



UNIUNEA EUROPEANĂ



GUVERNUL ROMÂNIEI



Instrumente Structurale  
2007-2013



# Platformă de e-learning și curriculum e-content pentru învățământul superior tehnic

**AEACD**

## **22. Preprocesari si Posprocesari OCR**

# OCR

- Optical Character Recognition
- Versiuni free: FreeOCR, Tesseract
- Versiuni comerciale – optiuni variate, detectie mai buna
- Scop: Detectia caracterelor dintr-o imagine pe baza unui dictionar setat la inceput sau nu
- Reconstructia caracterelor si compararea lor
- Rezultatele pot fi nesatisfacatoare, iar motorul OCR nu poate fi imbunatatit de catre utilizator

# OCR

- Apare nevoia imbunatatirii calitatii imaginii si a corectarii rezultatelor obtinute de OCR
- Procesarii OCR:
  - Preprocesari
  - Postprocesari
  - Procesari bazate pe feedback

# Preprocesari OCR

- Imbunatatirea imaginii:
  - Binarizare
  - Deskew
- Binarizarea este facuta special pentru imbunatatirea rezultatelor OCR
- Se foloseste o binarizare ce pastreaza cat mai mult din pixeli de foreground

# Postprocesare OCR

- Pe baza rezultatelor OCR, se folosește un dicționar și se reverify toate cuvintele găsite în acest dicționar
- Cele găsite corect sunt păstrate iar celelalte corectate

# Postprocesare OCR

- Aceasta corectura se face folosind trei operatii:
  - Adunare: se adauga fiecare litera din alfabet in cuvant pe fiecare pozitie pentru a obtine posibilele variante corecte
  - Scadere: se elimina pe rand fiecare litera din cuvant
  - Substitutie: se inlocuieste pe rand fiecare litera din cuvant cu fiecare litera din alfabet
- In final, vor rezulta mai multe variante corecte
- Folosind un sistem de votare bazat pe informatie semantica si sintactica se va gasi cuvantul corect

# Postprocesare OCR

- Algoritmul anterior ofera rezultate bune, insa timpul de rulare poate fi foarte mare in anuminte cazuri
- Din acest motiv, este necesar sa verificam doar literele din cuvinte ce pot fi gresite
- Fiecare litera are alocata un grad de confidenta din partea OCRului ce va fi folosit pentru selectia celor posibil eronate

# Procesare bazata pe feedback

- Combina pre si post procesarea OCR
- La sfarsitul pasului de postprocesare se verifica rezultatele obtinute si gradul de imbunatatire al acestora
- Daca rezultatele sunt nesatisfacatoare, se revine la pasul de preprocesare, incercand sa se imbunatateasca o anumita caracteristica
- Este necesara crearea unui sistem de evaluare automat satisfactor