthmann,daniel.schuster}@tu-dresden.de

ABSTRACT

Businesses and large organizations currently prefer scanners to incorporate paper documents into their electronic document archives. While cameras integrated into mobile devices such as smartphones and tablets are commonly available, it is still unclear how using a mobile device for document capture influences document content recognition. This is especially important for information extraction carried out on documents captured in a mobile scenario. Therefore this paper presents a set of experiments to compare automatic index data extraction from business documents in a static and in a mobile case. The paper shows which decline in extraction one can expect, explains the reasons and gives a short overview over possible solutions.

Categories and Subject Descriptors

I.7.5 [**DOCUMENT AND TEXT PROCESSING**]: Document Capture—*Document analysis, Optical character recognition (OCR)*

General Terms

EXPERIMENTATION

Keywords

Mobile tagging; Information extraction; Evaluation

1. INTRODUCTION

Capturing and processing paper documents is still of high importance for all kinds of business as well as non-profit organizations. Organizations need a fast and accessible way to add such documents to their electronic document archives and annotate them with metadata for easy retrieval if required. The metadata is usually provided in the form of index entries, simple key value pairs identifying the document's content and role within the company. An invoice document for example might have entries such as sender, receiver, amount or invoice items.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
DocEng'13, September 10–13, 2013, Florence, Italy.
Copyright 2013 ACM 978-1-4503-1770-2/13/09 ...$15.00.
http://dx.doi.org/10.1145/2494266.2494302.

High accessibility is only possible if the workflow from paper to the organization's electronic archives is as short as possible. In many cases this consists of a clerk using a scanner and a desktop application. Today however most people own smartphones and tablets that are equipped with high-quality cameras. Using these devices for document capture is quite appealing as they enable to capture documents anytime, anywhere and to immediately upload them to the document archive.

There already exist systems for indexing and archiving scanned documents like SmartFix [5] or the OpenText Capture Center [7]. Other systems provide mobile document capture capabilities such as the work by Patel [6] or Receipts2Go [4]. However there is yet no efficiency evaluation available on what it means to introduce mobile document images to an indexing system which used to work on scanned documents.

Especially considering the decrease of document capture quality, we want to know how much this influences the quality of extracted metadata index entries. Therefore this work presents an overview of the differences in quality we encountered using the same system for automatic document indexing in a mobile case compared to a static case. We present results for different mobile scenarios and discuss encountered errors as well as proposals for future solutions.

The remainder of this paper is organized as follows. At first we discuss different variables influencing mobile document capture. Then a short overview of the intelligent indexing approach is explained, together with some details on how mobile capturing might influence the results. Afterwards the main part of this paper gives an overview of our experimental setup and results. The results are discussed in detail in Section 4.3 after which the paper closes with conclusions and an overview of future work.

2. CAPTURE QUALITY

There is a multitude of variables which influences the quality of mobile document capturing. We used lighting, type of paper and tilt of the mobile device for the analysis in this paper.

Lighting is the amount of light a document receives while being captured. In the static case, a bright lamp integrated into a scanning device is used to achieve the best lighting possible. In the mobile case lighting depends on the time of day or the artificial light available. For our experiments we recorded documents within three common lighting scenarios, we expect to be near to real-world capturing. Table 1 shows an overview of those levels (with light measured in Lux).

Table 1: Lighting quality levels

Class	Lux	Scenario
L1	780–1097	Bright daylight
L2	116–430	Artificial direct lighting
L3	26–29	Artificial indirect lighting

The type of paper influences the contrast between the document's background and text. We expect darker recycling paper to cause more errors than high quality white paper.

Table 2: Paper quality levels

Class	Paper type	Scenario
P1	White	External documents
P2	Recycling	In-house documents

Lastly we also evaluated the influence of the tilt of the device. To measure bad device position we placed our capture device at an angle of 45° compared to the document plane and contrast those results to positioning the device flat and directly above the document.

Table 3: Tilt quality levels

Class	Position	Scenario
T1	Flat	Standing in front of document
T2	45°	Sitting in front of document

3. TEST SETUP

Our experiments are based on Intellix, a commercial quality extraction system which identifies similar looking documents out of the training set on a kNN basis. On this set of similar looking documents, three types of extraction steps are executed. First, it tries to identify fields with similar value across the set of similar looking documents which often applies for document type and sender. Second, it tries to identify index fields with values at nearly the same position (using bounding boxes of user-tagged values in training documents). Third, it uses context words in the surrounding of tagged fields in the training documents. Details about the extraction workflow can be found in [3, 8]. A similar workflow was first described by Cesarini et al. [1] and is likely to be found in other document processing systems as well.

