Systemy optycznego rozpoznawania znaków pisma

Mirosław Gajer

Podano podstawowe informacje o aktualnych kierunkach badawczych w zakresie systemów optycznego rozpoznawania znaków pisma (OCR). Omówiono zadania realizowane przez tego typu systemy na tle innych systemów przetwarzania języka naturalnego przez komputer. Omówiono w skrócie historię powstania i rozwoju systemów OCR oraz zwrócono szczególną uwagę na problemy związane z rozpoznawanie znaków pisma wybranych języków orientalnych. Opisano sposób realizacji przykładowego algorytmu rozpoznawania znaków pisma. Poruszono zagadnienia związane z segmentacją obrazu i ekstrakcją cech. Jako przykład klasyfikatora znaków przedstawiono propozycję zastosowania sztucznej sieci neuronowej typu Kohonena. Zaproponowany algorytm charakteryzuje się dużą uniwersalnością i może zostać wykorzystany do rozpoznawania znaków praktycznie dowolnego rodzaju pisma.

Obecnie można zaobserwować dynamiczny rozwój programów komputerowych przeznaczonych do przetwarzania tekstów zapisanych w językach naturalnych [1]. Wśród głównych kierunków reprezentowanych w dziedzinie komputerowego przetwarzania języka naturalnego można wymienić, między innymi, systemy syntezy mowy na podstawie zapisu tekstowego języka. Systemy te osiągnęły już dość wysoką jakość, pozwalającą na całkowite zrozumienie wypowiedzi generowanych przez komputer. Ponadto trwają intensywne prace badawcze nad systemami rozpoznawania mowy [3]. Niestety, w ich przypadku jakość działania pozostawia jeszcze bardzo wiele do życzenia. Innym ważnym kierunkiem badań jest przekład komputerowy, w przypadku którego w ostatnich latach poczyniono znaczne postępy, choć droga do całkowitego zastąpienia człowieka-tłumacza komputerem jest jeszcze bardzo daleka (najlepsze wyniki osiągane są w przypadku automatycznego tłumaczenia tekstów o ściśle sprecyzowanej tematyce) [2]. Kierunkiem badań spajającym wymienione technologie przetwarzania języka naturalnego są tzw. głosowe systemy dialogowe (*spoken dialog systems*), w przypadku których komputer ma za zadanie rozpoznanie skierowanej do niego wypowiedzi, dokonanie jej interpretacji, a następnie wygenerowanie na jej podstawie odpowiedzi, którą użytkownik może odsłuchać dzięki sprawnie działającemu modułowi syntezy mowy [10]. Być może w przyszłości właśnie za pośrednictwem głosowych systemów dialogowych będzie można dokonywać np. rezerwacji biletów lotniczych, miejsc w hotelach, realizować operacje na rynkach finansowych oraz uzyskiwać dostęp do informacji turystycznej, przy czym oczywiście tego typu serwisy mogą świadczyć usługi w wielu różnych językach [13, 14]. Warto zauważyć, że pomimo niezwykle dynamicznego rozwoju technik informatycznych teksty zapisane w językach naturalnych nie zawsze występują już w gotowej postaci elektronicznej, przeznaczonej do bezpośredniego przetwarzania przez komputer. Znaczna część tekstów (zwłaszcza starszych) istnieje wciąż jedynie w tradycyjnej wersji papierowej. Ręczna zamiana tych tekstów na postać elektroniczną wymagałaby wręcz niewyobrażalnych nakładów pracy i środków finansowych. Z tego powodu ważnym obszarem badawczym jest rozwijanie komputerowych technik wizyjnych przeznaczonych do optycznego odczytu tekstu i jego automatycznej konwersji do postaci elektronicznej OCR (*Optical Character Recognition*).

