Marcin PAZIO
POLITECHNIKA GDAŃSKA, WYDZIAŁ ELEKTRONIKI, TELEKOMUNIKACJI I INFORMATYKI,
KATEDRA SYSTEMÓW AUTOMATYKI

Lokalizacja tekstu w obrazie

Mgr inż. Marcin PAZIO

W 1989r. ukończył studia na Wydziale Elektroniki Politechniki Gdańskiej. Pracuje w Katedrze Systemów Automatyki wydziału Elektroniki, Telekomunikacji i Informatyki PG. Zajmuje się technikami przetwarzania i analizy obrazów oraz ich zastosowaniami w rozwiązaniach mogących nieść pomoc osobom niewidomym.

e-mail: Marcin.Pazio@eti.pg.gda.pl

Streszczenie

W naturalnym otoczeniu człowieka znajduje się duża ilość łatwo rozpoznawalnej informacji przedstawionej w postaci znaków graficznych i tekstu. Informacja taka jest bardzo przydatna w poruszaniu się w środowisku miejskim. Niestety, osoby z upośledzonymi funkcjami wzroku w sposób oczywisty pozbawione są możliwości korzystania z tego rodzaju przekazu. Istniejące na rynku systemy rozpoznawania tekstu (OCR) nie są niestety dostosowane do rozpoznawania tekstu zawartego w obrazach zarejestrowanych kamerą czy aparatem cyfrowym. W artykule przedstawiono, opracowane w Katedrze Systemów Automatyki Wydziału ETI PG, algorytmy wyszukiwania tekstu w obrazie oraz jego przetwarzania do postaci umożliwiającej rozpoznanie przez system OCR i odczytania treści za pomocą syntezatora mowy.

Słowa kluczowe: analiza obrazu, algorytmy lokalizacji tekstu.

Localization of text in images

Abstract

The system capable of localizing and reading aloud text embedded in natural scene images can be very helpful for blind and visually impaired persons - providing information useful in everyday life and increasing their confidence and autonomy. Even though the currently available optical character recognition (OCR) programs are fast and accurate, most of them fail to recognize text embedded in natural scene images. The goal of the algorithm described in this paper is to localize text-like image regions and pre-process them in a way that will make OCR work more reliably. The approach described in the paper is based on color image segmentation and segment shape analysis. Preliminary tests have shown that the proposed algorithm offers satisfactory detection rate and is pretty robust to typical text distortions, such as slant, tilt and bend.

Keywords: image analysis, text detection algorithms.

1. Wstęp

Człowiek wprowadza w swoje środowisko, szczególnie miejskie, dużą ilość informacji przedstawionej w postaci znaków graficznych i tekstu. Komunikaty te są pomocne w określeniu punktu w którym się znajdujemy lub pozwalają odnaleźć interesujące nas miejsca. Niejednokrotnie, informacja podawana jest pośrednio – na przykład poszukujemy sklepu znajdującego się obok reklamy, zawierającej określony tekst lub symbol.

Osoby z upośledzonymi funkcjami wzroku w sposób oczywisty pozbawione są możliwości korzystania z tych udogodnień a więc bardzo pomocne dla nich byłoby urządzenie pozwalające na zlokalizowanie napisów oraz znaków i symboli a następnie przekazujące ich treść za pomocą syntezatora mowy.

Przy projektowaniu systemu odczytującego napisy w naturalny sposób nasuwa się koncepcja połączenia urządzenia (w postaci kamery lub aparatu cyfrowego) rejestrującego obrazy, oprogramowania do rozpoznawania pisma OCR (ang. Optical Recognition System) oraz syntezatora mowy. Niestety, w praktyce okazuje się, iż programy OCR, powszechnie stosowane w skanerach, są nieskuteczne przy rozpoznawaniu tekstu zawartego w obrazach scen naturalnych, ponieważ tekst taki zwykle umieszczony jest na niejednorodnym tle, ma niedostatecznie duży kontrast w stosunku do tła, ustawiony jest ukośnie w stosunku do krawędzi obrazu i może być zdeformowany geometrycznie na skutek odwzorowania optycznego z przestrzeni trójwymiarowej na dwuwymiarową. Dlatego, dla skutecznego odczytania tekstu, przed jego analizą przy użyciu programu OCR, konieczna jest automatyczna lokalizacja tekstu, jego ekstrakcja oraz korekcja.

