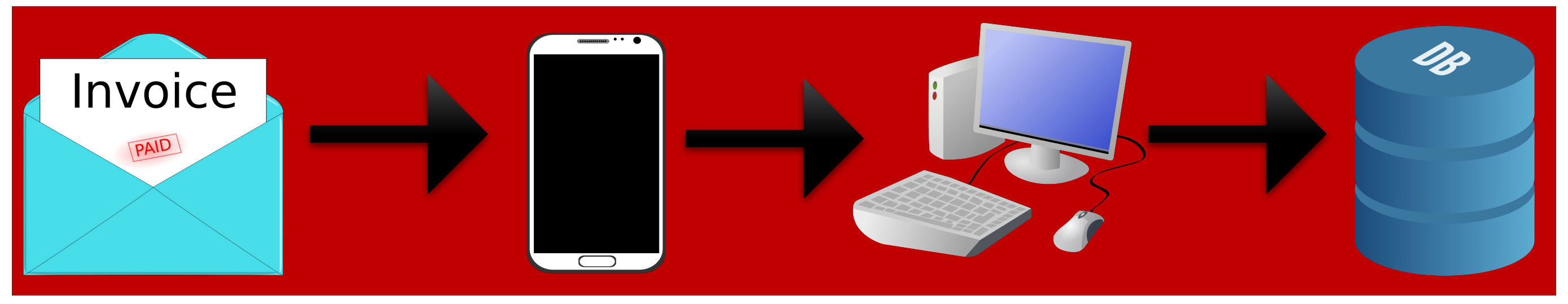
UPPSALA UNIVERSITET

ADC Invoice -Automatic reading and interpretation of paper invoices

Introduction

Today, most invoices are distributed in paper format. However many companies use bookkeeping software and online banking services. It is tedious work to manually type in all the numbers and other reduced information from an invoice.
That is why we have created ADC Invoice, an application that reads information from an image of an invoice.

After that, our idea was to make importation to bookkeeping software possible. In the beginning we thought about making a mobile application that takes a photo and sends it to a server for reading.



How it works

The process of the image in ADC Invoice can be split into four steps, where the image processing API, OpenCV is used:

Enhancing - The image scales down and a binary mask is applied, making the foreground objects white and the background becomes black. Erode and dilate filters are used to separate merged characters or combine those who are broken apart. The image will be scaled up again.

Tagging - Each text-block in the image is captured within a rectangular shape.

