UNIUNEA EUROPEANĂ

GUVERNUL ROMÂNIEI

Instrumente Structurale
2007-2013

Platformă de e-learning și curriculă e-content pentru învățământul superior tehnic

AEACD

18. Detectia elementelor constitutive ale documentelor: Caracterele

Elemente constitutive

- Reprezinta elementele de interes din document
- Fiecare document este impartit in pagini, iar acestea care sunt ori toate la fel, ori pot face parte din componente diferite ale documentului
- Ex: Paginile din introducere pot fi numerotate folosind caractere romane, iar celelalte vor fi numerotate folosind caractere arabe

Elemente Constitutive. Clasificare

- Aceste diferente intre pagini apar datorita elementelor continute de acestea
- Detectia elementelor constitutive poate duce, de exemplu, la detectia paragrafelor
- Aceste se clasifica in:
  - Caractere
  - Spatii albe
  - Separatori
  - Linii
  - Cadre si Tabele

Caracterele

- Reprezinta elementele de baza pentru sistemele de tip OCR
- In documente gasim o larga varietate de tipuri de caractere ce pot fi clasificate dupa:
  - font
  - marime
  - italic/bold
  - caractere speciale
- Prezinta informatie esentiala din punct de vedere al continutului
- Pot ajuta la recunoasterea tipului de paragraf:
  - Titlu
  - Subtitlu
  - Nota de subsol
  - Etc
- Exemplu: In cazul documentelor deteriorate informatii precum numerele paginilor si/sau numele documentului si al autorilor pot sa fie necesare

