MTA SZTAKI DSD

Department of Distributed Systems

Szkennelt szövegek digitalizálása során keletkező hibák elemzése magyar szövegek esetében

Pataki Máté

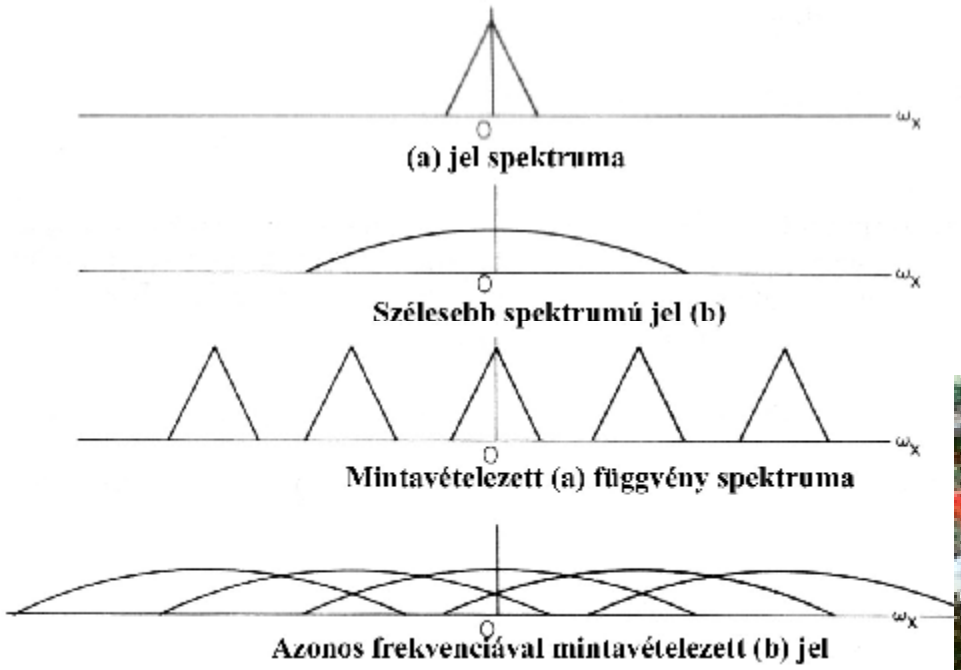
Tóth Zoltán

Tartalomjegyzék

Szöveges dokumentumok digitalizálása

1. Mintavételezés (szkennelés)
2. Kvantálás
3. Előfeldolgozás
4. Karakterfelismerés
5. Szófelismerés, szövegfeldolgozás

Mintavételezés



Ha nem teljesül a Nyquist feltétel, spektrumátfedési hiba lép fel (Moiré effektus)

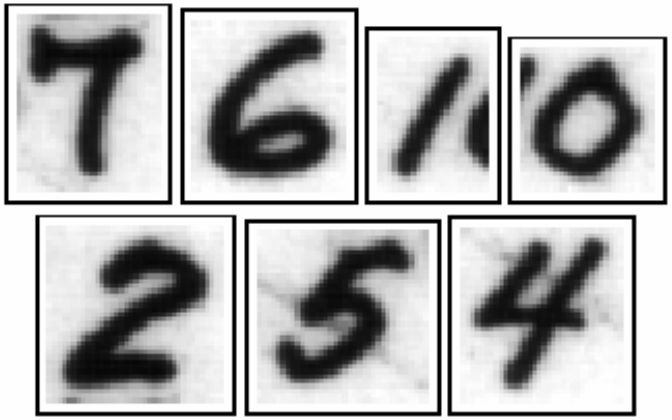
Kvantálás

Szürkeskálás kép 8 bites, 4 bites és 1 bites verziói

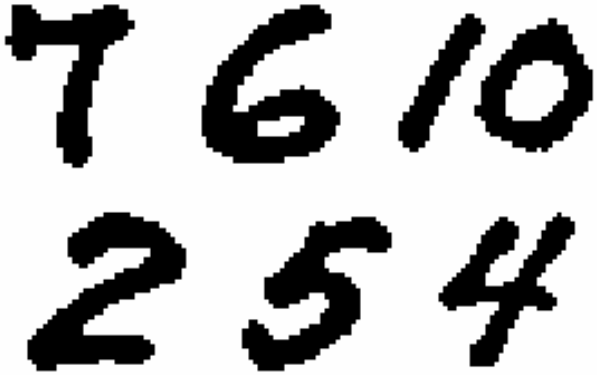
Előfeldolgozás I

- Zajszűrés
- Geometriai torzítás korrekciója
- Előtér háttér szeparáció
- Szegmentáció, szerkezetfelismerés
- Morfológiai képfeldolgozó operátorok alkalmazása
- Képi tulajdonságok kinyerése