dr inż. Mirosław Gajer
– Katedra Automatyki AGH

Rys historyczny rozwoju systemów OCR

Historia systemów optycznego rozpoznawania znaków pisma rozpoczęła się w 1929 r. od przyznania przez Niemiecki Urząd Patentowy austriackiemu inżynierowi G. Tauschekowi patentu na urządzenie mechaniczne przeznaczone do rozpoznawania znaków pisma. Zasada działania rozważanego urządzenia polegała na kolejnym umieszczaniu odpowiednich szablonów przed fotodetektorem. W przypadku zgodności kształtu rozpoznawanej litery z postacią szablonu detektor rejestrował maksymalną moc padającej na niego wiązki światła. Z kolei w 1953 r. Amerykanin D. Shepard opatentował zbudowany przez siebie system GIZMO, który następnie stał się pierwszym ko-

mercyjnym systemem OCR wykorzystywanym przez korporację Readers Digest. System ten opierał się na optycznej technice analizy obrazów i dopuszczał różne rodzaje czcionek i właściwie dowolne rozmieszczenie znaków w ramach skanowanego pola. Z kolei w roku 1965 system OCR zaczęła powszechnie stosować Poczta Amerykańska, a następnie brytyjska Royal Mail. Obecnie systemy odczytujące automatycznie teksty drukowane zapisane alfabetem łacińskim są dość powszechnie wykorzystywane w instytucjach rządowych i jednostkach administracyjnych, ponieważ przyczyniają się do przyśpieszenia procesu obiegu dokumentów oraz pozwalają na szybki dostęp do materiałów drukowanych, umożliwiając tym samym ich modyfikację i dalsze przetwarzanie.

Wyzwania stojące przed projektantami systemów OCR

Obecnie realizacja komputerowego rozpoznawania drukowanych znaków pisma łacińskiego nie stanowi już większego problemu, liczba rozpoznanych prawidłowo znaków drukowanych w większości przypadków osiąga poziom przynajmniej 98 %. Ponadto zastosowanie metod słownikowych, polegających na tym, że w przypadku błędnego rozpoznania jakiegoś znaku w słowniku poszukiwany jest wyraz najbardziej podobny do wyrazu błędnie odczytanego, pozwala na dalszą, znaczącą redukcję stopy błędów. Jednak wciąż dużym wyzwaniem dla badaczy pracujących w dziedzinie OCR jest realizacja komputerowego rozpoznawania znaków pisma odręcznego [7]. Niestety w tym przypadku uzyskiwane rezultaty już nie są tak wysokiej jakości i w bardzo wysokim stopniu zależą od stopnia staranności i czytelności pisma odręcznego. Trudno się zresztą temu specjalnie dziwić, ponieważ często napisy wykonane odręcznie są tak nieczytelne, że niekiedy nawet ich autor ma po jakimś czasie problemy z ich odszyfrowaniem. Z problemem tym borykają się na co dzień np. aptekarze, którzy nieustannie muszą domyślać się, co też lekarz zapisał na danej recepcie. W świetle tego trudno wymagać, aby maszyna w dziedzinie rozpoznawania pisma odręcznego miała być doskonalsza od człowieka. Jednak badania w dziedzinie OCR nie są prowadzone wyłącznie nad rozpoznawaniem pisma odręcznego. Wystarczy uświadomić sobie fakt, że zdecydowana większość mieszkańców naszego globu posługuje się na co dzień innym systemem zapisu niż alfabet łaciński, co może do pewnego stopnia wydawać się zaskakujące z naszej perspektywy. Jednak praktycznie cały obszar cywilizacji islamskiej (ponad 1 mld ludzi) bazuje na alfabecie arabskim (wyjątek stanowi Turcja, Malezja i Indonezja, gdzie obecnie stosowany jest alfabet łaciński). Podobnie cały obszar Półwyspu Indyjskiego wraz ze znaczną częścią Indochin (Birma, Tajlandia, Kambodża i Laos) używa różnych alfabetów wywodzących się ze starożytnych Indii (pism tych używa ponad 1,5 mld ludzi). Z kolei na tradycyjnym obszarze wpływów kultury chińskiej (Chiny, Korea i Japonia) teksty zapisuje się za pomocą wywodzących się ze starożytnych Chin ideogramów, będących w istocie formą pisma obrazkowego (liczba użytkowników tego rodzaju pisma przekracza 1,5 mld ludzi). Ponadto nic nie wskazuje na to, aby tradycyjne systemy zapisu tekstów w krajach o wysoko rozwiniętych i bardzo dawnych kulturach miały zostać wyparte przez alfabet łaciński. Jednocześnie biorąc pod uwagę niezwykle szybkie tempo rozwoju gospodarczego krajów takich jak Chiny i Indie, trzeba uznać, że opracowanie sprawnie funkcjonujących systemów OCR dla różnych egzotycznych form zapisu tekstów jest pilnym i ważnym zadaniem.

