BRONISŁAWÓW 2005

WOJCIECH BIENIECKI*

ANALIZA WYMAGAŃ DLA METOD PRZETWARZANIA WSTĘPNEGO OBRAZÓW W AUTOMATYCZNYM ROZPOZNAWANIU TEKSTU

1. Wprowadzenie

Automatyczne rozpoznawanie tekstu (ang. OCR – *optical character recognition*) jest ważnym zagadnieniem w dziedzinie przetwarzania obrazów. Zastosowanie tej techniki to przede wszystkim:

- archiwizacja i udostępnianie dokumentów: przekształcanie do postaci elektronicznej książek, gazet i pozostałych materiałów drukowanych [3], w tym także równań i symboli matematycznych [18], zapisu nutowego, itp;
- systemy identyfikacji, monitorowania i sterowania: np. rozpoznawanie tablic rejestracyjnych w pojazdach [13], [1], identyfikacja oznaczeń na obiektach [7];
- analiza zdjęć w diagnostyce medycznej [9];
- inne: systemy czytające dla osób niewidomych [14], interfejs komunikacji w palmtopach, systemy weryfikacji dokumentów wypełnianych ręcznie [4], [15].

Systemy rozpoznawania tekstu działają zwykle według pewnego schematu:

1) Przetwarzanie wstępne obrazu [7], [16], [5]. Zazwyczaj ogranicza się ono do wyrównania histogramu jasności obrazu.
2) Binaryzacja obrazu. Najczęściej obraz jest przekształcany do postaci binarnej przy użyciu jednej z metod progowania adaptacyjnego [20].
3) Segmentacja [12]. Etap ten jest bardzo ważny dla działania całego systemu. Segmentacja rozpoczyna się od określenia orientacji dokumentu i rozpoznania układu strony. Procedury segmentacji wydzielają ze strony obszary zawierające akapity tekstu, obrazy oraz tabele. Odpowiednie procedury mogą też dokonać przekształcenia na negatyw oraz korekty geometrii obrazu, zazwyczaj jest to obrót. Wydzielone obszary tekstu segmentuje się dalej. Rozdziela się linie tekstu, następnie słowa, a ostatecznie znaki w słowach.
4) Rozpoznawanie znalezionych znaków i przekształcenie dokumentu do postaci tekstowej. Proces rozpoznawania znaków jest algorytmem zazwyczaj dość złożonym, wykorzystującym wybrane algorytmy klasyfikacji nadzorowanej lub nienadzorowanej [8], [10], [19].

*mgr inż. Katedra Informatyki Stosowanej, Politechnika Łódzka, al. Politechniki 11, 90-924 Łódź. e-mail: wbieniec@kis.p.lodz.pl

5) Korekcja ortograficzna przy pomocy słownika. Algorytm sprawdza, czy rozpoznane słowa znajdują się w słowniku i koryguje błędy, które mogły wystąpić na skutek nieprawidłowego rozpoznania znaków występujących w słowie.
6) Zapisanie tekstu i obrazów w żądanym formacie PDF, DOC, LaTeX lub HTML. W przypadku programów takich jak FineReader lub Recognita można zażądać zapisania kompletnego układu strony, lub ograniczyć się tylko do zachowania kroju i wielkości czcionek, można też zrezygnować z zapisywania wszelkiej innej niż tekst treści. Format PDF dopuszcza także, aby słowa (lub znaki), co do których nie ma pewności, czy są prawidłowo rozpoznane, zapisywać jako obraz.

2. Cel wprowadzenia dodatkowych algorytmów przetwarzania wstępnego

Rozważając algorytmy przetwarzania wstępnego obrazu należy zadać pytanie, jakie są zadania tych algorytmów i w jakim stopniu prawidłowa realizacja tych zadań ma wpływ na dalszy proces przetwarzania strony.

Przede wszystkim wybór odpowiednich algorytmów zależy od samego materiału źródłowego. Załóżmy od razu, że mamy do czynienia z tekstem drukowanym – systemy automatycznego rozpoznawania pisma odręcznego wymagają innego podejścia niż opisane we wstępie [2], [11], [17]. Jakość materiału ma zasadniczy wpływ na jakość obrazu w postaci cyfrowej. Najczęściej przetwarzane dokumenty to:

- książki;
- pojedyncze kartki z drukarki laserowej, z drukarki igłowej, z faksu, kserokopie;
- wygenerowany komputerowo tekst w formie grafiki rastrowej (format PS, PDF, TIFF, JPEG).