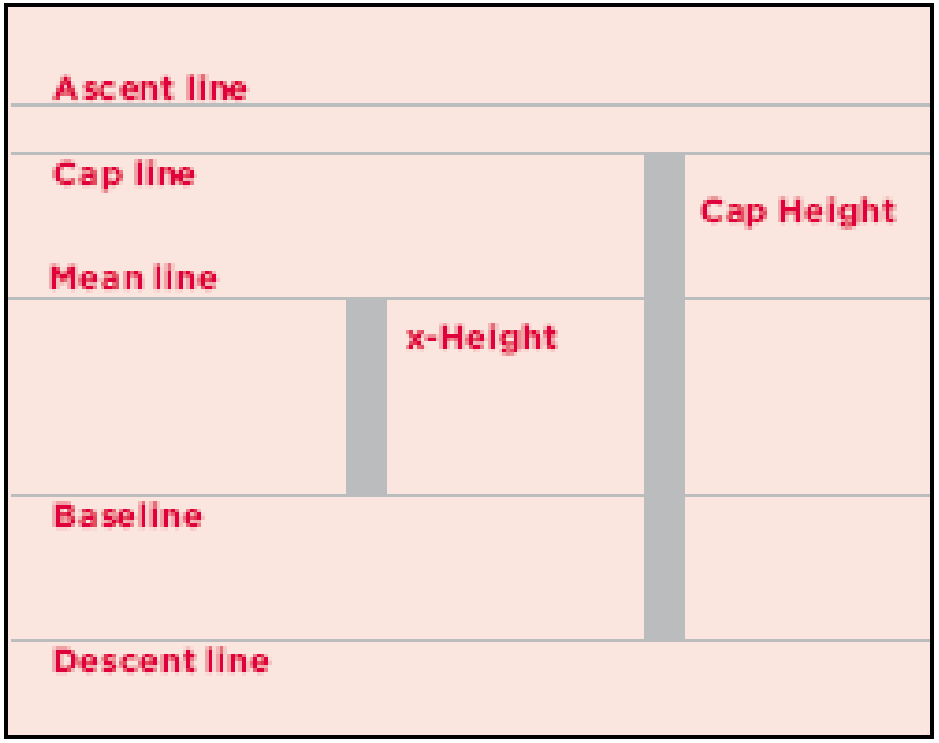
Caracterele

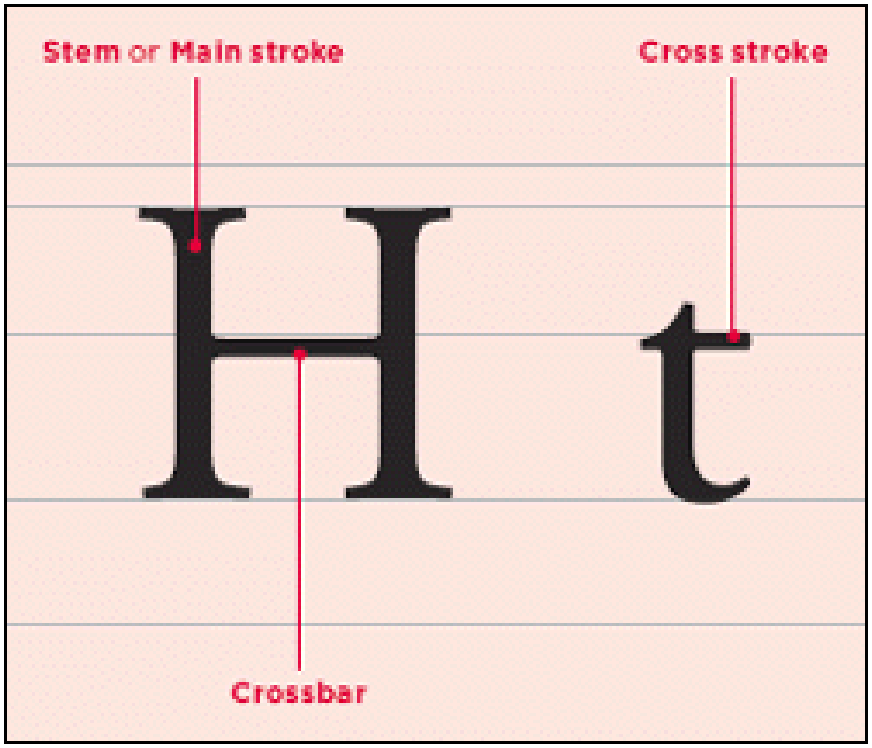
- Sunt usor de gasit
- Se porneste de la un pixel negru si se gasesc toti pixelii interconectati – entitate
- Fiecare entitate reprezinta un caracter sau un alt element constitutiv ce trebuie determinat
- Initial entitate = caracter
- Pentru filtrarea lor se folosesc 3 tipuri de filtre
  - “Inside” filter
  - “Merge” filter
  - “Width” filter

Reconstructia caracterelor

- Scopul acestei procesari este de a obtine caracterele din imagine intregi si nu impartite in bucati
- Se folosesc 3 filtre:
  - “Inside”
  - “Merge”
  - “Width”