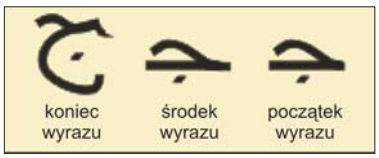
Realizacja OCR dla języków orientalnych

Okazuje się, że w przypadku innych form zapisu tekstów niż alfabet łaciński, realizacja optycznego systemu rozpoznawania znaków dla języków orientalnych staje się o wiele bardziej skomplikowanym zadaniem, nad rozwiązaniem którego pracuje wiele czołowych ośrodków badawczych świata [15]. W dziedzinie tej ukazuje się również bardzo dużo publikacji w specjalistycznych czasopismach poświęconych technikom rozpoznawania wzorców i maszynowej inteligencji [6]. W przypadku rozpoznawania znaków pisma służących do zapisu tekstów języków orientalnych, problem polega albo na samej postaci pisma, albo na mnogości różnego typu znaków, które należy za pomocą systemu OCR prawidłowo zidentyfikować. Przykładem pierwszego rodzaju trudności są wszelkie teksty zapisane za pomocą pisma wywodzącego się z klasycznego języka arabskiego [7]. Obecnie tym systemem pisma zapisywane są nie tylko teksty arabskie, ale również perskie (dodano 4 specjalne znaki dla głosek nie występujących w języku arabskim [11]) oraz między innymi teksty w językach urdu i hausa (język ten bywa również opcjonalnie zapisywany alfabetem łacińskim). Wszelkie systemy zapisu wywodzące się z alfabetu arabskiego są bardzo trudne do komputerowego rozpoznawania [9]. Przyczyna tego faktu tkwi w samej postaci pisma, które ma charakter kursywny, co oznacza, że jego forma drukowana nie różni się praktycznie od zapisów odręcznych. Przykłady napisów w językach arabskim, perskim i urdu zamieszczono na rys. 1.

Problem z realizacją sprawnie działających systemów rozpoznawania znaków wszelkich rodzajów form za-

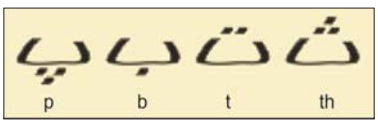
Rys. 1. Przykłady napisów w językach: arabskim, perskim i urdu

pisu opartych na alfabecie arabskim polega głównie na dokonaniu automatycznej segmentacji napisu na obszary reprezentujące poszczególne znaki. Ponadto większość ze znaków alfabetu arabskiego przyjmuje odmienne formy graficzne w sytuacji, gdy występuje na początku, w środku bądź na końcu wyrazu. Zjawisko to zostało zobrazowane na rys. 2.



Rys. 2. Różne formy graficzne znaku arabskiego w początkowej, środkowej i końcowej pozycji wyrazu

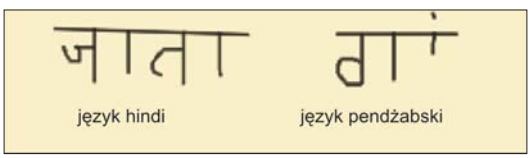
Ponadto wiele znaków alfabetu arabskiego jest do siebie bardzo podobnych, przez co tworzą szeregi znaków, które różnią się jedynie liczbą kropek stawianych nad znakiem lub pod nim (niekiedy również w środku znaku), czego przykład został zamieszczony na rys. 3.



