Optical Character Recognition and How it Works

Henrique Santos Franco

October 29, 2014

Presentation Resources

- Presentation and extra material is available at: http://goo.gl/yu6bNR?gdriveurl

- It contains a text with more information and an example using MATLAB

Table of Contents

- Optical Character Recognition, or **OCR**, is a technology capable of converting different kinds of documents into editable and searchable data

- Such documents can be scanned documents, PDF files or images captured by a digital camera

- OCR tools recognize texts in **images** and convert them to a format that computers are able to read, edit or store as **text**

- OCR researches can be traced back to the 1950s
  - Big scanners were used
  - Difficult and inefficient processing tools
  - Very expensive service

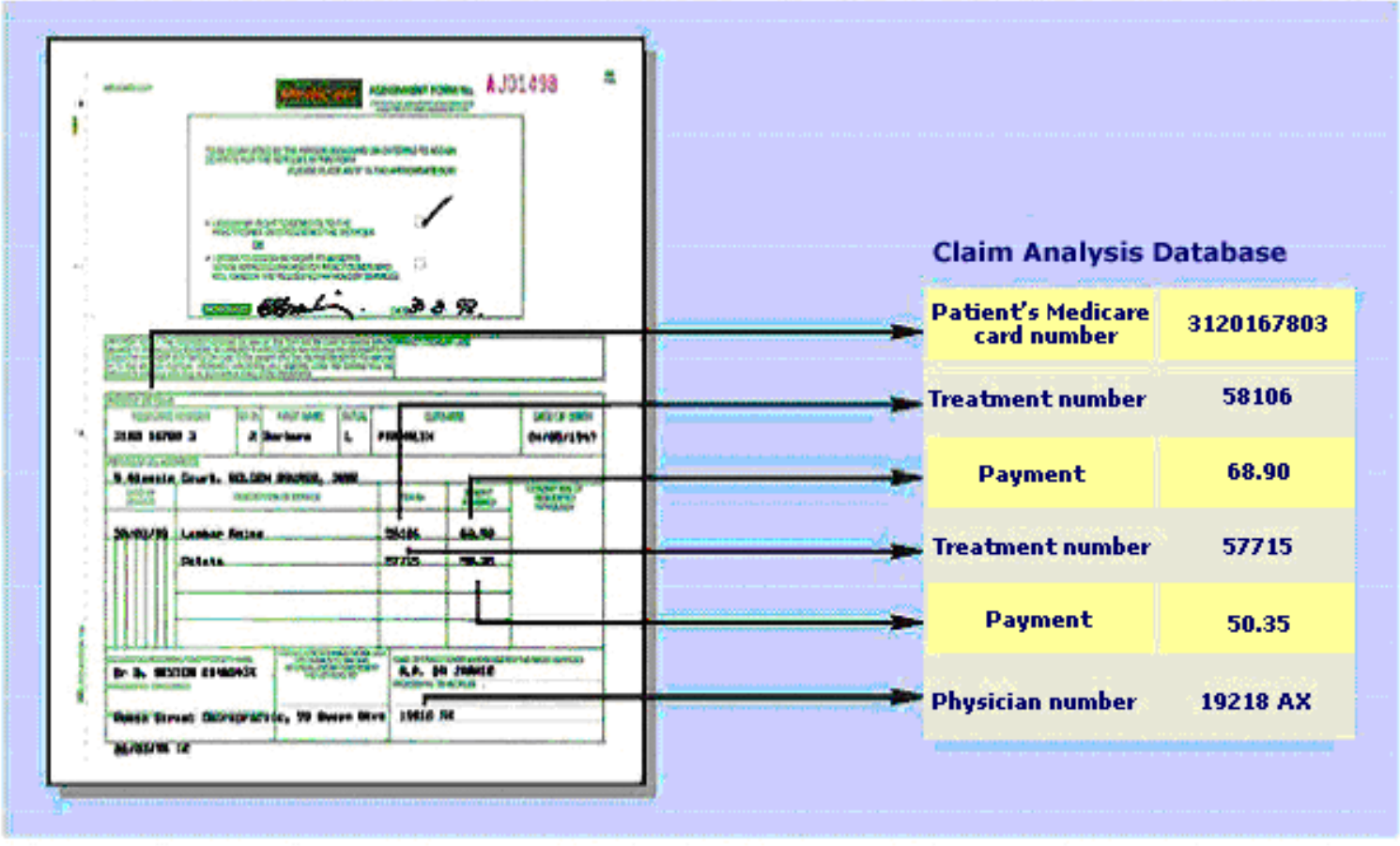
- By the 1960s and 1970s
  - Important hardware advances
  - OCR applications sprang to companies, newspaper publishers, banks and many others
  - Still an error-prone process