2. Metody lokalizacji tekstu w obrazie

W procesie poszukiwania napisów w obrazie możemy założyć, że stanowią one grupy obiektów „literopodobnych" o pewnej regularności przestrzennej oraz wspólnych dla grupy cechach, takich jak barwa, położenie w obrazie i wysokość.

W chwili obecnej znanych jest wiele metod wyszukiwania tekstu w obrazie lub sekwencji wideo. Metody te można podzielić na trzy podstawowe grupy:

- metody działające w oparciu o analizę krawędziową [1],
- metody wykorzystujące analizę tekstur [2],
- metody analizujące segmenty przetworzonego obrazu [3].

W systemie opracowanym w Katedrze Systemów Automatyki zastosowane zostało trzecie podejście, w którym poszukiwanie segmentów spełniających określone kryteria uwzględniające kolorystykę i wzajemne geometryczne relacje, pozwala na lokalizację krótkich sekwencji, składających się z kilku liter.

3. Algorytm wyszukiwania tekstu w obrazie

Konstrukcja algorytmu wyszukiwania tekstu oparta została na założeniu, iż napisy rozmieszczone są wzdłuż linii prostych, napisane zostały czcionką o takiej samej wysokości, oraz że wszystkie znaki w obrębie napisu mają ten sam kolor. W praktyce, założenia te spełnione są przez większość tablic informacyjnych i oznaczeń miejsc w postaci nazwy sklepów, punktów usługowych itp.

Opracowany algorytm lokalizacji napisów w obrazie składa się z następujących kroków:

- segmentacja obrazu,
- filtracja segmentów,
- klasteryzacja,
- obróbka końcowa. [4]

Segmentacja obrazu

Zadaniem segmentacji jest podział obrazu na zbiór rozłącznych obszarów zwanych segmentami. W opracowanym algorytmie zastosowana została metoda segmentacji przez rozrost obszaru, do przeprowadzenia której niezbędne jest wskazanie tak zwanych pikseli zarodkowych. Typowanie pikseli zarodkowych odbywa się poprzez odnajdywanie takich pikseli, których parametry odpowiadają maksymalnym wartościom trójwymiarowego histogramu obrazu. Przetwarzanie obrazu oraz analiza histogramu wykorzystuje model barw CIE LAB. Zastosowanie tego modelu barw

pozwala na ustalenie takich kryteriów segmentacji, w których parametry barwy oraz jasności traktowane są odrębnie, dzięki czemu możliwe jest uodpornienie algorytmu na lokalne zmiany jasności, które mogą prowadzić do deformacji segmentów.

Segmentacja obrazu jest procesem czasochłonnym, a ponieważ dokładność odwzorowania kształtu znaków może znacząco wpływać na czas przetwarzania obrazu, konieczny jest kompromis pomiędzy jakością odwzorowania obiektów a szybkością działania algorytmu. Kształt znaków jest korygowany w ostatnim kroku algorytmu.

Filtracja segmentów

Filtracja segmentów ma na celu redukcję liczby segmentów poddawanych dalszemu przetwarzaniu. Na podstawie analizy geometrycznych cech segmentów eliminowane są te z nich, które nie odpowiadają znakom alfanumerycznym. Kryteria filtracji muszą dobrane w taki sposób aby nie odrzucić obiektów niosących użyteczną informację. W szczególności, należy uwzględnić zakłócenia i zniekształcenia występujące w obrazie, które mogą mieć wpływ na deformację kształtu znaku. Niekorzystne zjawiska, takie jak rozmycie obrazu, mały kontrast lokalny, nierównomierne oświetlenie oraz skróty perspektywiczne często prowadzą do powstawania segmentów odpowiadających kilku znakom lub do podziału znaków na odrębne segmenty. Sytuację dodatkowo komplikuje fakt, że w zarejestrowanym w naturalnym otoczeniu obrazie mogą znajdować się obiekty „literopodobne", które nie są literami.