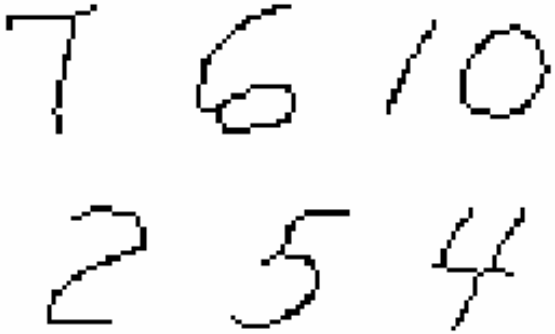
Előfeldolgozás II



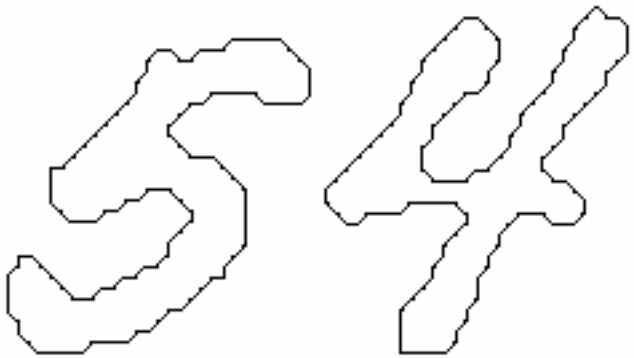
Szegmentálás



Binarizálás

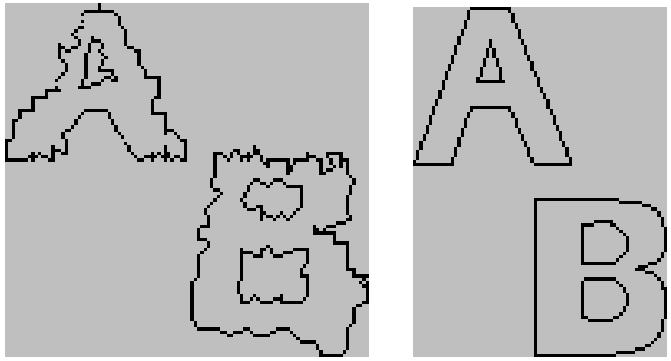


Vázosítás



Kontúrdetekció

Előfeldolgozás III



Poligonillesztés

Konvex befoglaló (és az eredeti objektum különbségének) meghatározása

Karakterfelismerés (OCR)

- **Mintaegyeztetés**
  A mintát a betű képére illesztjük, és megmérjük az egyezés mértékét
- **Tulajdonság alapú**
  A karakterek speciális sajátosságainak, szabályainak vizsgálata
- **Hierarchikus, komplex módszer**

Szófelismerés, szövegfeldolgozás

- Célja, hogy nyelvtani szabályok felhasználásával OCR hibákat szűrjön ki
- További hibákat is bevihet a rendszerbe
- Nyelvfüggő

a (95%)	**b** (90%)	**i** (88%)	**a** (89%)	**k** (96%)
o (83%)	**d** (79%)	**l** (85%)		**h** (62%)
á (80%)		**I** (80%)		
		1 (76%)		

Tesztek - tesztrendszer

- Humán tesztek
- Gépi teszt
  - Dokumentum nyomtatása
  - Mesterséges hibák generása
  - Karakterfelismerés
  - Összehasonlítás
    - Kinyomtatott szöveg
    - Szkennelés eredménye

Mesterséges hibák

Mesterséges hibák

Kávéfoltos szöveg

A karakterfelismerés hibái

- Ékezethibák

  veréb/véreb, alma/álma, hó/hő
- Írásjelek tévesztése (- – — , . ; :)
- Betűcserék (M m, é e)
- Az i betű felismerési problémái (í i I 1)
- Számok és betűk keverése (g 9, J 3, O 0)
- Az o és ö betű felismerési problémái

Department of Distributed Systems

DSD

Leggyakrabban előforduló hibás karaktercserék

Orig	OCR	Count
M	m	124103
-	—	82358
é	e	75882
á	a	71436
-	NULL	55990
		43263
V	v	42109
g	9	40713
,	,	40180
NULL	-	30378
o	õ	21321
ó	o	18301
NULL		15324
í	i	13992
”	"	13975
W	w	11401
–	-	10428
`	-	10251