The extraction workflow works on structured OCR output produced by a commercial OCR engine which we consider as a black box. We believe this is a common approach in document indexing as commercial OCR engines offer the best OCR quality available. We will discuss later on which part of the errors are due to a decrease in OCR accuracy and could possibly be diminished by tuning OCR parameters.

To access the Intellix indexing service a static and a mobile client were provided. The static client is a simple Java-based command line tool capable of sending a whole corpus of documents to the Intellix service and evaluating the results. The mobile client is an application written for Apple iOS. The client is able to take pictures from documents and detects the document's border using the well known Hough transform. In addition a homography matrix rectifies the image. The mobile client is also able to present the results and provide user corrections for the extracted values via the same Web service. Figure 1 shows our client at work.

Figure 1: Capturing a business document with our iOS application. The current view shows the Intellix edge detection algorithm.

Our initial document corpus consists of 12,500 documents, captured with a customary scanner and tagged and corrected according to commonly used fields in document archiving. Beside a minimal set of fields to enable structured archiving (*document type*, *recipient*, *sender*, *date*), we added further popular fields like *amount*, *document number*, and *subject* based on a survey carried out by our project partner DocuWare. Out of this initial set, we randomly generated a subset of 1,000 static documents (C_1) and captured them using the internal camera of an iPad 3 (5-megapixel with autofocus) in combination with our Intellix iOS application (C_2). Therefore we established perfect circumstances to ensure a high recording quality.

To evaluate the influence of different quality levels, we randomly selected another subset of 200 documents. These documents were captured by our iPad 3 multiple times using different lighting, paper, and orientation scenarios. Altogether we created five scenarios resulting in five corpora (C_A-C_E), whose configurations can be seen in Table 4.

Table 4: Configurations

Category	Lighting	Paper	Tilt
A	L1 (good)	P1 (white)	T1 (flat)
B	L2 (medium)	P1 (white)	T1 (flat)
C	L3 (bad)	P1 (white)	T1 (flat)
D	L1 (good)	P2 (recycling)	T1 (flat)
E	L1 (good)	P1 (white)	T2 (45°)

For evaluation, we used the common metrics precision, recall, and F1-measure adopted by Chinchor and Sundheim [2] to the domain of information extraction for MUC-5. As the system user expects only correct results, we ignored error class "partial" and tackled this kind of extractions as wrong. Overall values were calculated using a micro-averaging approach by averaging single results across all recognized labels. To ensure significant evaluation results, each test was repeated ten times with changing document order.

4. EXPERIMENTS AND RESULTS

The following experiments mainly try to answer two research questions. First, how do documents captured on a

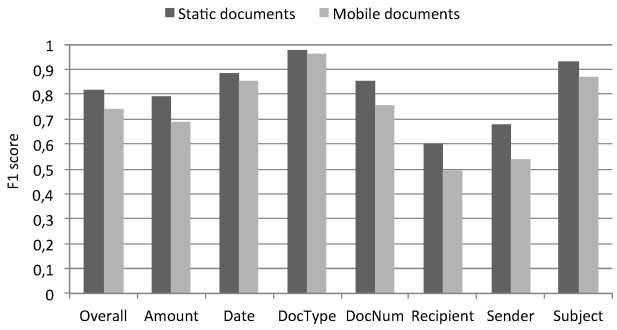

Figure 2: Comparison of overall and field-by-field results using static and mobile captured documents.

mobile device compare to documents captured by a scanner regarding the efficiency of the information extraction? Second, how big is the expected further decrease in information extraction efficiency, if the mobile capturing was done under bad environmental conditions as described in Section 2?

4.1 Capture Mode Comparison

To be able to compare the two capture modes, namely scanner vs. mobile device, we pre-trained our system with 4,000 static documents. Afterwards we evaluated the system twice using the same set of 1,000 documents captured by a scanner (C_1) as well as captured by the iPad client (C_2). Figure 2 shows the system behavior.

Altogether we reach a F1 score of 81% for documents captured by a scanner and 73% for the same documents captured by a mobile device. As expected, the results of mobile captured documents are below the ones of static documents. The differences range between 1 and 14 percentage points, depending on the type of index field. Especially the differences for the document type are typically small due to the fact that this index field is detected by a classification, which is less prone to OCR errors. Sender and recipient contain the largest spread (12 and 14 percentage points). The values of such index fields are typically much longer than others, which increases the probability of OCR errors.