Rys. 3. Podobieństwo form graficznych wybranych znaków alfabetu arabskiego

Trzeba jeszcze pamiętać, że w przypadku pisma arabskiego (jak również i innych pism semickich), zapisowi podlegają jedynie spółgłoski i samogłoski długie. Samogłoski krótkie nie są w zasadzie w ogóle odnotowywane w piśmie, co sprawia, że często dany wyraz staje się wieloznaczny, ponieważ może być odczytany na co najmniej dwa różne sposoby [6]. Uwaga powyższa dotyczy zwłaszcza języka perskiego, którego użytkownicy zaadoptowali alfabet arabski, nie zważając na fakt, że alfabet ten nie jest zbyt dobrym sposobem na oddanie fonetyki jakiegokolwiek języka indoeuropejskiego, co rozważaną wieloznaczność form pisanych dodatkowo potęguje [11]. W sytuacjach wątpliwych stosuje się niekiedy dodatkową wokalizację, polegająca na użyciu specjalnych znaków diakrytycznych na oddanie wymowy również samogłosek krótkich. Automatyczne wyodrębnienie takich znaków z napisu i ich prawidłowa interpretacja stanowią bardzo duże wyzwanie dla twórców systemów OCR dla alfabetu arabskiego.

W przypadku rozpoznawania znaków należących do różnych alfabetów będących w użyciu na terytorium Indii, sytuacja jest zdecydowanie prostsza niż w przypadku realizacji systemu OCR dla znaków alfabetu arabskiego. Na rys. 4. zamieszczono przykład napisów



Rys. 4. Przykłady napisów w językach hindi i pendżabskim

sporządzonych pismem devanagari (język hindi) oraz pismem gurmukhi (język pendżabski).

Cechą charakterystyczną rozważanych alfabetów jest pozioma kreska stanowiąca podstawę zapisu wyrazu, na której są jak gdyby zawieszone poszczególne znaki, co znacznie ułatwia segmentacje napisu. Każdy ze znaków reprezentuje jedną sylabę otwartą, zakończoną samogłoską "a" (np. sa, ma, ba itp.) W przypadku gdy sylaba kończy się na inną samogłoskę, dodawane są specjalne znaki diakrytyczne, które system OCR musi oczywiście prawidłowo rozpoznać.

Inny rodzaj trudności w opracowaniu systemu OCR występuje w przypadku podjęcia próby optycznego rozpoznawania znaków tekstów sporządzonych za pomocą ideogramów chińskich [4]. Służą one do zapisu nie tylko tekstów w różnych językach chińskich, ale również zapisywany jest nimi język japoński. Tutaj zasadnicza trudność nie polega na segmentacji tekstu, która jest znacznie ułatwiona z uwagi na występowanie izolowanych znaków, ale przyczyna jej tkwi w mnogości różnorodnych form znaków [8]. Na przykład podstawowy zbiór ideogramów wykorzystywanych do zapisu języka japońskiego liczy aż 1947 znaków, a jest on i tak dalece nie kompletny. W przypadku realizacji systemu OCR dla tego rodzaju pisma, krytycznym komponentem systemu, od którego zależy w głównej mierze poprawność jego funkcjonowania, jest klasyfikator, który musi bezbłędnie rozpoznawać kilka tysięcy klas obiektów, co z oczywistych względów nie jest łatwym zadaniem.

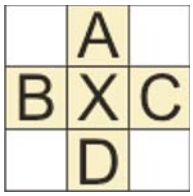
Przykładowy algorytm optycznego rozpoznawania znaków pisma

Opracowany przez autora algorytm optycznego rozpoznawania znaków pisma jest algorytmem uniwersalnym, w tym sensie, że możliwe jest jego zastosowanie do rozpoznawania wszelkiej postaci znaków pisma, jak również i innych wzorców graficznych. W rozważanym algorytmie można wyróżnić fazy związane z komputerowym przetwarzaniem obrazów, ekstrakcją cech oraz klasyfikacją wzorców.