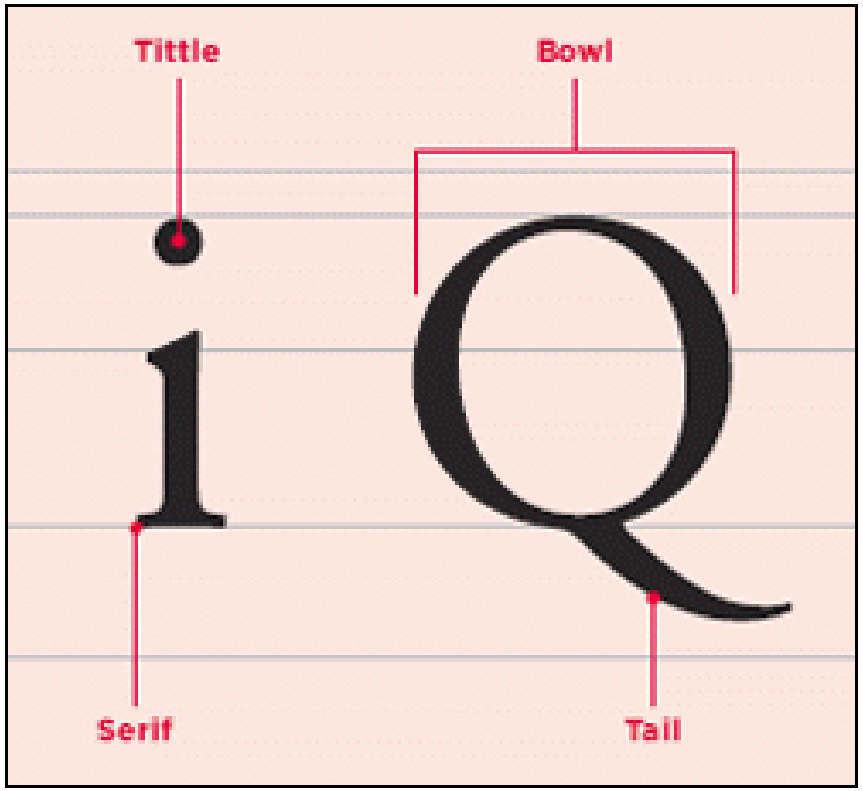
Caracteristicile caracterelor

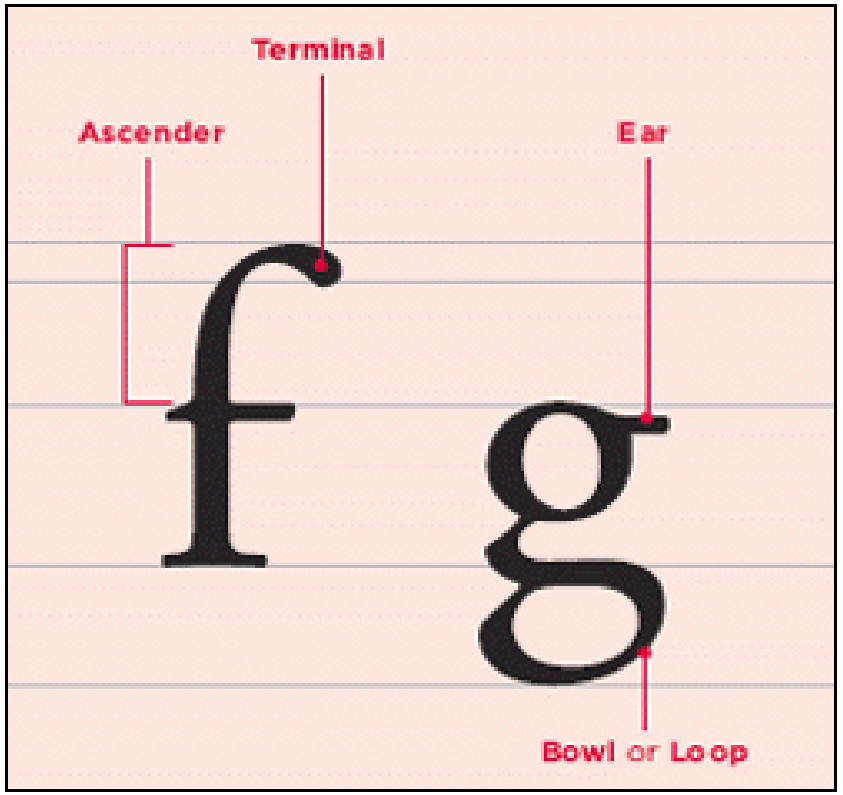




http://www.dynamicgraphics.com/dgm/Article/28539/

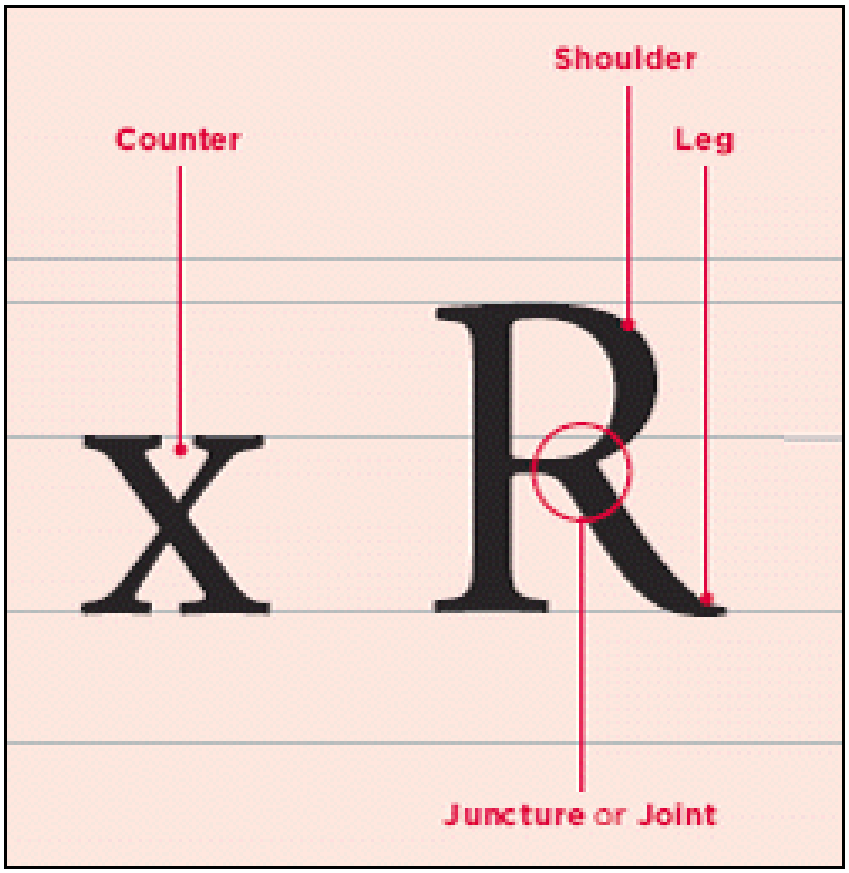