- OCR manufacturers pushed a standardization of printing fonts, paper and ink qualities
  - Creation of OCRA and OCRB fonts
  - Better quality of printing among companies

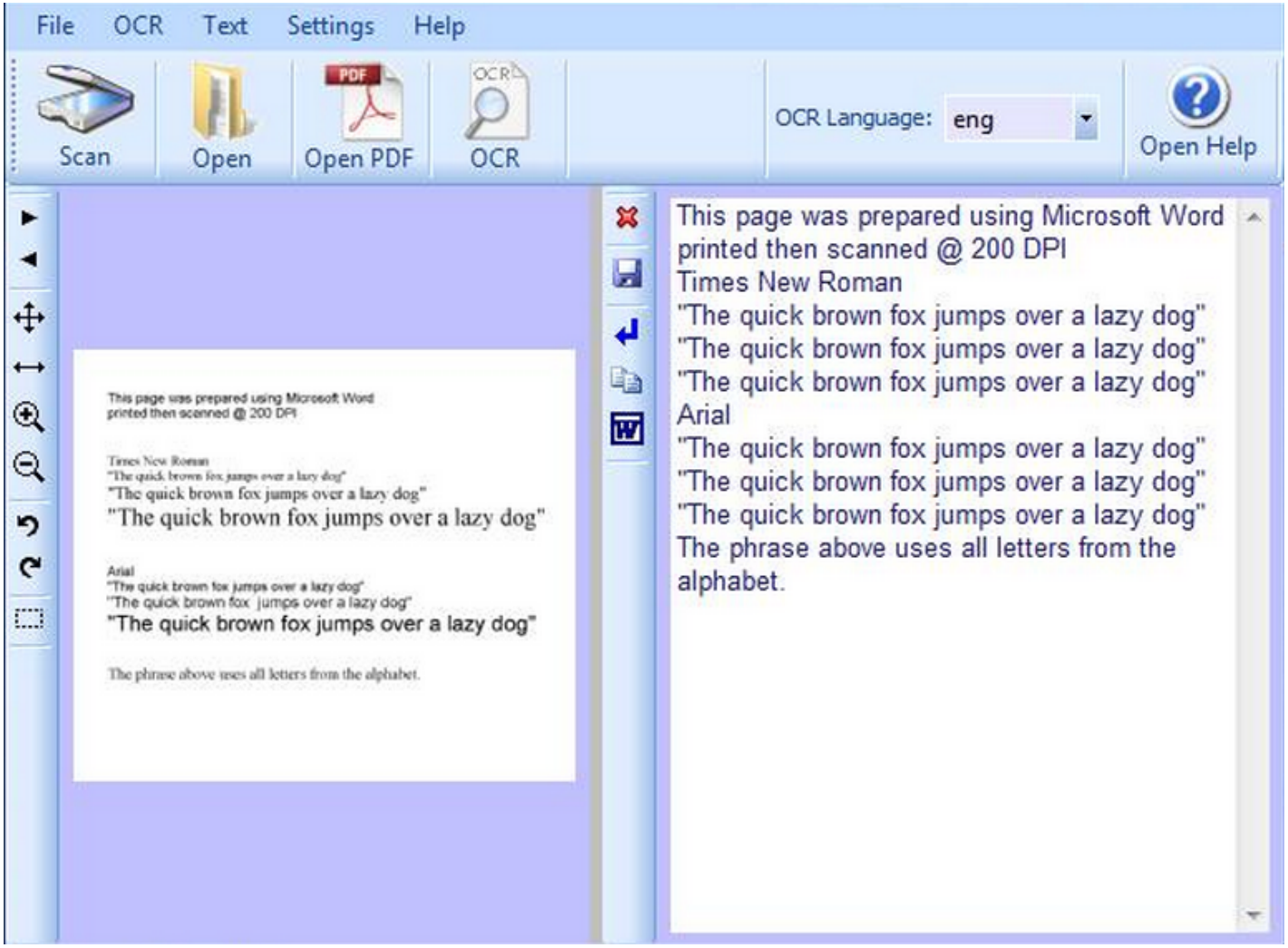
- Demand on handwriting recognition increased
  - Still difficult for the current techniques (1980s)
  - Hand print models and boxes