W procesie filtracji segmentów analizowane są proste, niewymagające czasochłonnych obliczeń, cechy kształtu, takie jak:

- powierzchnia segmentu,
- względna wysokość segmentu,
- stosunek wysokości do szerokości prostokąta opisanego na segmencie,
- współczynnik wypełnienia prostokąta opisanego na segmencie.

Zakresy wartości wymienionych cech określone zostały na podstawie analizy obiektów będących obrazami liter, cyfr, połączeń kilku znaków lub fragmentami napisów i znaków.

Powierzchnię segmentu określa liczba pikseli należących do segmentu. Na tym etapie filtracji eliminowane są segmenty bardzo małe (mniej niż 10 pikseli), które są zbyt małe do prawidłowego rozpoznania przez OCR oraz segmenty bardzo duże, których rozmiar jest porównywalny z rozmiarem obrazu, a więc z dużym prawdopodobieństwem można uznać, że nie są fragmentem litery.

Względną wysokość segmentu określa wzór:

$$\alpha = \frac{S}{w}, \tag{1}$$

gdzie S jest powierzchnią segmentu, zaś w- szerokością prostokąta opisanego na segmencie. W procesie filtracji odrzucane są te segmenty, dla których $\alpha \geq \alpha_{\max}$, gdzie $\alpha_{\max}$ jest wartością progową wyznaczoną doświadczalnie, co pozwala na wyeliminowanie prostokątnych elementów, które są zbyt małe, aby mogły być rozważane jako znaki alfanumeryczne.

Stosunek wysokości do szerokości prostokąta opisanego na segmencie obliczany jest ze wzoru:

$$\beta = \frac{h}{w}, \tag{2}$$

gdzie h jest wysokością zaś w szerokością prostokąta opisanego na segmencie. W tej fazie filtracji eliminowane są segmenty, dla których $\beta \geq \beta_{\max}$ (zbyt wysokie i zbyt wąskie), a które mogłyby być łatwo mylone z literą „I".

Współczynnik wypełnienia prostokąta obliczyć można jako iloraz:

$$\gamma = \frac{S}{h \cdot w}, \tag{3}$$

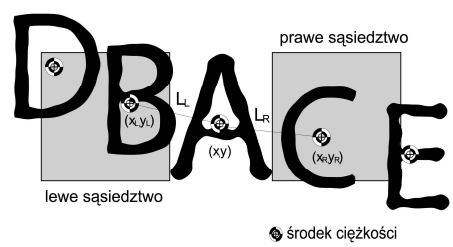
gdzie S jest powierzchnią segmentu, zaś h i w oznaczają odpowiednio wysokość i szerokość prostokąta opisanego na segmencie. Analiza tego współczynnika pozwala na wyeliminowanie tych segmentów, dla których $\gamma \leq \gamma_{\min}$, które często są elementami tła, takimi jak np. gałęzie.

Szybkość działania algorytmu poszukiwania tekstu w obrazie w znaczącym stopniu zależy od efektywności filtracji. Typowy obraz o wielkości 1 miliona pikseli dzielony jest na dwa do trzech tysięcy segmentów. Filtracja redukuję liczbę segmentów poddawanych dalszej obróbce o 80 - 85%.

Grupowanie

Grupowanie pozostałych po filtracji segmentów odbywa się dwuetapowo. W pierwszym etapie, wyizolowane „literopodobne" elementy łączone są w struktury „tekstopodobne", czyli klastry elementarne, a następnie klastry te łączone są w łańcuchy tekstowe.

