OCR

Obróbka Cyfrowa Materiału

Plan prezentacji

- Podstawowe pojęcia
- Historia OCR
- Zastosowanie OCR
- Opis dostępnych silników
- Opis procesu OCR
- Pozytywne i negatywne czynniki wpływające na jakość OCR

Podstawowe pojęcia

OCR – Obróbka cyfrowa materiału

OCR

- OCR – Optical Character Recognition (optyczne rozpoznanie znaków)
- Poprzez proces OCR rozumiemy przypisanie konkretnego znaku z alfabetu pikselom w zeskanowanym dokumencie
- Jest to konwersja obrazu do formy tekstowej
- W dokumencie tekstowym wyświetlanie znaków odbywa się poprzez odwoływanie się deklaracji w pamięci do przedstawienia znaku w pliku czcionki
- Czcionka jest specyficznie deklarowaną grafiką wektorową
- Dokument tekstowy zajmuje znacznie mniej miejsca niż jego kopia w formacie grafiki rastrowej czy wektorowej

ICR, IWR

- ICR – Intelligent Character Recognition, rozpoznawanie różnych rodzajów pisma wraz z właściwościami czcionki (rodzaj czcionki, krój pisma, akapit, kolumny)
- ICR to też rozpoznanie struktury dokumentu
- Większość silników typu ICR jest związana również z douczaniem programu
- Silnik OCR pozwalający na douczanie jest budowany na bazie sieci neuronowej
- IWR – Intelligent Word Recognition, rozpoznawanie słów
- Rozpoznawanie słów zamiast znaków wymaga większej mocy obliczeniowej, zastosowanie ma w modelach rozpoznawania pisma odręcznego
- OMR – Optical Music Recognition (optical?)
- SR – Speech Recognition, rozpoznawanie mowy

Historia OCR

OCR – Obróbka cyfrowa materiału

Historia OCR

- W 1914r. rozpoczęto prace nad maszyną konwertująca tekst drukowany na kod telegraficzny (tzw. Optofon)
- W 1929r. w Niemczech przyznano pierwszy patent na system OCR (rozpoznawanie z fotografii)
- W 1933r. Pierwszy patent zatwierdził UP USA
- W 1949r. w USA opracowano pierwszą metodę mechanicznego odczytywania tekstu
- Firma IMR opracowała pierwszy komercyjny system OCR w 1953 r.
- Od 1965r. Urząd Pocztowy USA stosuje OCR do sortowania listów
- W 1974 roku R. Kurzweil opracował metodę odczytywania tekstu z różnych czcionek (wykupiony przez firmę Xerox)

Historia OCR c.d.

- W 1980r. Rozpoczęły się prace nad rozpoznawaniem tekstu odręcznego
- W 1985r. powstał silnik Tesseract, od 2005 w open source, rozwijany przez Google od 2006 roku
- W 1987r. Stworzono silnik Read I.R.I.S., rozwijany do 2009 roku, własność firmy I.R.I.S.
- ABBYY Finereader (od 1989r.)
- W 1990r. Zaaplikowano do tabletów odczyt pisma odrecznego
- W 1991r. Stworzono program MIDISCAN (OMR)
- W 1996r. pojawia się pierwszy darmowy silnik OCR
- W 1997r. Urząd Pocztowy USA wprowadza system rozpoznawania ręcznie zapisanego adresu

Zastosowanie OCR

OCR – Obróbka cyfrowa materiału

Jakie są zalety zawarcia tekstu w dokumencie?

- Możliwość zmniejszenia wielkości pliku
- Możliwość wykorzystywania rozpoznanego tekstu w innych programach
- Możliwość wyszukiwania słów kluczowych w treści dokumentu
- Możliwość odczytania tekstu za pomocą syntetyzatora mowy
- Tzw. Text Mining (eksploracja tekstu), który przydaje się w wyszukiwaniu informacji

Pozostałe zalety OCR

- Przekształcanie tekstu odręcznego w tabletach
- Automatyzacja pracy za pomocą analizy treści dokumentu
- Odczytywanie tekstu z plików wideo czy archiwalnych materiałów zdjęciowych (m.in. wykorzystanie w policji np. przy poszukiwaniu nr rejestracyjnego samochodu poprzez kamery miejskie)

Opis procesu OCR

OCR – Obróbka cyfrowa materiału

Proces OCR

- Proces zaczyna się od wydzielenia w pliku grafiki rastrowej sekcji zawierających tekst – tworzenie struktury dokumentu
- Na podstawie deklaracji rodzaju sekcji dalsze rozpoznanie będzie formatowane wg narzuconego schematu
- Każdy silnik rozpoznawania tekstu z pliku grafiki rastrowej musi sekcje pikseli przyporządkować konkretnemu znakowi w zestawie czcionek
- W pliku czcionki każdy schemat jest czarnobiały
- Dokument zostaje więc sprowadzony do wersji czarnobiałej
- Czarnobiałe piksele są następnie przyporządkowane wektorom zawartym w silniku
- Następnym etapem jest weryfikacja kilku znajdujących się obok siebie znaków w oparciu o słownik

III poziomy OCR

- Rozpoznanie układu dokumentu
- Rozpoznanie znaków
- Weryfikacja słownikowa słów
- Douczanie silnika OCR

- Weryfikacja gramatyki
- Tzw. Neocognitron – stosowanie dwóch rodzajów wzorów, prostego i złożonego

K. Fukushima, *A hierarchical neural network model for selective attention,* [w:] „Neural computers", R. Eckmiller, C. Von der Malsburg, s. 81-90, 1987.

