Skanowanie dokumentów i techniki rozpoznawania znaków

Tomasz Lewicki

WWSIS, Wrocław

czerwiec 2007

Skanowanie

Skanowanie to proces przekształcania nieruchomego dwywymiarowego obrazu lub obiektu trójwymiarowego do postaci cyfrowej przy pomocy urządzenia zwanego ogólnie **skanerem optycznym**.

Obraz w postaci cyfrowej

Obrazy cyfrowe (zdigitalizowane) mogą być tworzone **bezpośrednio** przy pomocy cyfrowych urządzeń rejestracji obrazu, np. cyfrowych aparatów fotograficznych lub cyfrowych kamer video lub **pośrednio**, np. skanując obrazy analogowe za pomocą skanerów optycznych.

Obraz cyfrowy tworzy się próbkując obraz analogowy z określoną rozdzielczością. Przyjmuje się założenie, że obraz analogowy tworzy siatkę punktów (pikseli). W trakcie próbkowania otrzymuje się zbiór punktów siatki opisanych dla każdego piksela jego wartością tonalną (biały, czarny, odcień szarości lub kolor) w zapisie bitowym.

Z obrazu cyfrowego można — po odpowiednich przekształceniach — odtworzyć obraz analogowy, np. na drukarce lub ploterze.

Rodzaje skanerów optycznych

Przykłady klasyfikacji

Ze względu na sposób użycia:

- ręczne
- płaskie, inaczej stołowe (*flatbed*)
- bębnowe

Ze względu na przeznaczenie:

- skaner klisz i slajdów
- skaner dokumentów
- skaner kodów kreskowych

Parametry skanerów

„Jakość” skanera opisuje kilka podstawowych parametrów:

- głębia kolorów (*color depth*)
- rozdzielczość optyczna (*optical resolution*)
- zakres gęstości (*density range*)

Głębia kolorów

Głębią kolorów nazywamy ilość bitów potrzebnych do opisania barwy piksela na obrazie. Wyższe wartości tego parametru wskazują, że możliwe jest oddanie większej ilości barw i odcieni na obrazie. Najmniejsza głębia kolorów wynosi 1 bit i umożliwia zapisanie każdego piksela w dwóch kolorach: czarnym lub białym. Kolejne wartości głębi kolorów powszechnie stosowane w grafice komputerowej to:

- 8 bitów: $2^8 = 256$; zarówno w skali szarości, jak i w ograniczonej palecie kolorów
- 15 bitów: $(2^5)^{\cdot 3} = 2^{15} = 32768$
- 16 bitów: $2^{5+6+5} = 2^{16} = 65536$; tzw. *Hi-color*
- 24 bity: $(2^8)^{\cdot 3} = 2^{24} = 16777216$; tzw. *True-color*
- 32 bity: $(2^8)^4 = 2^{32}$

Ta ostatnia wartość to w rzeczywistości *Hi-color* z dodatkowymi ośmioma bitami używanymi najczęściej do zapisania informacji o przezroczystości.

Rozdzielczość obrazu

Rozdzielczość optyczna i interpolowana

Rozdzielczość obrazu cyfrowego to łączna liczba pikseli tworzących obraz. Rozdzielczość należy rozpatrywać w odniesieniu do konkretnego urządzenia, np. ekranu monitora, na którym wyświetlany jest obraz.

W odniesieniu do skanerów wyróżnia się dwie rozdzielczości: optyczną oraz interpolowaną. **Rozdzielczość optyczna** zwana również **fizyczną** to rzeczywista ilość elementów światłoczułych (CCD lub CIS) skanera podzielona przez maksymalną szerokość obiektu, jaką skaner jest w stanie zeskanować. **Rozdzielczość interpolowana** to programowe zwiększenie rozdzielczości optycznej. Interpolacja w przypadku skanera polega na wstawianiu pomiędzy dwa rzeczywiste punkty zeskanowanego obrazu dodatkowych pikseli, których wartości (barwa i jasność) są wyliczane na podstawie pewnego modelu z analogicznych wartości sąsiednich pikseli.

Zakres gęstości

Zakres gęstości, zwany też zakresem dynamiki (*dynamic range*) jest miarą zdolności skanera do rozróżniania barw w naciemniejszych i najjaśniejszych obszarach obrazu. Oznaczany jest dużą literą D. Mówiąc ściślej jest to różnica pomiędzy czułością przy najwyższej (D_{max}; najciemniejsze piksele) i najniższej (D_{min}; najjaśniejsze piksele) gęstości.

Wartości D zawierają się w umownym przedziale od 0 do 4, gdzie 0 oznacza przezroczystą błonę fotograficzną, a 4 — błonę całkowicie zaczernioną. Gęstość jest mierzona w skali logarytmicznej o podstawie 10, czyli gęstość $D = 2$ jest 10-krotnie mniejsza niż $D = 3$. Skanery wysokiej klasy (większość bębnowych, niektóre płaskie) mają wyższe zakresy gęstości oraz wartości D_{min} i D_{max}. Wyższe wartości D mają znaczenie zwłaszcza przy obróbce materiałów transparentnych (np. klisz fotograficznych).

Przechwytywanie obrazu przez skaner
CCD

CCD jest skrótem od *Charge Coupled Device*, czyli **urządzenie o sprzężeniu ładunkowym**. Jest to matryca elementów światłoczułych działających na zasadzie wewnętrznego efektu fotoelektrycznego.
W skanerach płaskich elementy te są umieszczone liniowo (dla każdej z barw składowych RGB osobno) na ruchomej belce przesuwającej się wewnątrz skanera. W rejestracji obrazu pomaga układ luster i soczewek oraz zimna lampa katodowa dająca barwę światła zbliżoną do naturalnej. Taki układ ma stosunkowo dużą głębię ostrości.