Klastry elementarne tworzone są w oparciu o analizę barwy segmentów, wysokości prostokątów opisanych na segmentach oraz wzajemnego położenia środków ciężkości segmentów. Etapy tworzenia klastra możemy prześledzić na przykładzie pokazanym na rys. 1.



Rys. 1. Tworzenie klastrów elementarnych
Fig. 1. Formation of elementary text clusters

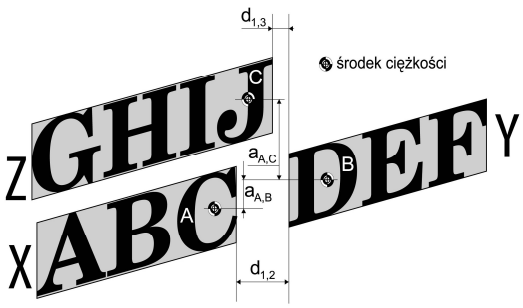
Proces ten rozpoczyna się od wyodrębnienia segmentów o jednakowej barwie (A,B,C,D,E). Następnie wyznaczane jest lewe i prawe sąsiedztwo rozpatrywanego segmentu (A); o rozmiarze tych sąsiedztw decyduje wysokość prostokąta opisanego na badanym segmencie. Na podstawie wzajemnego położenia środków ciężkości przylegających segmentów (X_L, Y_L) i (X_R, Y_R) wyznaczane są współczynniki kierunkowe linii L_R i L_L. Jeżeli istnieją oba sąsiednie segmenty, współczynniki te są uśredniane dla lewego i prawego segmentu.

W wyniku przeprowadzonej analizy podanych wyżej kryteriów, stwierdzić można, że segmenty A, B, C, D oraz E spełniają kryteria barwy i wysokości. Segment D zostanie jednak odrzucony, gdyż segment B jest bliżej A, zaś segment E musi zostać wyeliminowany z uwagi na to, że jego środek ciężkości znajduje się poza obszarem sąsiedztwa.

Segmenty, tworzące klastry elementarne, są „zarodkami" łańcuchów zawierających tekst. Klastry łączone są iteracyjnie w większe grupy w oparciu o:

- kolor elementów klastra, wysokości elementów klastra,
- wzajemne położenie środków ciężkości przylegających elementów sklejanych klastrów,
- współczynniki kierunkowe, wyznaczone na etapie tworzenia klastrów elementarnych.

Sposób tworzenia klastrów wskazujących tekst ilustruje rys. 2.



Rys. 2. Tworzenie łańcuchów
Fig. 2. Formation of text chains

Aby dwa sąsiednie klastry elementarne zostały połączone w łańcuch, odległości środków ciężkości przylegających elementów sklejanych klastrów *d* (w poziomie) oraz *a* (w pionie) muszą być dostatecznie małe w porównaniu ze średnią wysokością segmentów w klastrze. W przykładzie pokazanym na rys. 2. w łańcuch połączone zostaną klastry X oraz Y zaś klaster Z zostanie odrzucony, ponieważ odległość $a_{A,C}$ jest porównywalna z wysokością segmentu. Proces łączenia klastrów trwa do momentu, gdy wartość funkcji kryterialnej, wiążącej wymienione cechy, przekroczy wyznaczoną doświadczalnie wartość progową. Dalszej obróbce poddawane są klastry co najmniej dwuelementowe. Średnia wartość współczynników kierunkowych elementów klastra wyznacza linię bazową tekstu oraz jego nachylenie.

Obróbka końcowa

Zadaniem obróbki końcowej jest takie przetworzenie zlokalizowanych obszarów „tekstopodobnych", aby umożliwić poprawne działanie programu OCR. Wymaga to:

- korekty położenia napisu w stosunku do krawędzi obrazu,
- ponownej binaryzacji obszarów zawierających tekst,
- eliminacji zniekształceń geometrycznych powstałych w wyniku skrótów perspektywicznych.