Dokumenty przekształcane są do postaci cyfrowej najczęściej przy użyciu skanera płaskiego, kamery lub aparatu cyfrowego, kamery analogowej wideo. Zwykle obrazy przed wprowadzeniem do pamięci programu analizującego zapisywane są w plikach rastrowych. Zapisany obraz charakteryzuje: rozdzielczość, liczba kolorów, rodzaj kompresji (stratna/bezstratna).

Ustalając warunki przeprowadzenia eksperymentu należy wskazać, jakie defekty obrazu mogą być potencjalnie przyczyną błędów w działaniu programu OCR. Wstępnie wskazano następujące defekty:

Zaszumienie obrazu. Szum w obrazie może wpływać na zdolność programu OCR do rozpoznawania znaków. Dodatkowa kropka lub kreska w pobliżu danego znaku może spowodować błędy klasyfikacji. Szum pojawia się często w dokumentach pochodzących z faksu lub kserografu, a także może on powstać przy akwizycji obrazu z kamery lub aparatu przy niedostatecznym oświetleniu. Istnieje wiele metod przetwarzania wstępnego dla usuwania szumu z obrazu, najczęściej wykorzystuje się odpowiednio dobrany filtr medianowy, operacje morfologiczne erozji i dylatacji oraz filtry dolnoprzepustowe. Programy OCR zwykle nie dają możliwości wprowadzania dodatkowej filtracji obrazu.

Zbyt mała rozdzielczość obrazu. Algorytmy segmentacji i rozpoznawania zaimplementowane w systemie OCR nie są w stanie prawidłowo oszacować kształtu znaków, jeśli rozdzielczość nie jest wystarczająca. Dla dokumentów drukowanych normalną czcionką (10 lub 12 pt.) zwykle minimalna rozdzielczość gwarantująca prawidłową pracę systemu

wynosi 300 DPI. Niedostateczna rozdzielczość może wystąpić w przypadku, gdy obraz jest pozyskiwany z kamery analogowej lub aparatu cyfrowego ze znacznej odległości. Na rozdzielczość obrazu ma również wpływ format zapisu obrazu w pliku graficznym. Ograniczona głębia koloru lub ograniczona liczba odcieni szarości (np. do 16) może spowodować błędy binaryzacji. Zapis obrazu z użyciem kompresji stratnej (JPEG) może powodować błędy w rozpoznawaniu kształtu znaków, ponieważ algorytmy kompresji stratnej powodują rozmycie krawędzi. Obraz o zbyt małej rozdzielczości można próbować powiększyć korzystając z algorytmów interpolacji liniowej, kwadratowej lub wykorzystującej przekształcenia fraktalowe.

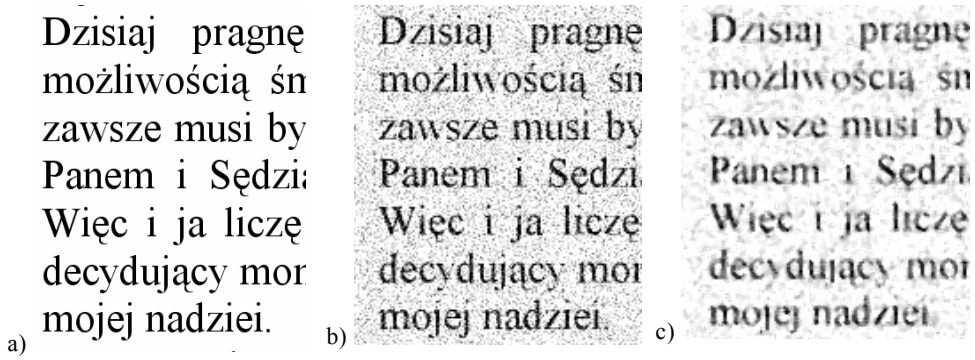
Niejednorodne tło. Niejednorodność tła w dokumencie tekstowym może mieć zasadniczy wpływ na etap binaryzacji obrazu. Dokumenty archiwalne mogą być zaplamione, papier może być pożółkły. Tekst przeznaczony do rozpoznania nie zawsze jest drukowany czarną czcionką na białym papierze, tło może być kolorowe podobnie jak litery. W przypadku wykonywania kopii książki część środkowa może zostać gorzej naświetlona. Problem nierównomiernego oświetlenia jest również charakterystyczny dla wszystkich materiałów uzyskiwanych metodą fotografii dokumentów. Usunięcie defektów obrazu jest zwykle trudną operacją niedostępną w programach OCR ogólnego stosowania. Filtr musi być ściśle dopasowany do rodzaju występującego szumu. Filtracja wyrównująca tło o nierównomiernym naświetleniu lub niejednorodnym kolorze polega na tzw. odjęciu tła. Zaprojektowanie takiego filtru wymaga uzyskania obrazu referencyjnego, czyli takiego, z którego zostały usunięte przeznaczone do rozpoznania znaki. Przybliżony obraz referencyjny można uzyskać poprzez filtrację dolnoprzepustową, metody analizy tekstury lub operacje morfologiczne. Po operacji odjęcia obrazu referencyjnego otrzymuje się poprawiony obraz z jednorodnym tłem. Alternatywną dla progowania metodą binaryzacji w przypadku, gdy rozpoznaje się tekst na obrazie barwnym może być jedna z metod klasyfikacji pikseli.