4.2 Mobile Capture with Bad Quality

To prove our expectations according to the extraction of documents with different quality levels, we pre-trained our system using 1000 mobile documents (C_2) captured with perfect circumstances. For each quality category, we evaluated the performance on a fresh system using 200 test documents (C_A-C_E). Figure 3 shows the results of this test.

As expected, documents from category A perform best. Guaranteeing ideal lighting and orientation conditions and using white paper for documents results in extraction rates around 75% F1-measure. While the kind of paper, a document is printed on, does only marginally effect the rating (D - L1/P2/T1 - 72%), changes in lighting circumstances significantly reduce the overall results. Documents captured with medium lighting (B - L2/P1/T1) reach F1-measure around 70%, while documents recorded with bad lighting (C - L3/P1/T1) can be extracted with 67%. The location of the camera according to the document has the biggest influence on the extraction results. Capturing at an angle of 45° while sitting, even after optimizing the picture by our mobile application, results in a F1-measure 17 percentage points lower (E - L1/P1/T2 - 58%) than the results of our best category A.

If we look at the field-by-field results in Figure 3, we can see nearly constant values for index fields document type and sender. As already mentioned, for both fields we provide algorithms, which rather classify than extract these values. OCR errors do not effect the results as much as the results of other fields.

Within the results for category E (L1/P1/T2), we can see an interesting characteristic. While outcome for "amount" is quite comparable to the other categories, other index fields, especially the ones that occur in the top half of the documents perform much lower. We expect this behavior because of the limitation of mobile device cameras to use only one region as focal point. While the bottom of the captured document is sharp enough to produce good OCR results, the top of the document gets more and more diffuse.

4.3 Error Analysis

To get an overview of errors that occur by using mobile captured documents, we analyzed the results from our comparative test between static and mobile versions. Therefore we manually tagged each erroneous extraction that has occurred in a mobile document but was correctly identified in the appropriate static document. Altogether we could identify six error classes (ordered by occurrence):

1. *Wrong extraction (30%):* The extracted value is completely wrong. There is no similarity between extracted and valid value (i.e. "02/21/2013" vs. "invoice").
2. *OCR - Character (25%):* Extracted and valid value differ only by OCR character errors (i.e. "invoice" vs. "invuice").
3. *Partial extraction (19%):* The extracted and valid value overlap. Either too many or too few characters for valid value were extracted (i.e. "ACME Ltd." vs. "ACME").
4. *Missing extraction (15%):* No value was extracted, although a valid one exists (i.e. "02/21/2013" vs. "-").
5. *OCR - Tokenization (9%):* Extracted and valid value differ only by wrong detected word tokenization (i.e. "ACME Ltd." vs. "AC MELtd.").
6. *Spurious extraction (2%):* A value was extracted, although there is no value for this field (i.e. "-" vs. "02/21/2013").

Two-third of failures while processing mobile documents are partly based on our algorithms. While we use kNN search and layout-based extraction algorithms that mainly focus on the position of fields within the document, movements while capturing are hard to handle for them. Wrong, missing and spurious extractions (47%) tend to be constituted by the detection of similar looking documents.

Partial extractions, which are responsible for 19% of all errors, are based on movements while capturing. Although our iPad application tries to normalize recorded documents, our algorithms produce this kind of error due to small movements and deformations.

Only one-third of analyzed extraction errors are based on OCR. 9% of all failures lead to a wrong tokenization of words. Missing or extended spaces are typically seen in sender and recipient names.