Caracteristicile caracterelor

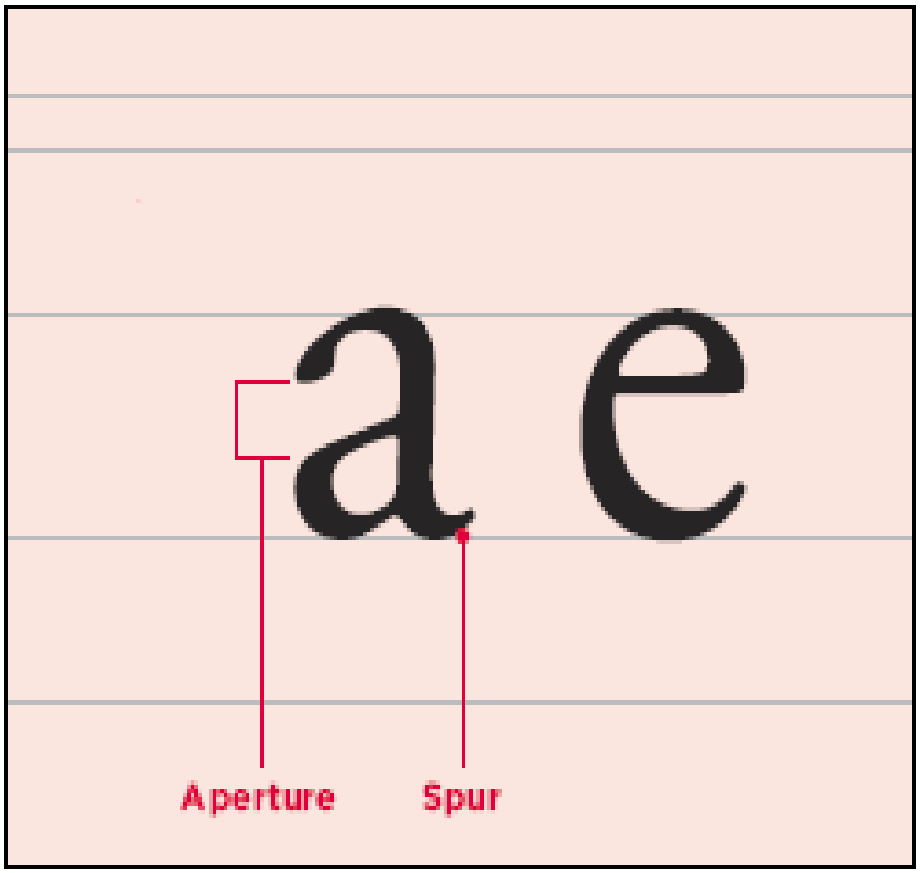




http://www.dynamicgraphics.com/dgm/Article/28539/

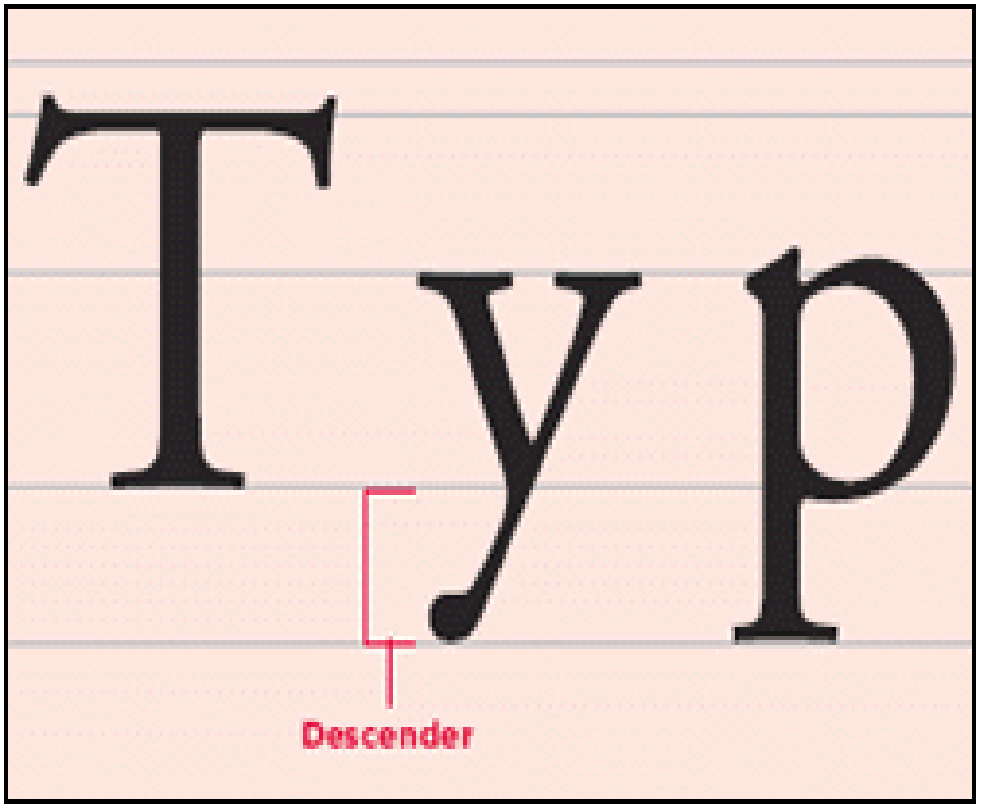
Caracteristicile caracterelor

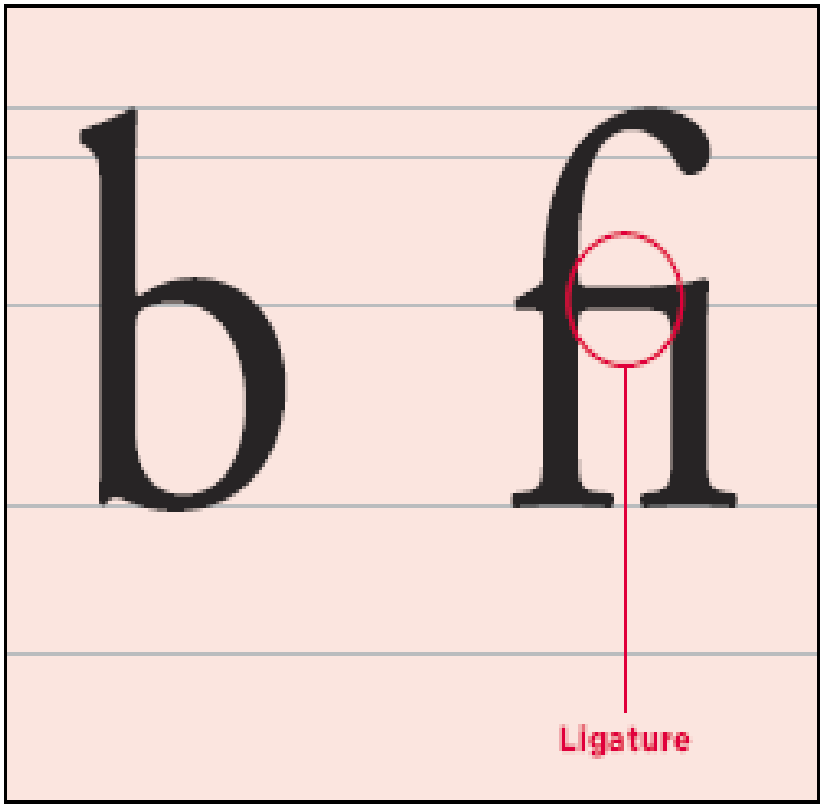