- Analiza ruchów pisma (rozpoznawanie pisma odręcznego)

Metody douczania

- Douczanie może odbywać się na każdym z wymienionych wcześniej etapów
- Silnik ABBYY pozwala użytkownikowi na douczanie rozpoznawania znaków i dodawanie wyrażeń słownikowych
- Czy weryfikacje metodą słownikową możemy nazwać IWR?
- IWR to rozpoznawanie całych słów z bazowych pikseli, nie zaś weryfikacja znaków
- IWR bazuje na metodzie neokognitronowej
- IWR zastosowanie ma głównie w rozpoznawaniu pisma odręcznego, ale podejrzewa się, że w przyszłości będzie to główna metoda rozpoznawania tekstu we wszelkich dokumentach

Pozytywne i negatywne czynniki wpływające na proces OCR

OCR – Obróbka cyfrowa materiału

Badania nad jakością OCR

Dostępne silniki OCR, zwłaszcza komercyjne, zostały poddane ocenie przez ośrodki badawcze na całym świecie

Zestawienie wyników m.in. w:

- Holley, Rose. „How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs". *D-lib Magazine* 15, nr 3/4 (2009). http://www.dlib.org/dlib/march09/holley/03holley.html.
- Mühlberger, Günter. „Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text- und Strukturerkennung (OCR)". *Zeitschrift für Bibliothekswesen und Bibliographie* 58, nr 1 (2011): 10–18.
- Kalota, Tomasz, Rafał Raczyński, i Paweł Rękar. „Przetwarzanie i OCR czasopism drukowanych gotykiem krok po kroku". Zredagowane przez Cezary Mazurek, Maciej Stroiński, i Jan Węglarz. *Polskie biblioteki cyfrowe 2010*. Poznań: Ośrodek Wydawnictw Naukowych, 2011.
- Powell, Tracy, i Gordon Paynter. „Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images". *D-lib Magazine* 15, nr 3/4 (2009).
- Tanner, Simon, Trevor Muñoz, i Pich Hemy Ros. „Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th Century Online Newspaper Archive". *D-lib Magazine* 15, nr 7/8 (2009). http://www.dlib.org/dlib/july09/munoz/07munoz.html.

Czynniki

Proces	Wpływ
Dobór materiału	Jakość dokumentu Grubość strony (przebijanie)
Skanowanie	Format pozyskanego pliku Wady techniczne lampy skanującej Rozdzielczość skanowania
Korekta graficzna	Profil kolorów Głębia bitowa Kompresja

Czynniki

Proces	Wpływ
Rozpoznawanie struktury	Kadrowanie Prostowanie Różnorodność układu stron Jasne przestrzenie pomiędzy kolumnami tekstu Brak zanieczyszczeń w skrajnych częściach strony
Analiza powierzchni należącej do jednego znaku	Kontrast-Gamma-Jasność Przyciemnienie pikseli przedstawiających znak Kontrast pomiędzy pikselami należącymi do znaku a tłem DPI Tresholding

Czynniki

Proces	Wpływ
Dopasowywanie grupy pikseli do wzorcowych obrazów	Ilość wzorców w bazie Możliwość tworzenia nowych wzorców Algorytm przyporządkowujący
Weryfikacja słownikowa	Zawartość słownika Algorytm dopasowujący wyrażenia do słów Możliwość zmieniania słownika Możliwość importowania słowników z innych programów
Douczanie oprogramowania	Czas

Czynniki

Proces	Działanie
Dobór materiału	Wykonywanie digitalizacji na materiałach najbardziej popularnych – ochrona zbiorów Dobór dobrej jakości egzemplarza
Skanowanie	Wybór formatu kompresji bezstratnej Rozdzielczość min. 300DPI Obraz niekadrowany w kolorze
Korekta graficzna	Wyostrzenie tekstu Jeżeli obraz w skali szarości – konwersja do czarnobiałego Ocena skanu

Czynniki

Proces	Działanie
Rozpoznawanie struktury	Prostowanie/kadrowanie Zmiana kontrastu z uwzględnieniem granic pomiędzy sekcjami Nie można poprawić układu stron zwykłą korektą graficzną
Analiza powierzchni należącej do jednego znaku	Kontrast-Gamma-Jasność Przyciemnienie pikseli przedstawiających znak Kontrast pomiędzy pikselami należącymi do znaku a tłem DPI Tresholding

Czynniki

Proces	Działanie
Dopasowywanie grupy pikseli do wzorcowych obrazów	Obór oprogramowania Powiększanie bazy wzorców Douczanie
Weryfikacja słownikowa	Import słowników z różnych programów Dodawanie wyrażeń słownikowych
Douczanie oprogramowania	Współpraca zespołu Tworzenie schematów dla poszczególnych drukarni