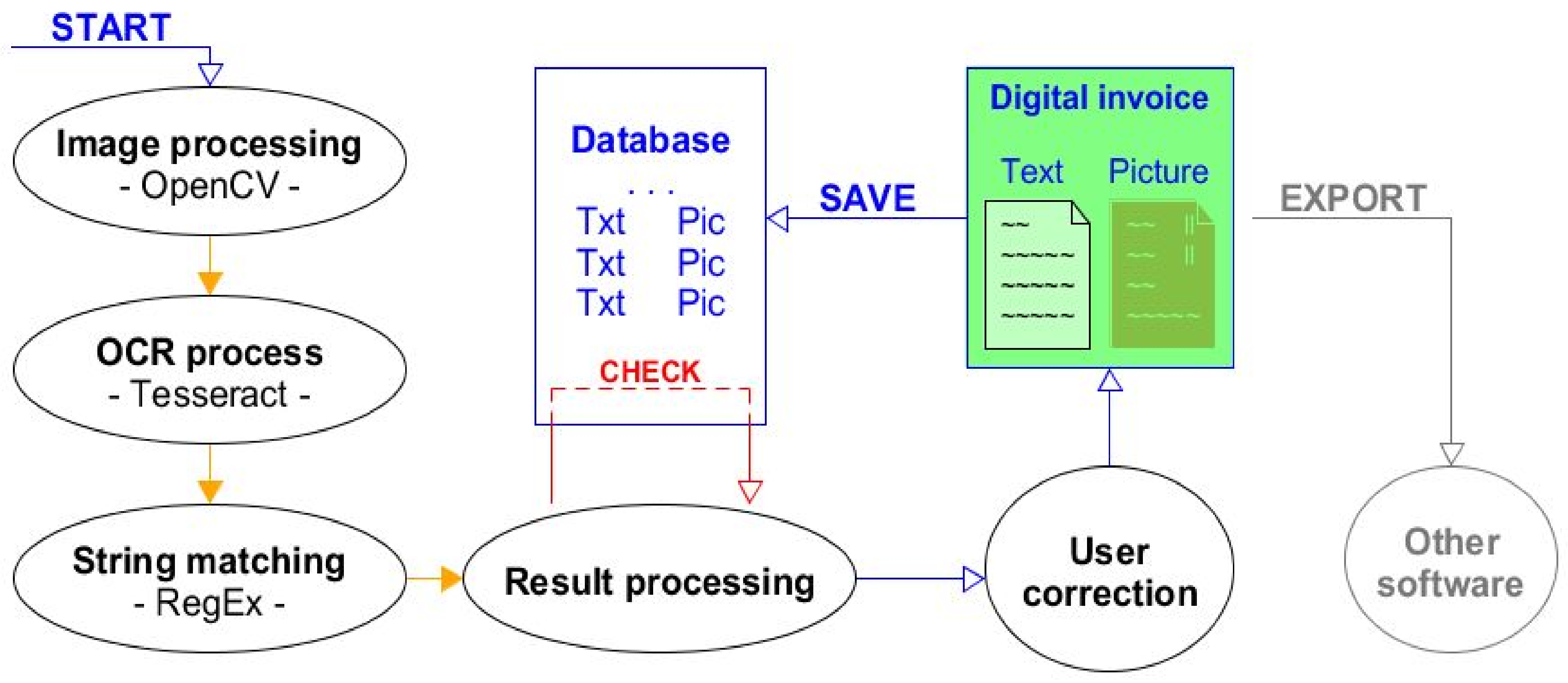

START

Image processing
- OpenCV -

OCR process
- Tesseract -

String matching
- RegEx -

Result processing

Database
. . .
Txt Pic
Txt Pic
Txt Pic
CHECK

SAVE

Digital invoice
Text Picture

EXPORT

User correction

Other software

Reading - The Tesseract OCR (*optical character recognition*) API is used to read the text within a rectangle applied in the tagging step. The text will be returned as strings. Most of the ASCII-characters are supported.

Matching - The strings returned by the OCR are interpreted by a regular expressions algorithm utilizing patterns to identify the context of the strings. Every pattern is checked in a predefined order for every block of text. The order is longest pattern first, to prevent possible matches where a pattern is also a sub-pattern to a longer one, causing mismatches. As an example, IBAN numbers typically have twenty-four characters out of which the first two are letters and the last twenty-two are digits. Whereas VAT numbers only have 14 tailing digits.

Result

Time - The application reads an invoice in about 35 seconds, however the invoice needs to be scanned, possibly converted and found by the application before the reading can take place. Ideally the user will scan multiple invoices at a time, reducing the time spent walking back and forth between the scanner and computer. As this time can differ greatly between users the time to scan an invoice has not been measured.

Acuraccy - The application successfully extracts **[Nytt test]**% of all targeted fields of information. Given the unpredictable nature of OCR the user is an integral part of the applications accuracy. The Results are also affected by repetition, if invoices from a given service provider have been received previously the application will likely be more accurate as it saves information about previous services providers.

Future work

Layout - To support invoice layouts that use different formats and have information structured in other ways than those tested, the system would need a better model of a generic invoice. It is hard to create an OCR process that can make any guarantee of the results. To increase the chance of an acceptable result, a more thorough study of common invoice layouts could be performed, to attempt to create one or more accurate template models for both content and formats.

Batch-processing - The purpose of the ADC invoice software is to make the invoice digitization more efficient. To this end, a work queue would be able to focus the human interaction to more practical intervals. Considering the OCR process can be performed within a minute, a user could leave the system to perform the workload in the queue and later return to confirm the accumulated results.

Company Dictionary – An invoice is saved (after necessary correction), the financial - and address information for the service provide is saved to a file. Next time an invoice from the same service provider is processed, missing and failed labels (after OCR and matching) will be autocompleted if some of the fields matches the saved information.

Authors:
Boström, Carl
Herelius, Johan
Hugosson, Mathias
Maleev, Sergej