- After years of development, recognition techniques advanced rapidly
  - More powerful computers
  - Free handwriting (no longer necessary to use boxes)

- Specific-task readers can perform a specific action. In the image you can see that this OCR tool converts only the text used to fill the form

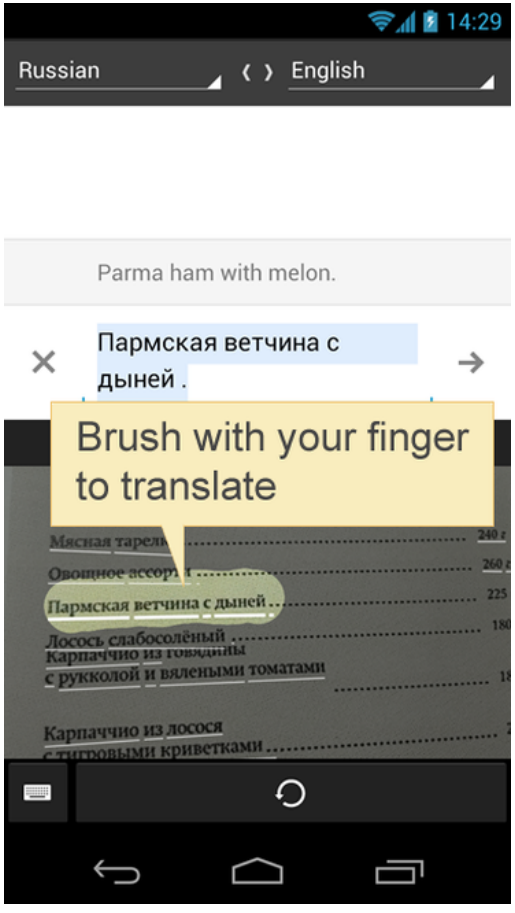


- General purpose readers can simply transform any text in an image into editable data

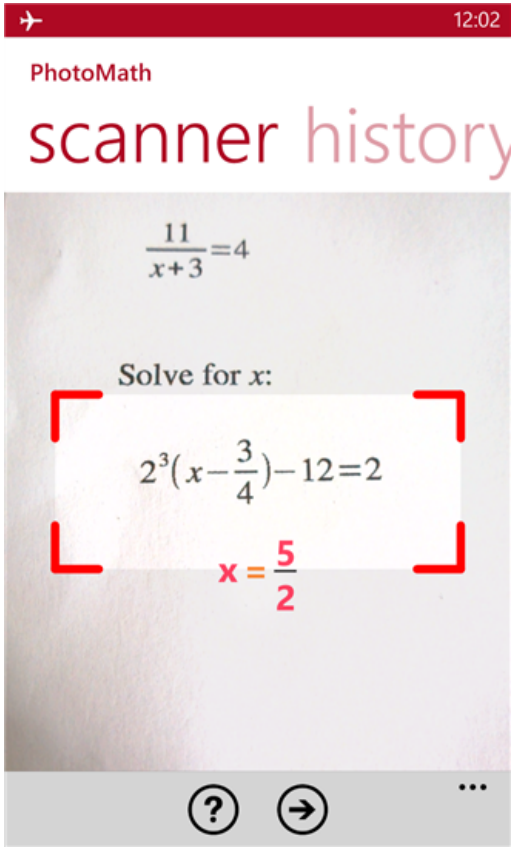


- Nowadays OCR tools have many different purposes

- The application is becoming more and more related to what we do to the read data than simply storing it
  - As an example, we could cite the Google Translate smartphone app or the new PhotoMath app