Zniekształcenia geometryczne. Podstawowym rodzajem zniekształcenia geometrycznego jest obrót. Obrót nawet o niewielki kąt może spowodować błędy w segmentacji linii tekstu lub komórek tabeli. Zwykle nowoczesne programy OCR są wyposażone w algorytmy automatycznej korekcji obrotu. W przypadku kopiowania książki może się pojawić problem zniekształcenia obrazu przy wewnętrznym marginesie książki, ponieważ kartka się zawija. Jeśli zniekształcenie nie jest duże, algorytmy rozpoznawania tekstu radzą sobie z tym zadaniem. Obrazy pozyskiwane poprzez fotografowanie charakteryzują się zazwyczaj największymi zniekształceniami geometrii. Można wyróżnić zazwyczaj trzy powody takich zniekształceń: efekt perspektywy zbieżnej (obiektyw nie jest ustawiony prostopadle do fotografowanej powierzchni), efekt rybiego oka (zniekształcenie powodowane przez optykę aparatu przy zbyt małej odległości od fotografowanej powierzchni) oraz efekt spowodowany wygięciem powierzchni (powstaje zwłaszcza przy fotografii książki lub odczytywaniu napisu namalowanego na przestrzennym obiekcie). Duże zniekształcenia geometrii mogą całkowicie uniemożliwić rozpoznanie tekstu, w testowanych programach OCR poza korekcją obrotu nie ma żadnych innych procedur kompensacji zniekształceń. Zaprojektowanie algorytmów korekcji geometrii będzie przedmiotem dalszych badań autora artykułu, prace będą prowadzone na podstawie metody zaproponowanej w [6].

3. Eksperyment

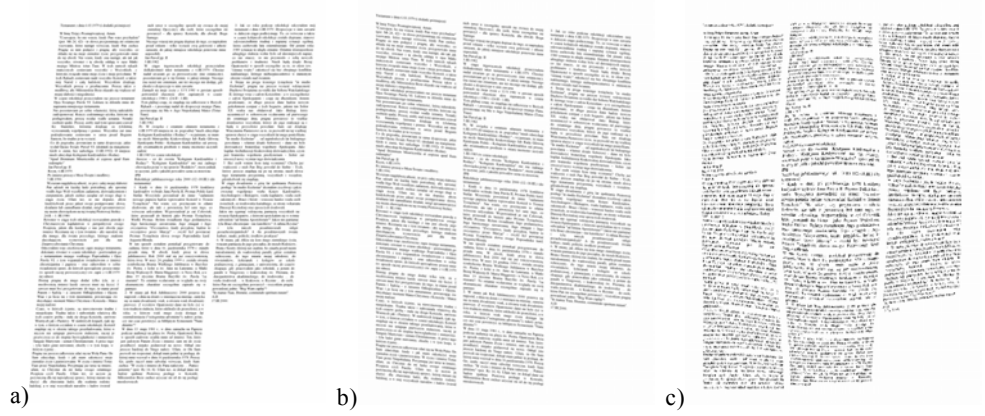
Testowaniu podlegał program FineReader 7.0 w wersji Office[1]. Materiałem testowym była jedna strona gazety zapisana w pliku PDF z rozdzielczością 600 DPI. Wielkość czcionki to w przybliżeniu 8pt. Oryginalny tekst jest w języku polskim i łacińskim. Zawiera 1772 słów i 10091 znaków (nie licząc odstępów). W każdym z testów porównywano oryginalny tekst zapisany w pliku tekstowym z tekstem otrzymanym przez proces rozpoznawania badanego obrazu. Porównanie następowało słowo po słowie, gdzie zliczano różniące się na określonych pozycjach znaki. Do porównywania użyto własnego, zaprojektowanego w tym celu programu.