Orig	OCR	Count
I	i	10130
U	u	10048
Ú	ú	9804
¢	-	8412
D	B	8108
ú	u	7896
J	3	7617
NULL	•	7444
õ	Ó	7438
NULL	'	6744
NULL	.	6531
u	Û	6469
	NULL	6268
Ö	Ö	5831
õ	O	5689
Z	Z	5671
i	L	5627
Í	Í	5574

Orig	OCR	Count
Õ	õ	5488
“	"	5442
-,	NULL	5337
í	l	5270
□	o	5167
£	t	5091
	-	5025
NULL	,	4635
¡	-	4619
e	é	4503
a	á	4248
.	NULL	4198
û	ü	3959
É	E	3913
j	J	3283
,	NULL	3184
o	ó	3112
"	„	3105

Ö és Ő betűk felismerésének problémája

Orig	OCR	Count
o	õ	21321
ó	o	18301
õ	ó	7438
Ö	ö	5831
õ	o	5689
Õ	õ	5488
o	ó	3112
ó	Ó	1361
o	ö	1213

Leggyakoribb szóhibák

Szó	Eredeti	OCR	Különbség
a	5762018	5716296	45722
és	1319840	1281757	38083
s	38423	5498	32925
hogy	1171612	1153779	17833
de	479068	461786	17282
az	1980365	1965373	14992
Úgy	34743	20643	14100
nem	1091302	1080086	11216
még	289016	278288	10728
egy	705763	695371	10392
Így	24386	14514	9872
már	303129	293412	9717
Ő	19078	11164	7914
is	762166	754575	7591
És	123332	117331	6001
jó	93441	88244	5197
mag	768	5906	-5138
4	14706	20248	-5542
d	18261	23842	-5581

Szó	Eredeti	OCR	Különbség
d	18261	23842	-5581
e9y	2	5606	-5604
11	4947	10760	-5813
gy	1580	7675	-6095
nt	2381	8567	-6186
ban	12811	19055	-6244
z	1741	8305	-6564
val	2454	9396	-6942
mar	615	7688	-7073
ho9y	4	7194	-7190
st	1661	9171	-7510
ao	27	7575	-7548
lt	269	7825	-7556
úgy	137825	145848	-8023
ra	4158	12658	-8500
p	7252	16220	-8968
c	10013	20989	-10976
rt	2130	13652	-11522
ny	478	12991	-12513

Szavak ragozott alakjainak száma

Szó	Ragozott alakok száma
láb	173
hív	169
fog	162
él	157
vár	157
ember	156
szív	156
áll	155
szó	151
kéz	150
ér	146
barát	145
úr	145

Szó	Ragozott alakok száma
tesz	140
mond	139
beszél	139
talál	137
fej	137
város	137
tart	137
ruha	135
út	134
hall	132
apa	129
néz	129
lát	129

Szó	Ragozott alakok száma
álom	128
nyom	128
dolog	128
ad	128
hajó	126
ház	126
hely	126
fal	125
maga	123
olvas	122
ismer	121
ír	120

Szavak ragozott alakjainak száma

1. lábak
2. lábam
3. lábadra
4. lábamat
5. lábáig
6. lábánál
7. lábacskáját
8. lábammal
9. lábukkal
10. lábakra

Láb	
1	lábak
2	lábam
3	lábadra
4	lábamat
5	lábáig
6	lábánál
7	lábacskáját
8	lábammal
9	lábukkal
10	lábakra
11	lábtól
12	lábunkon
13	lábatokig
14	lábaidhoz
15	lábaknál
16	lábatokat
17	lábakig
18	lábára
19	lábbeli
20	lábukig
21	lábukra
22	lábadon
23	lábaikon

24	lábaid
25	lábaimról
26	lábadnak
27	lábért
28	lábacskája
29	lábadhoz
30	lábairól
31	lábacskának
32	lábaiba
33	lábbelimen
34	lábbelivel
35	lábaimra
36	lábuknak
37	lábán
38	lábaidat
39	lábait
40	lábai
41	lábuk
42	lábukon
43	lábaiddal
44	lábamig
45	lábacskáikkal
46	lábától
47	lábaikkal

48	lábamnak
49	lábunk
50	lábakhoz
51	lábacskáit
52	lábacskájával
53	lábacskám
54	lábaidnál
55	lábé
56	lábukból
57	lábaimból
58	lábainál
59	lába
60	lábával
61	lábáról
62	lábaival
63	lábon
64	lábból
65	lábaira
66	lábbelije
67	lábadról
68	lábakat
69	lábamon
70	lábaknak
71	lábába

Köszönöm a figyelmüket!

WEB: *http://dsd.sztaki.hu*

Email: *Mate.Pataki@sztaki.hu*

2007. április 12.