Na wstępie pozyskany ze skanera obraz poddawany jest procesowi binaryzacji, w którym każdemu z pikseli przypisywany jest kolor czarny (reprezentowany binarnie przez zero) lub biały (reprezentowany binarnie przez jedynkę). Celem binaryzacji jest wyodrębnienie obrazu znaku, który powinien otrzymać kolor czarny, z jego tła, któremu powinien zostać przypisany kolor biały. Jednak proces binaryzacji obrazu nigdy nie jest doskonały i zawsze na białym tle obrazu znajdzie się pewna liczba czarnych pikseli, które muszą następnie zostać usunięte w procesie filtracji dolnoprzepustowej obrazu. Powodem tego zjawiska są trudności z optymalnym doborem wartości progu, powyżej którego piksele stają się białe. Próg binaryzacji jest zwykle ustalany automatycznie na podstawie analizy histogramu obrazu.

Jak już wspomniano, po binaryzacji obrazu konieczne jest jeszcze odfiltrowanie ewentualnych zakłóceń i szu-

mów. Można w tym celu wykorzystać złożenie dwóch filtracji nieliniowych, przy czym najpierw musi zostać użyty filtr minimalny, a następnie należy zastosować filtr maksymalny. Maski filtru minimalnego i maksymalnego mają identyczną postać, która została pokazana na rys. 5 i obejmują swym zasięgiem 5 sąsiednich pikseli.



Rys. 5. Maska filtru minimalnego i maksymalnego

Działanie filtru minimalnego polega na tym, że jeśli piksel znajdujący się w polu X ma kolor czarny, wówczas pozostanie on czarny jedynie w przypadku, gdy sąsiednie piksele (znajdujące się w polach A, B, C i D) również wszystkie mają kolor czarny. W przypadku przeciwnym pikselowi znajdującemu się w polu X przypisany zostanie kolor biały. Takie postępowanie prowadzi do usunięcia z obrazu izolowanych czarnych pikseli występujących na białym tle, jak również niewielkich zgrupowań takich pikseli. Efektem ubocznym jest zmniejszenie grubości zawartych w polu obrazu obiektów. W celu przywrócenia ich pierwotnej grubości należy wykonać filtrację maksymalną, w przypadku której pikselowi znajdującemu się w polu X, jeśli jest biały, zostaje przypisany kolor czarny, gdy przynajmniej jeden z jego sąsiadów A, B, C lub D jest koloru czarnego.

Przygotowany w ten sposób obraz może zostać poddany procesowi segmentacji mającej za zadanie wyodrębnienie obszaru obrazu zawierającego dokładnie jeden znak pisma. W przypadku gdy znaki pisma są wzajemnie od siebie odizolowane, nie sprawia to większych trudności, bowiem wystarczy, przeglądając kolejne wiersze i kolumny obrazu, odnaleźć górną, dolną, prawą i lewą krawędź obszaru znaku. Tak wyodrębniony znak zostaje wpisany w prostokąt, którego szerokość równa jest odległości krawędzi prawej od lewej, a wysokość równa jest odległości krawędzi górnej od dolnej. W kroku kolejnym rozważany prostokąt zostaje przeskalowany tak, aby przyjął kształt kwadratu. Operacja ta wykonywana jest w ten sposób, że krótszy bok prostokąta zostaje przemnożony przez stosunek długości boku dłuższego do krótszego. W opisanym procesie obraz znaku zostaje zdeformowany, tzn. jest rozciągnięty w pionie lub w poziomie.

Na tym etapie można przystąpić już do ekstrakcji cech obrazu. Proces ten polega na tym, że kwadrat, w który wpisany został obraz znaku, zostaje podzielony na np. 25 kwadratowych sektorów o identycznych wymiarach, co pokazano na rys. 6.