W pierwszym eksperymencie zbadano zdolność rozpoznawania badanego tekstu programem FineReader w zależności od rozdzielczości obrazu. Drugi eksperyment polegał na zaszumieniu obrazu, a następnie odszumieniu filtrem medianowym z maską 25 punktów (Rys. 1). Kolejne badanie polegało na zamodelowaniu najczęściej występujących zniekształceń obrazu, które mogą powstać zarówno w czasie pracy na skanerze, jak i z aparatem fotograficznym (Rys. 2). Ostatnie badanie miało na celu sprawdzenie wpływu niejednorodnego naświetlenia tła skanowanego obrazu na jakość rozpoznania. Nierównomierne oświetlenie jest problemem występującym w przypadku użycia aparatu fotograficznego (Rys. 3). Wyniki w postaci liczby popełnionych pomyłek przedstawiono w Tabelach 1-4.

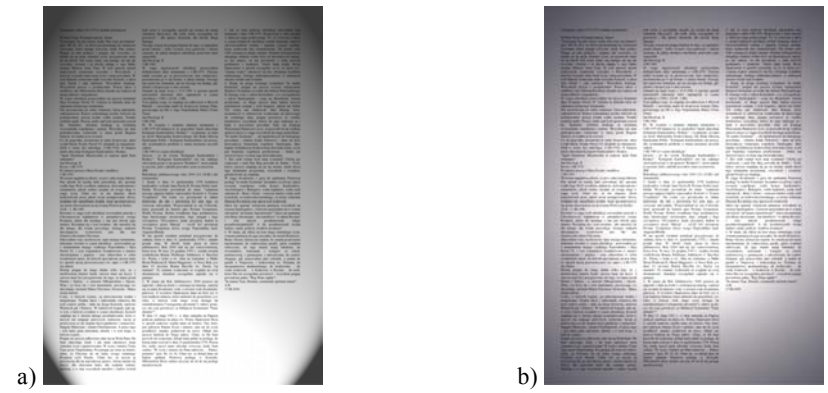


Rys. 1. Zaszumianie obrazu a) fragment obrazu oryginalnego, b) dodany szum Gaussa 30%, c) zastosowanie filtru medianowego

[1] © 1996-2005 ABBYY Software House, http://www.abbyy.com/



Rys. 2. Zniekształcenie obrazu: a) skos w prawo 10°, b) skos w dół 10°, c) „poduszka” 20%



Rys. 3. Modelowanie oświetlenia obrazu: a) model 1, b) model 2

Tabela 1

Liczba błędów rozpoznawania dla różnych rozdzielczości

Obraz	Rozdziel-czość	Liczba pomyłek	Błąd względny [%]
Oryginalny	600 DPI	18	0,17
Pomniejszony 2x	300 DPI	52	0,52
Pomniejszony 4x	150 DPI	190	1,88
Pomniejszony 6x i powiększony 3x	300 DPI	126	1,25

Tabela 2

Liczba błędów rozpoznawania dla obrazów zaszumionych

Szum	Rozdziel-czość	Liczba pomyłek	Błąd względny [%]	Uwagi
10%	300 DPI	148	1,47	
20%	300 DPI	1442	14,29	
30%	300 DPI	5602	55,51	Trudności z określeniem obszarów tekstu
10% + mediana	300 DPI	534	5,29	
20% +mediana	300 DPI	578	5,73	
30% + mediana	300 DPI	4486	44,46	
10%	600 DPI	25	0,25	
20%	600 DPI	144	1,43	
30%	600 DPI	320	3,17	
10% + mediana	600 DPI	41	0,41	
20% +mediana	600 DPI	132	1,31	
30% + mediana	600 DPI	454	4,50	

Tabela 3

Rozpoznawanie obrazu zniekształconego (600 DPI)

