

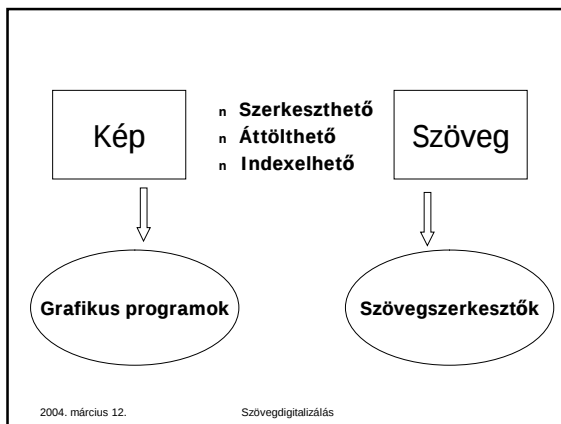


Szövegdigitalizálás, szövegfelismerő programok

- Mi az OCR?
- Karakterfelismerési módszerek
- Az optikai karakterfelismerés menete
- A karakterfelismerés folyamata
- A karakter felismerése
- A Recognita termékcsalád

2004. március 12.

Szövegdigitalizálás



2004. március 12.

Szövegdigitalizálás

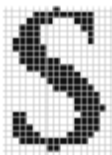
OCR

Optikai karakterfelismerő program.

- n Optikai karakterfelismerés: az a folyamat, melynek segítségével egy nyomtatott oldal képéből kinyerhető a szöveg.
- n Az oldal képe a szöveg – és az esetleges egyéb elemek, mint a cím vagy ábrák – elektronikus lenyomata.
- n „Apró pontok tömege”.
- n A számítógépes program elemzi a képen található képpontok geometriáját.

2004. március 12.

Szövegdigitalizálás



2004. március 12.

Szövegdigitalizálás

Karakterfelismerési módszerek

Mintaillesztés (*Matrix matching*)

- n Minden egyes potenciálisan előforduló *karakter mintája* a szoftverben
- n A minták összehasonlítása a dokumentum karakterképeivel, pontos megfeleltetésig
- n Azonosítás, rögzítés.
- n Gyors, jó, viszonylag olcsó.
- n Hátránya: ha a szoftver nincs felkészítve valamilyen betűtípusra, a felismerés nem működik.

2004. március 12.

Szövegdigitalizálás

Karakterfelismerési módszerek

Kontúr (vagy topológiai) analízis

- n A karaktereket *általános jellemzőik alapján* (görbületek, hurok száma, jellemző pontok helyzete) ismeri fel a szoftver, betűtípustól és stílustól függetlenül.
- n Tanítási opció a szokatlan formájú karakterek felismerésére.
- n A programok nagy része a lapszerkezet felismerésére és elmentésére is alkalmas.

2004. március 12.

Szövegdigitalizálás

Karakterfelismerési módszerek

SAT-technológia

(Self-Assertion Technology)

- n A fenti módszerek ötvözte.
Kiindulópontja: a lapon található karakterek csak korlátozott számú betűtípust tartalmaznak.

2004. március 12.

Szövegdigitalizálás

SAT-technológia

Menete:

- n Karakterazonosítás kontúranalízissel
- n Tudásbázis felépítése
- n Újraelemzés

2004. március 12.

Szövegdigitalizálás

I. Képelőfeldolgozás lépései

1. Szöveg behelyezése
2. Rotáció
3. Ferdeségkorrekció
4. Fölösleges elemek eltávolítása

2004. március 12.

Szövegdigitalizálás

I. Képelőfeldolgozás lépései

5. Automatikus fényerőszabályozás
6. Lapszerkezet azonosítása
7. Zónázás
8. Lapszerkezetelemzés

2004. március 12.

Szövegdigitalizálás

Az optikai karakterfelismerés menete

- n Szöveg beolvasása, előkészítése
- n Felismerés
- n Kódkká alakítás

2004. március 12.

Szövegdigitalizálás

II. A karakterfelismerés folyamata

1. Szegmentálás

- n A karaktert négyzetbe foglalja
- n Körbejárja a betűt és számolja a fordulókat
- n Dönt

2004. március 12.

Szövegdigitalizálás

II. A karakterfelismerés folyamata

2. Jellemzők kigyűjtése (feature extraction)

- n Visszatér a bizonytalan karakterekhez
- n Összehasonlítva dönt tulajdonságvektorral

2004. március 12.

Szövegdigitalizálás

Az ideális tulajdonságvektor

Megkülönböztet:

- n Hasonló alakú karaktereket
- n Azonos karakterek különböző variációit

2004. március 12.

Szövegdigitalizálás

Helyesírás-ellenőrzés

- n Bizonytalan szavak, karakterek megjelölése
- n Szótárak, helyesírásellenőrzők

2004. március 12.

Szövegdigitalizálás

Recognita – Omnipage



SAT + LA

2004. március 12.

Szövegdigitalizálás

- n Ötszintű gyorsaság/pontosság-beállítás
- n Soknyelvűség
- n Táblázatkezelés
- n Tanítás
- n Ellenőrzés
- n Lapkezelő nézet (több oldalas dokumentumok közti gyors eligazodás)
- n Karakterfelismerés 6-72 pontig
- n Szókereső
- n Szűrőkészítés képek kezelése (szövegbe ágyazva is elmenthetők)
- n Közvetlen kapcsolat windowsos szövegszerkesztőkkel és táblázatkezelőkkel
- n Több mint 50 kimeneti formátum

2004. március 12.

Szövegdigitalizálás

A Recognita termékcsalád története

- n 1989 –: SZKI – Recognita Rt.-fejlesztés: Recognita Reader, később: Recognita Plus.
- n 1990-91: leányvállalatok (No., USA)
- n 1993: MAVÁ
- n 1994: OEM + *upgrade* modell
- n 1995: pénzügyi bukás + technológiai fejlesztés

2004. március 12.

Szövegdigitalizálás

A Recognita termékcsalád története

- n 1995-96: független elemzők szerint is világelső a program; ISO 9001 minőségbiztosítási tanúsítvány megszerzése
- n 1996 végén: fúzió az amerikai Caere Co-val. Magyar függetlenség megmarad. Nemzetközi díjak elnyerése
- n 1997-99: nyereség; díjak
- n 2000 március: egyesülés a ScanSoft céggel.

2004. március 12.

Szövegdigitalizálás

A Recognita termékcsalád története

- n 2002: OmniPage Pro 12 Office (114 nyelv)
- n 2004: OmniPage Pro 14 Office (PDF-konverzió)

2004. március 12.

Szövegdigitalizálás