Korekta położenia w stosunku do krawędzi obrazu jest możliwa w oparciu o informacje o współczynnikach kierunkowych prostych, wokół których leżą środki ciężkości segmentów stanowiących klastry. Współczynniki te obliczane są jako średnie wartości współczynników kierunkowych wyznaczonych dla klastrów elementarnych.

Ponowna binaryzacja dokonuje się w oparciu o uściślone w procesie klasteryzacji parametry kolorystyczne elementów klastra i jego tła. Lokalnie określona barwa pozwala na zdecydowanie lepsze odtworzenie kształtu znaków, niż przyjęte w procesie segmentacji globalne oszacowanie dla całego obrazu.

W pierwszym etapie obróbki końcowej dla każdego klastra wyznaczany jest opisujący go czworokąt. Segmenty znajdujące się wewnątrz wyznaczonego czworokąta i nie będące elementami klastra, są uznawane za elementy tła. Dla tych wszystkich segmentów ponownie obliczany jest średni kolor, ale z uwzględnieniem wyłącznie tych pikseli, które znajdują się wewnątrz wyznaczonego czworokąta.

W drugim etapie, obszar obrazu wyznaczony przez czworokąt jest obracany o kąt wynikający z wartości współczynników kierunkowych prostych, wokół których leżą środki ciężkości segmentów będących elementami klastra.

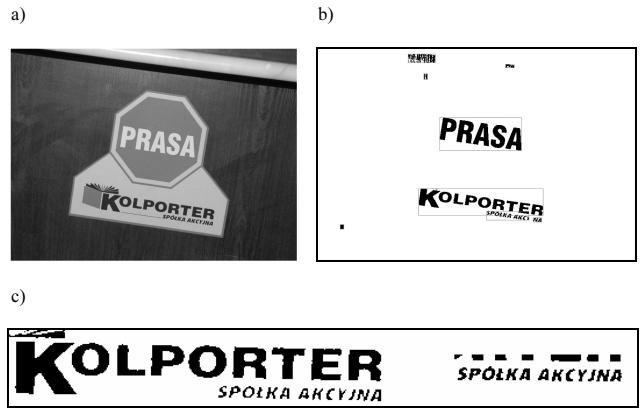
W trzecim etapie obrócony obszar obrazu poddawany jest binaryzacji. Do binaryzacji wykorzystywane są parametry kolorystyczne segmentów należących do klastra oraz wyznaczone w pierwszym etapie obróbki końcowej średnie kolory segmentów należących do tła. Dla każdego piksela obróconego fragmentu obrazu znajdowany jest segment najbliższy w przestrzeni Lab spośród segmentów znajdujących się w obszarze wyznaczonego czworokąta. Jeżeli najbliższy segment jest elementem tła, to piksel również uznawany jest za element tła, jeżeli zaś najbliższy segment jest elementem klastra – to piksel uznawany jest za element napisu.

Zniekształcenia geometryczne powstałe w wyniku skrótów perspektywicznych można wyeliminować stosując metodę polegająca na ocenie charakteru zniekształcenia poprzez analizę kształtu obwiedni tekstu oraz analizę histogramów jego rzutów poziomych [5].

4. Wyniki doświadczalne

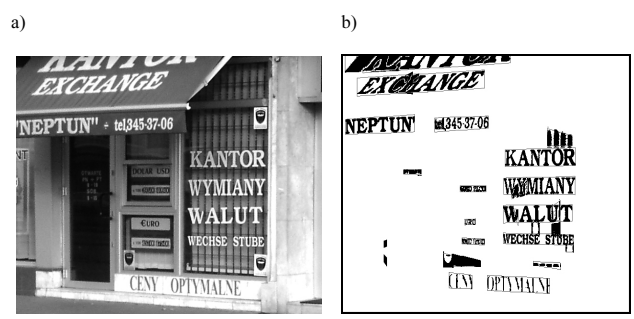
Opracowane algorytmy przetestowane zostały w procesie wyszukiwania napisów w obrazach niepoddawanych korekcji położenia i zniekształceń, co przedstawiono na rys. 3., rys. 4. i rys. 5. Obszary zawierające zlokalizowane teksty otaczane są ramką a rozpoznane litery i cyfry przedstawiane są w postaci czarnych znaków.