Zniekształcenie	Segmentacja	Liczba pomyłek	Błąd wzgl. [%]
Obrót 2°	Automatycznie obrócony obraz	18	0,18
Obrót 5°	Automatycznie obrócony obraz	18	0,18
Obrót 10°	Nie wyznaczono granic tekstu	-	-
Skos w prawo 5°	Źle wyznaczono podziały wierszy	18	0,18
Skos w prawo 10°	Źle wyznaczono podziały wierszy	45	0,45
Skos w dół 5°	Stronę obrócono tak, aby linie były poziome	18	0,18
Skos w dół 10°	Stronę obrócono tak, aby linie były poziome	45	0,45
Poduszka 10%	Miejscami źle wydzielony tekst	560	5,55
Poduszka 20%	Źle wydzielony tekst	1875	18,58

Tabela 4

Rozpoznawanie obrazu źle oświetlonego (600 DPI)

Model	Segmentacja	Liczba pomyłek	Błąd wzgl. [%]
Model 1	Poprawna	28	0,28
Model 2	Poprawna	33	0,33

4. Analiza wyników i wnioski

Pierwszy eksperyment pokazał, że rozdzielczość ma istotnie duży wpływ na jakość rozpoznawania znaków. Poniżej 300 DPI (przy wielkości czcionki takiej, jak w eksperymencie) program w zasadzie nie jest w stanie pracować. Segmentacja linii tekstu przebiega poprawnie także przy mniejszej rozdzielczości. Obraz pomniejszony do rozdzielczości 100 DPI został przez program FineReader odrzucony. Ciekawy jest przypadek, gdy obraz został najpierw pomniejszony do rozdzielczości 100 DPI, a następnie powiększony trzykrotnie. Liczba błędów jest dwukrotnie większa niż w przypadku obrazu zmniejszonego normalnie do rozmiaru 300 DPI, lecz nadal akceptowalna. Stąd wniosek, że przy zastosowaniu odpowiedniego algorytmu resamplingu można poprawić jakość rozpoznawania pomimo, iż obraz wejściowy był w niedostatecznej rozdzielczości.

Zaszumienie obrazu ma zasadniczy wpływ na jakość rozpoznawania. Im większy szum, tym gorzej działa mechanizm rozpoznawania, lecz także występują zakłócenia w procesie wyznaczania obszarów tekstu. Filtr medianowy zastosowany w celu rekonstrukcji obrazu okazuje się nie wystarczający, wręcz może pogorszyć sytuację. Operacja mediany może spowodować rozmycie krawędzi liter lub nawet usunięcie fragmentów czcionek, w szczególności znaków interpunkcyjnych.

Zniekształcenia obrazu również wpływają na pracę programu OCR, w szczególności na proces segmentacji linii tekstu. W przypadku zbyt dużego kąta pochylenia program nie jest w stanie określić kierunku teksu i skompensować obrotu. Wprowadzenie zniekształceń typu skos również może powodować błędy segmentacji linii, jednak nie ma wpływu na jakość rozpoznawania poszczególnych liter. W przypadku zniekształceń nieliniowych program nie potrafi przeprowadzić prawidłowej segmentacji linii tekstu, co skutkuje tym, że dalsze przetwarzanie nie jest możliwe. Zniekształcenia geometryczne są możliwe do skompensowania przez proste algorytmy mapowania pikseli, jednak pewną trudność może stanowić automatyczna analiza parametrów tego przekształcenia. Konstrukcja takich algorytmów będzie stanowić dalszy etap prowadzonych przez autora artykułu badań.

Nierównomierne oświetlenie obrazu ma znikomy wpływ na jakość zarówno segmentacji obszarów tekstowych, jak i automatycznego rozpoznawania znaków. Wprowadzone modele oświetlenia były wygenerowane sztucznie. Istnieje jednak podejrzenie, w naturalnym obrazie obszary niedostatecznie naświetlone będzie charakteryzował większy szum, co może powodować błędy.

Podsumowując przeprowadzone eksperymenty należy stwierdzić, że najważniejszym etapem przetwarzania wstępnego jest korekcja geometrii obrazu. Zniekształcenia geometrii

są atrybutem obrazów powstających przy użyciu urządzeń mobilnych, czyli kamer i aparatów fotograficznych. Są to jednocześnie urządzenia w stosunku do skanera dość szybkie, a ich główną przewagę stanowi możliwość wykonywania zdjęć z różnej odległości.