Następnie w ramach każdego z sektorów zliczane są czarne piksele obrazu. W zależności od kształtu rozpoznawanego znaku pisma w pewnych sektorach piksele czarnych będzie stosunkowo dużo, a w innych sektorach będzie ich mniej bądź mogą nie występować tam wcale. Uzyskany podczas zliczania czarnych pikseli 25 elementowy wektor liczb stanowić może podstawę do klasyfikacji znaków. W celu uniezależnienia procesu rozpoznawania znaków od ich wymiarów bezwzględnych należy każdą ze składowych wektora podzielić przez liczbę wszystkich pikseli zawartych w obrębie kwadratu, w który wpisany został dany znak. Znormalizowane w ten sposób wektory cech stanowić już mogą dane wejściowe dla modułu automatycznej klasyfikacji znaków pisma.

Rys. 6. Sposób podział obrazu znaku pisma na 25 równych sektorów

Klasyfikator neuronowy

Spośród wielu możliwych systemów automatycznej klasyfikacji wzorców, na który warto zwrócić uwagę, jest klasyfikator oparty na sztucznej sieci neuronowej typu Kohonena [5]. Sieć Kohonena jest sztuczną siecią neuronową zbudowaną z tylko jednej warstwy liniowych neuronów. Jest to sieć, której trening realizowany jest bez nauczyciela, natomiast istotną rolę odgrywa w nim proces konkurencji, bowiem procesowi treningu poddawany jest jedynie ten neuron, którego reakcja na zadany wzorzec była największa. Zatem celem treningu jest jeszcze dodatkowe wzmocnienie tej reakcji. Wektory cech wyznaczone na podstawie komputerowej analizy obrazów znaków przed podaniem na wejście sieci Kohonena muszą zostać poddane normalizacji tak, aby ich długość wynosiła dokładnie jeden. Wartość wektora X po normalizacji obliczana jest ze wzoru (1), gdzie indeks i przebiega po wszystkich składowych wektora X.

$$\bar{X}_i = \frac{X_i}{\sqrt{\sum_{j=1}^{N} X_j^2}} \tag{1}$$

Z kolei sam proces modyfikacji wartości współczynników wagowych neuronów przebiega zgodnie ze wzorem (2), gdzie podano, jak wartości współczynników wagowych, w kolejnym kroku treningu, zależą od wartości w kroku poprzednim oraz od wartości sygnału podanego na wejście sieci. Ponadto współczynnik η determinuje szybkość procesu uczenia sieci.

$$w_{n+1} = w_n + \eta(x_n - w_n) \tag{2}$$

W wyniku procesu uczenia wagi neuronów sieci Kohonena zostają dostrojone w ten sposób, że sieć zaczyna grupować wektory do siebie podobne. W rezultacie tego poszczególne neurony sieci Kohonena uczą się rozpoznawać różne rodzaje wzorców. Liczba neuronów sieci Kohonena musi być jednak ponad dwukrotnie większa od liczby rozpoznawanych wzorców. Dzieje się tak dlatego, że proces treningu jest procesem stochastycznym, w wyniku którego kilka neuronów może nauczyć się rozpoznawać ten sam typ wzorca, a pewne neurony mogą nie nauczyć się rozpoznawać żadnego wzorca. Dlatego po zakończeniu procesu treningu konieczne jest przeprowadzenie walidacji sieci w celu ustalenia, które neurony nauczyły się rozpoznawać jakiego typu wzorce. Jeśli rozpoznawanie każdego ze wzorców zo-

stało nauczone przez co najmniej jeden z neuronów sieci, wówczas tak wytrenowana sieć może zostać wykorzystana do rozpoznawania znaków. Jeśli warunek ten nie jest spełniony, wówczas proces treningu należy powtarzać aż do skutku.