http://www.dynamicgraphics.com/dgm/Article/28539/

Caracteristicile caracterelor

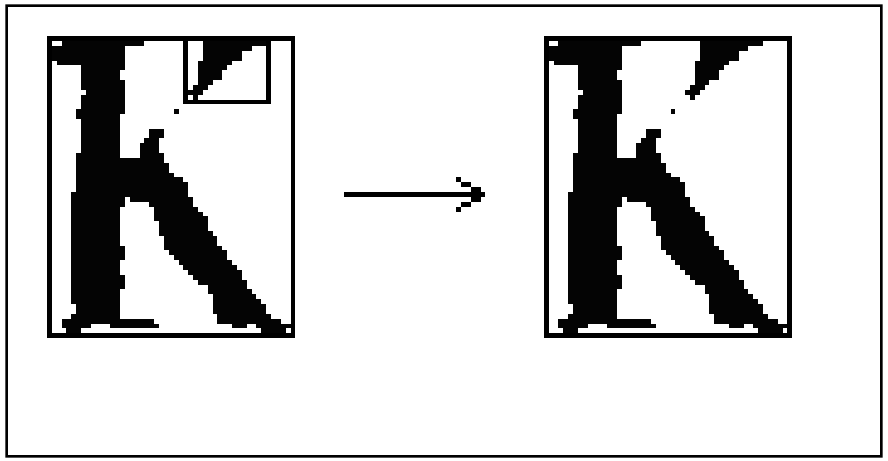




http://www.dynamicgraphics.com/dgm/Article/28539/

“Inside” filter

- Acest filtru verifica daca exista alte dreptunghiuri ce incadreaza alte entitati mai mici in interiorul dreptunghiului de incadrare al entitatii curente
- Daca gaseste o astfel de entitate o adauga la entitatea curenta