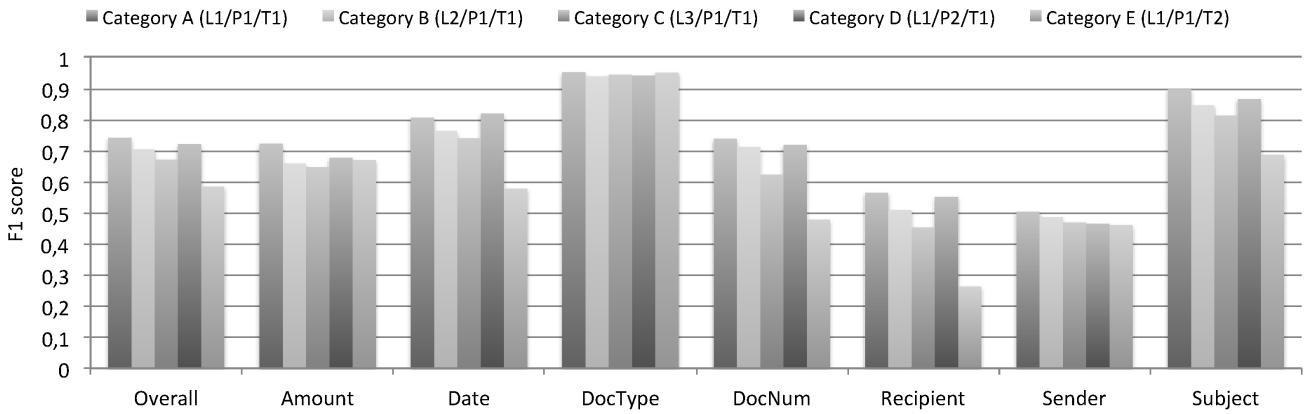

Figure 3: Comparison of overall and field-by-field extraction results for each quality category.

4.4 Discussion

The best way to overcome bad extraction results for mobile recorded documents is quite easy. Ensuring perfect circumstances according to lighting, paper and position of the mobile device guarantee high extraction rates. In practice perfect conditions are mostly rare. To minimize the trade-off we propose different solutions.

The limitations of mobile device cameras to focus only single regions leads to bad OCR results, when captured with an angle of 45°. Automatically capturing a document multiple times with different focuses and adding it afterwards to a completely sharp version can be a solution.

Altogether mobile applications have to get smarter. Right before hitting the shutter, the user has to be informed how the quality of the capture can be increased. Possible solutions are measurements of lighting level or identification of depth information to avoid partly blurred documents.

5. CONCLUSION

This paper presents a comparison of the efficiency of index data extraction on business documents captured either by a scanner or by a mobile device using the commercial quality extraction system Intellix. It shows that quality drops around 8% points F1 score, when capturing with mobile device cameras in an optimal way. The quality decrease will raise up to 20% and more if bad lighting or tilt compared to the paper plane occur during capturing. The problems for this loss in quality are an increase in OCR errors, movements, and distortions.

While the concrete numbers may vary from system to system, our error analysis shows the different types of errors and thus offers clues to tackle at least part of the quality decrease by improving OCR and extraction algorithms. We plan to work in this direction to improve the capturing experience for mobile users. Furthermore, improved user interfaces telling the user how to get the best extraction results on capturing would be of great help.

Nevertheless, the results described in this paper already show the feasibility of capturing business documents with mobile devices. With expected ongoing improvements in OCR quality, extraction algorithms and quality of camera hardware, we believe mobile capturing to be soon the preferred way for archiving paper documents.

6. ACKNOWLEDGMENTS

This research was funded by the German Federal Ministry of Education and Research (BMBF) within the research program "KMU Innovativ" (fund number 01/S12017). We thank our project partners from DocuWare for insightful discussions and providing us with the document corpus used for evaluation.

7. REFERENCES

[1] F. Cesarini, S. Marinai, and G. Soda. Retrieval by layout similarity of documents represented with mxy trees. In *Document Analysis Systems*, 2002.

[2] N. Chinchor and B. Sundheim. Muc-5 evaluation metrics. In *Proceedings of the 5th conference on Message understanding*, MUC5 '93, pages 69–78, 1993.

[3] D. Esser, D. Schuster, K. Muthmann, M. Berger, and A. Schill. Automatic indexing of scanned documents - a layout-based approach. In *Document Recognition and Retrieval XIX (DRR)*, San Francisco, CA, USA, 2012.

[4] B. Janssen, E. Saund, E. Bier, P. Wall, and M. A. Sprague. Receipts2go: the big world of small documents. In *Proceedings of the 2012 ACM symposium on Document engineering (DocEng '12)*, 2012.

[5] B. Klein, A. Dengel, and A. Fordan. smartfix: An adaptive system for document analysis and understanding. *Reading and Learning*, pages 166–186, 2004.

[6] X. Liu and D. Doermann. Mobile retriever: access to digital documents from their physical source. *International Journal of Document Analysis and Recognition (IJDAR)*, 11(1):19–27, 2008.

[7] Opentext. Opentext capture center. http://www.opentext.com/2/global/products/products-capture-and-imaging/products-opentext-capture-center.htm, 2012.

[8] D. Schuster, K. Muthmann, D. Esser, A. Schill, M. Berger, C. Weidling, K. Aliyev, and A. Hofmeier. Intellix - end-user trained information extraction for document archiving. In *Document Analysis and Recognition (ICDAR)*, Washington, DC, USA, 2013. (accepted for publication).