W aparatach cyfrowych CCD ma postać macierzy dwuwymiarowej — obraz jest rejestrowany w całości w jednym momencie.

Przechwytywanie obrazu przez skaner — c.d.

CIS

CIS to skrót od *Contact Image Sensor*. Jest to technika rejestracji młodsza niż CCD. Źródłem światła są diody półprzewodnikowe LED, dzięki czemu pobór prądu jest mniejszy niż w przypadku skanerów CCD. Zwierciadło jest zbędnym elementem układu, ponieważ sensory są umieszczone bliżej skanowanego obiektu. Całość jest umieszczona na ruchomej belce, podobnie jak w przypadku skanerów opartych o układy CCD. Skanery z układami CIS mają małą głębię ostrości.

Przechwytywanie obrazu przez skaner — c.d.

PMT

PMT jest skrótem od *Photomultiplier Tube* i oznacza **fotopowielacz**. Urządzenie ma postać lampy próżniowej, zbudowanej z anody i katody oraz co najmniej jednej dynody (zazwyczaj więcej niż jednej), której zadaniem jest powielenie elektronu wybitego z fotokatody przez foton na nią padający.

Fotopowielacze charakteryzują się dużymi czułościami i wzmocnieniami, ale wymagają do działania wysokiego napięcia (od kilkuset do kilku tysięcy woltów) i są drogie, przez co znajdują zastosowanie niemal wyłącznie w profesjonalnych urządzeniach, np. w skanerach bębnowych (po jednym fotopowielaczu dla każdej z barw składowych RGB).

Przygotowanie do druku

Podczas przygotowywania zdigitalizowanego obrazu do druku należy mieć na uwadze fakt, iż rozdzielczość drukarki różni się od rozdzielczości skanera. Pamiętać trzeba również o tym, że drukarka dysponuje tylko czterema kolorami (CMYK), przy pomocy których musi oddać wszystkie kolory widoczne na ekranie (o ile oczywiście mamy do czynienia z obrazem barwnym).

Jeśli obraz ma zbyt niską rozdzielczość, drukarka będzie miała za mało informacji potrzebnych do wiernego oddania detali i barw poszczególnych pikseli. Jest obraz jest w bardzo wysokiej rozdzielczości, przed drukiem należy dokonać tzw. *downsamplingu*, czyli zmniejszenia częstości próbkowania sygnału (tutaj: obrazu).

Rozpoznawanie znaków

OCR

OCR jest skrótem od *Optical Character Recognition* i oznacza **optyczne rozpoznawanie znaków**, potocznie zwane **rozpoznawaniem tekstu**. Jest to technika pozwalająca z obrazu w postaci cyfrowej wyekstrahować tekst, ewentualnie cechy czcionek użytych do jego złożenia (krój, wielkość) i inne elementy dokumentu, np. formatowanie, tabele, formularze.

Rozpoznawanie znaków

ICR

ICR to skrót od *Intelligent Character Recognition*, czyli **inteligentne rozpoznawanie znaków**. Podstawowym zadaniem systemów ICR jest rozpoznanie znaków alfanumerycznych zapisanych odręcznie. Do rozpoznawania używane są mechanizmy sieci neuronowych.

Rozpoznawanie znaków

OMR

OMR oznacza *Optical Mark Recognition* — **optyczne rozpoznawanie znaczników**. Polega na rozpoznawaniu znaków innych niż alfanumeryczne, np. pól wyboru lub kodów kreskowych. Czytniki OMR znacznie ułatwiają analizę dużej ilości zestandaryzowanych formularzy oraz umożliwiają kontrolę poprawności ich wypełnienia.

Rozpoznanie znacznika polega na zmierzeniu ilości światła odbitego lub przechodzącego przez ściśle określony fragment skanowanego dokumentu.

Skuteczność technik rozpoznawania tekstu i znaków

Systemy OCR, ICR i OMR znajdują zastosowanie na różnych polach. Z tego względu ich skuteczność musi być badane według odmiennych kryteriów. Jeśli przyjąć za podstawowe kryterium odsetek prawidłowo odczytanych i przeanalizowanych danych, to najbardziej skuteczne są techniki OMR (nawet 99,9%), mniej skuteczne są systemy OCR, a największym procentem błędnych interpretacji cechują się systemy ICR. Dwa ostatnie systemy (OCR i ICR) również mogą osiągać około 99% skuteczności, ale tylko w bardzo ściśle określonych, niemal „laboratoryjnych" warunkach oraz po ręcznej edycji błędów.

„Sztuczki i kruczki” dla celów OCR

Przygotowując dokumenty papierowe do digitalizacji i rozpoznania tekstu dobrze jest pamiętać o kilku zasadach, które mogą znacząco podnieść skuteczność rozpoznania:

- papier powinien być jasny i — jeśli to możliwe — pozbawiony skaz, zagnieceń, włókien z pulpy papierowej itp. defektów
- dokument należy skanować z rozdzielczością optyczną co najmniej 300 dpi, czasami wyższą
- obraz przeznaczony do OCR powinien być zapisany w odcieniach szarości, najlepiej w formacie nieskompresowanym
- należy przyjrzeć się ustawieniom jasności i kontrastu w programie skanującym, ewentualnie wykonać kilka skanów próbnych i poddać je procesowi OCR
- najwięcej błędów zdarza się przy interpretacji drobnych znaków, np. kropek, przecinków, średników, myślników