Zakończenie

Automatyczne rozpoznawanie znaków pisma z zastosowaniem technik analizy obrazów stanowi ważny dział badań związanych z komputerowym przetwarzaniem języka naturalnego. Technika optycznego rozpoznawania znaków pisma pozwala na przekształcenie dokumentów istniejących w postaci drukowanej do równoważnej im postaci elektronicznej. Etap ten jest o tyle istotny, że umożliwia następnie zastosowanie innych technik przetwarzania tekstu wypracowanych przez lingwistykę komputerową. Na przykład dokumenty po skanowaniu mogą być automatycznie klasyfikowane ze względu na tematykę ich treści. Również w zbiorze takich dokumentów mogą być automatycznie poszukiwane informacje, których użytkownik potrzebuje. Ponadto dzięki zastosowaniu technik translacji automatycznej rozważane dokumenty mogą być tłumaczone przez komputer również na inne języki [14]. Obecnie dużym wyzwaniem jest opracowanie efektywnie działających systemów optycznego rozpoznawania znaków pisma dla wielu języków orientalnych, gdzie napotykane są dodatkowe trudności, które nie są spotykane w przypadku rozpoznawania znaków alfabetu łacińskiego. Fakt ten związany jest zarówno z mnogością różnych form znaków, jak i z ich specyficznym kształtem. Zastosowanie nowoczesnych technik rozpoznawania bazujących na metodach sztucznej inteligencji, takich jak np. sztuczne sieci neuronowe, pozwala jednak żywić pewną nadzieję na skuteczne przezwyciężenie tych problemów w przyszłości.

Bibliografia

1. Allen J. F.: *Natural language understanding*, The Benjamin/Cummings Publishing Company, New York, 1995.
2. Arnold D., Balkan L., Meijer S., Humphreys R. L., Sadler L.: *Machine translation: an introductory guide*, NCC Blackwell, London, 1994.
3. Axelrod S., Goel V., Gopinath R. A., Olsen P. A., Visweswariah K.: *Subspace constrained gaussian mixture models for speech recognition*, IEEE Transactions on Speech and Audio Processing, vol. 13, no. 6, 2005, ss. 1144–1160.
4. Baker H., Ho P. K.: *Cantonese*, Hodder & Stoughton, London, 2002.
5. Gajer M.: *Zastosowanie wybranych typów sieci neuronowych do rozpoznawania obrazów*, Pomiary Automatyka Robotyka, 1/2001, ss. 5–10.
6. Guidère M.: *Toward corpus-based machine translation for standard Arabic*, Translation Journal, vol. 6, no. 1, 2002.
7. Hamid A., Haraty R.: *A neuro-heuristic approach for segmentating hadwritten Arabic text*, ACS/IEEE International Conference on Computer Systems and Applications, 2001, ss. 110–113.
8. Le S., Youbing J., Lin D., Yufang S.: *Word alignment of English-Chinese bilingual corpus based on chunks*, Chinese Academy of Sciences, Chinese Information Processing Center, Institute of Software, Peking, 2000.
9. Matthews D., Dalvi M. K.: *Urdu*, Hodder & Stoughton, London, 2002.
10. Ney H., Niessen S., Och F. J., Sawaf H., Tillmann C., Vogel S.: *Algorithms for statistical translation of spoken language*, IEEE Transactions on Speech and Audio Processing, vol. 8, 2000, ss. 24–35.
11. Rahnama K. P.: *Język perski*, Wydawnictwo Akademickie DIALOG, Warszawa, 1999.
12. Scurfield E.: *Chinese*, Hodder & Stoughton, London, 2002.
13. Souvingnier B., Keller A., Rueber B., Schramm H., Seide F.: *The thoughtful elephant: strategies for spoken dialog systems*, IEEE Transactions on Speech and Audio Processing, vol. 8, ss. 47–67.
14. Waibel A., Geatner P., Tomokiyo L. M., Schultz T., Woszczyna M.: *Multilinguality in speech and spoken language systems*, Proceedings of the IEEE, vol. 88, 2000, ss. 1297–1313.
15. Xu M., Dong J.: *Generating new styles of Chinese strokes based on statistical model*, Task Quarterly, vol. 11, no 1–2, ss. 129-136.