Literatura

[1] Bubliński Zb., Mikrut Zb.: *Lokalizacja tablic rejestracyjnych pojazdów*. Automatyka, t. 7, zeszyt 3, Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków 2003.

[2] Chaudhuri B. B., Pal U., Mitra M.: *Automatic recognition of printed oriya script*. Sadhana (A journal of Indian Academy of Sciences), vol. 27, Feb. 2002, pp. 23 – 34.

[3] Gatos B., Mantzaris S. L., Perantonis S. J., Tsigris A.: *Automatic page analysis for the creation of a digital library from newspaper archives*. International Journal on Digital Libraries (IJODL),vol. 3(1), pp. 77 – 84, 2000.

[4] Gorski N., Anisimov V., Augustin E., Baret O., Maximov S.: *Industrial bank check processing: the A2iA CheckReader*. Internationl Journal on Document Analysis and Recognition, 3: 196-206, Springer-Verlag, 2001.

[5] Herceg P., Huyck B., Johnson C., Kundu A., Van Giulder L.: *Optimizing OCR accuracy for bi-tonal, noisy scans of degraded Arabic documents*. Technical Paper, The MITRE Corporation, 2005.

[6] Hujka P.: *Model of geometric distortion caused by lens and method of its elimination*. ElectronicsLetters.com - http://www.electronicsletters.com, ISSN 1213-161X, vol. 1/4/2004, pp. 1 – 3.

[7] Jung K., Kim K. I., Jain A. K.: *Text information extraction in images and video: a survey*. PR(37), No. 5, May 2004, pp. 977 – 997.

[8] Kauniskangas H.: *Document image retrieval with improvements in database quality*. Oulu University Library, ISBN: 951-42-5313-2. ISBN: 951-42-5313-2, 1999.

[9] Kou Z., Cohen W. W., Murphy R. F.: *Extracting Information from Text and Images for Location Proteomics*. BIOKDD 2003: 2 – 9.

[10] Lu Z., Bazzi I., Kornai A., Makhoul J., Natarajan P., Schwartz R.: *A robust, language-independent OCR system*. Proc. 27th AIPR Workshop: Advances in Computer-Assisted Recognition SPIE Proceedings 3584 1999.

[11] Ma H., Doebermann D.: *Adaptive, hindi OCR using generalized Hausdorff image comparison*. ACM Transactions on Asian Language Information Processing (TALIP), Volume 2, Issue 3, September 2003, pp. 193 – 218.

[12] Mao S., Rosenfeld A., Kanungo T.: *Document structure analysis algorithms: a literature survey*. Proc. SPIE Electronic Imaging. 2003 Jan; 5010: 197 – 207.

[13] Parker J. R., Federl P.: *An Approach To License Plate Recognition*. Vision Interface '97, Kelowna, B.C., May 20-22, 1997. Department of Computer Science Report #96/591/ 11.

[14] Pasalkar N. B., Joshi C. V., Tasgaonkar M.: *Script to speech conversion for Marathi language*. TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region, 15-17 Oct. 2003, Vol. 4, pp. 1262 – 1266.

[15] Pawlik P., Mikrut Zb., Bubliński Zb.: *System Wizyjnej Analizy Ankiet*. Automatyka, t. 8. z. 3. Uczelniane Wydawnictwa Naukowo-Dydaktyczne, Kraków 2004.

[16] Shi Z., Setlur S., Govindaraju V.: *Digital Image Enhancement Using Normalization Techniques and their Application to Palmleaf Manuscripts*. SPIE 2005.

[17] Steinherz T., Rivlin E., Intrator N.: *Off-line cursive script word recognition: A survey*. International Journal of Document Analysis 533 and Recognition, 2(2):90 – 110, 1999.

[18] Suzuki M., Tamari F., Fukuda R., Uchida S., Kanahori T.: *INFTY – an integrated OCR system for mathematical documents*. Proceedings of ACM Symposium on Document Engineering 2003, Grenoble, Ed. C.Vanoirbeek, C.Roisin, E. Munson, 2003, pp. 95 – 104.

[19] Tanghva K., Borsack J., Condit A.: *Evaluation of model-based retrieval effectiveness with OCR text*. ACM Transactions on Information Systems, 14(1): 64 – 93, January 1996.

[20] Thillou C., Gosselin B.: *Robust thresholding based on wavelets and thinning algorithms for degraded camera images*. Proceedings of ACIVS 2004, (2004).