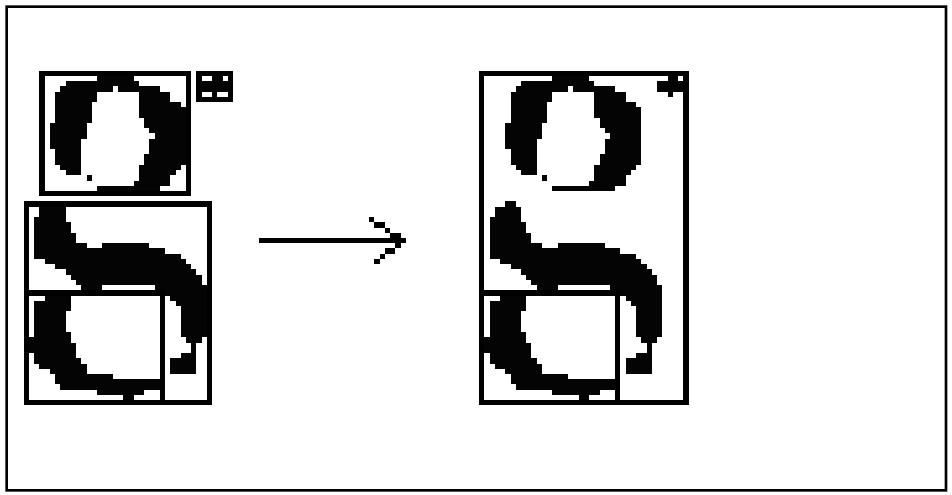


“Merge” filter

- Acest filtru reconstruieste caracterul din mai multe bucati
- Consideram 2 entitati care se afla una deasupra celeilalte (A deasupra lui B)
- Pentru fiecare entitate determinam dreptunghiurile de incadrare, ce au ca si parametri: x1,x2 – limita din stanga/drepata de pe abcisa si y1,y2 – limita de sus/jos de pe ordonata

“Merge” filter

- Daca pentru doua entitati se indeplinesc conditiile
  - Centrul fiecareia este incadrat de marginile celeilalte – B.x1<(A.x1+A.x2)/2<B.x2 si A.x1<(B.x1+B.x2)/2<A.x2
  - Distanta dintre ele nu depaseste un anumit prag, este destul de mica – A.y2 - B.y1
- Atunci atunci acestea se unesc



“Width” filter

- Scopul acestei filtrari este de a obtine doar caractere pentru procesare nu si imagini sau semne de punctuatie
- In cazul in care caracterul are una din dimensiuni mult mai mare ca cealalta si raportul de umplere (“fill ratio”) este mai mare de 80% atunci acea entitate este eliminata din lista caracterelor

Caracterele

- Au diferite proprietati: font, boldness, italic
- Fontul se determina verificand inaltimea dreptunghiului de incadrare al fiecarui caracter (entitatii)
- Valoare fontului pe acel paragraf va fi reprezentat de varful histogramei create cu valorile obtinute
- Pe aceeasi linie detectata, caractere cu fonturi diferite vor apartine paragrafelor diferite
- Mai mult, si spatiul dintre caractere si cuvinte poate sa fie folosit ca o masura pentru detectia fontului

Caracterele

- Boldnessul si dimensiunea fontului pot reprezenta o caracteristica importanta deoarece pot separa titlul de restul paragrafului
- Un algoritm pentru determinarea boldness-ului compara raportul dintre numarul de pixeli de pe contur si numarul de pixeli negri totali ai entitatii; pentru caraterele bold acest raport o sa fie mai mic
- Alt algoritm calculeaza dimensiunea “creionului”
  - Se gaseste cea mai mica linie pe directiile orizontale si verticale plecand din fiecare pixel al entitatii
  - Se realizeaza o histograma cu aceste valori si se alege maximul histogramei ca valoare predominanta
  - Aceasta va fi valoarea exacta de boldness
- Primul algoritm este folosit pentru a face o comparatie intre caractere, iar al doilea este folosit pentru a gasi masura exacta de boldness

Caracterele

- Italic (gradul de inclinarea) poate duce la aceleasi diferentieri ca si fontul sau grosimea
- Si in acest caz se pot utiliza doi algoritmi:
  - Primul este denumit metoda “latimii” si calculeaza latimea dreptunghiului de incadrare
  - Apoi se roteste cu -16 si +16 grade entitatea si se calculeaza de fiecare data latimea noului dreptunghi obtinut
  - Caracterele italice vor avea o latime initiala mai mare, iar cele neitalice mai mica

Caracterele

- Al doilea algoritm foloseste o metoda asemanatoare
- Roteste caracterul cu -16 si +16 grade si calculeaza in toate cele 3 cazuri cel mai lung segment vertical
- Acesta va apartine caracterului neitalic
- De la aceste reguli fac abstractie caracterele considerate italice prin natura, ex: A