Przykład na zdjęciu z rys. 3. przedstawia dość prostą sytuację - zlokalizowany został napis PRASA oraz KOLPORTER i wyeliminowane zostały obiekty niebędące napisami. Rys. 3c ilustruje wynik działania obróbki końcowej, wraz z prostowaniem tekstu.



Rys. 3. Przykład lokalizacji prostego tekstu na szyldzie. Obraz oryginalny (a), wyodrębniony tekst (b) i obrócone napisy (c)
Fig. 3. Example of a simple board and the result of the text detection Original image (a) the isolated text (b) and rotated inscriptions

Zawartość obrazu przedstawionego na rys. 4. jest bardziej skomplikowana, ponieważ znajduje się tam stosunkowo dużo tekstu na niejednorodnym tle. Tekst został zlokalizowany i wyodrębniony ale konieczna jest dalsza obróbka, ponieważ nie wszystkie znaki będą mogły być rozpoznane przez system OCR.



Rys. 4. Obraz oryginalny (a) i obraz z wyodrębnionym tekstem (b)
Fig. 4. Original image (a) and the isolated text (b)

Natomiast na rys. 5 przedstawiony jest bardzo przydatny w życiu codziennym przypadek, a mianowicie zlokalizowanie i rozpoznanie numeru domu i nazwy ulicy na tablicy umieszczonej na budynku.

Rys. 5. Wynik działania algorytmu lokalizacji w przypadku napisów na tablicy z numerem domu
Fig. 5. The result of text detection for the board with the house number and the name of the street

Znaki znajdujące się na zdjęciu tej tablicy mają stosunkowo mały rozmiar oraz zmniejszoną, poprzez zanieczyszczenia, czytelność (w dolnej części tablicy). Zastosowany algorytm ponownej binaryzacji pozwolił na prawidłowe odwzorowanie kształtu znaków.

5. Wnioski

Wstępne badania testowe wykazały, że opracowane algorytmy skutecznie wyszukują i odwzorowują napisy na zarejestrowanych obrazach. Istotnym problemem, który będzie przedmiotem dalszych badań, jest nadmiar pozyskiwanej informacji tekstowej. Dlatego konieczne będzie wprowadzenie odpowiedniej filtracji danych, tak aby osobie niewidomej przekazywane były tylko te informacje, które są w danej chwili istotne.

Praca naukowa finansowana ze środków na naukę w latach 2005-2007 jako projekt badawczy.

6. Literatura

[1] N. Ezaki, M. Bulacu and L. Schomaker: Text detection from natural scene images: towards a system for visually impaired persons, Proc. 17th IEEE International Conference on Pattern Recognition, ICPR'04, 2004, pp. 683–686
[2] S. Ferreira, C. Thilou and B. Gosselin: From picture to speech: an innovative application or embedded environment, Proc. of the 14th ProRISC Workshop on Circuits, Systems and Signal Processing, ProRISC'03, 2003
[3] M. Pazio, M. Niedźwiecki, R. Kowalik, J. Lebiedź: Text detection system for the blind, EUSIPCO 2007, pp. 272-276
[4] M. Pazio: Analiza przydatności wybranych współczynników kształtu do oceny podobieństwa do litery, Zeszyty Naukowe ETI PG, Nr 5, T. 13 (2007), s. 539-546
[5] Myers, et al., "Recognition of Text in 3-D Scenes," Proc. 4th Symp on Document Image Understanding Technology, Apr. 23-25, 2001, pp. 85–99

Artykuł recenzowany