A simple example of Google Translate being used to read a menu and translate the content to English



PhotoMath is capable of reading equations and solving them for the user

Pre-Processing

The pre-processing stage enhances the quality of the input image and locate the data of interest

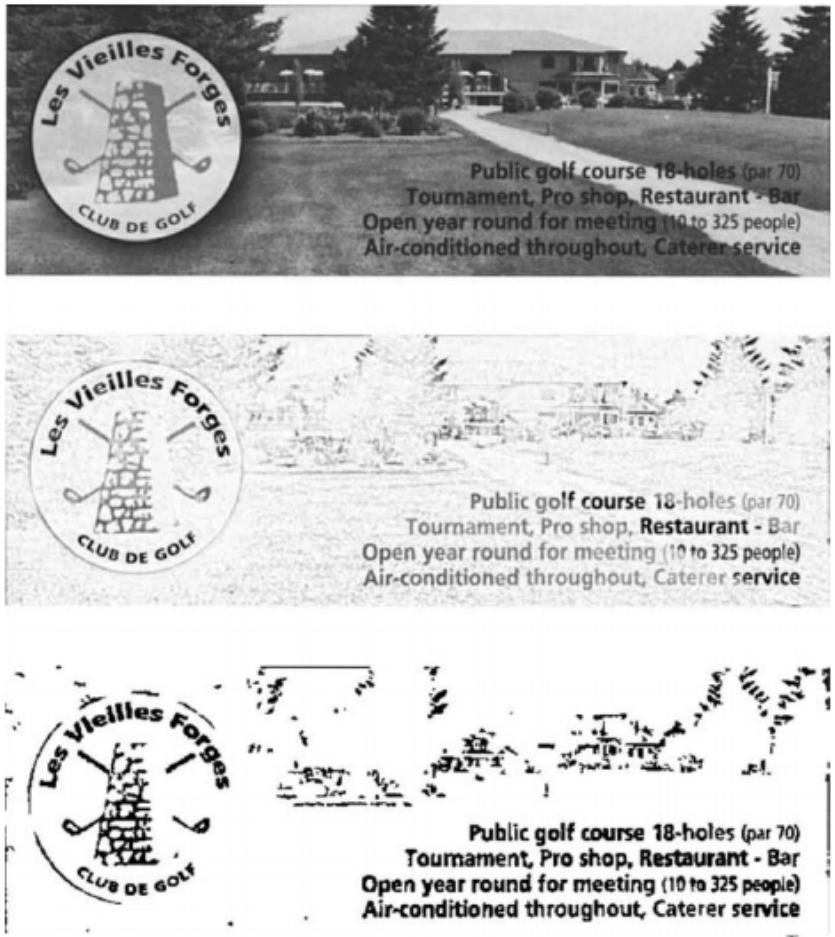
Feature Extraction

The feature extraction stage captures the distinctive characteristics of the digitized characters for recognition

Classification

The classification stage processes the feature vectors to identify the characters and words

- The pre-processing task is very important in order to acquire a very clear and distinct text



- Classification consists in converting the extracted patterns into the correct characters
- There are many different techniques that can be used to perform this
  - Template matching (first OCRs)
  - Artificial Neural Networks (from the 1980s)
  - Support Vector Machines (from the late 1990s)
- The techniques above will output the identified character
- Now you can decide what is going to be done to the characters identified (post-processing)

Thanks for your attention
I hope you enjoyed!