

Documentatie

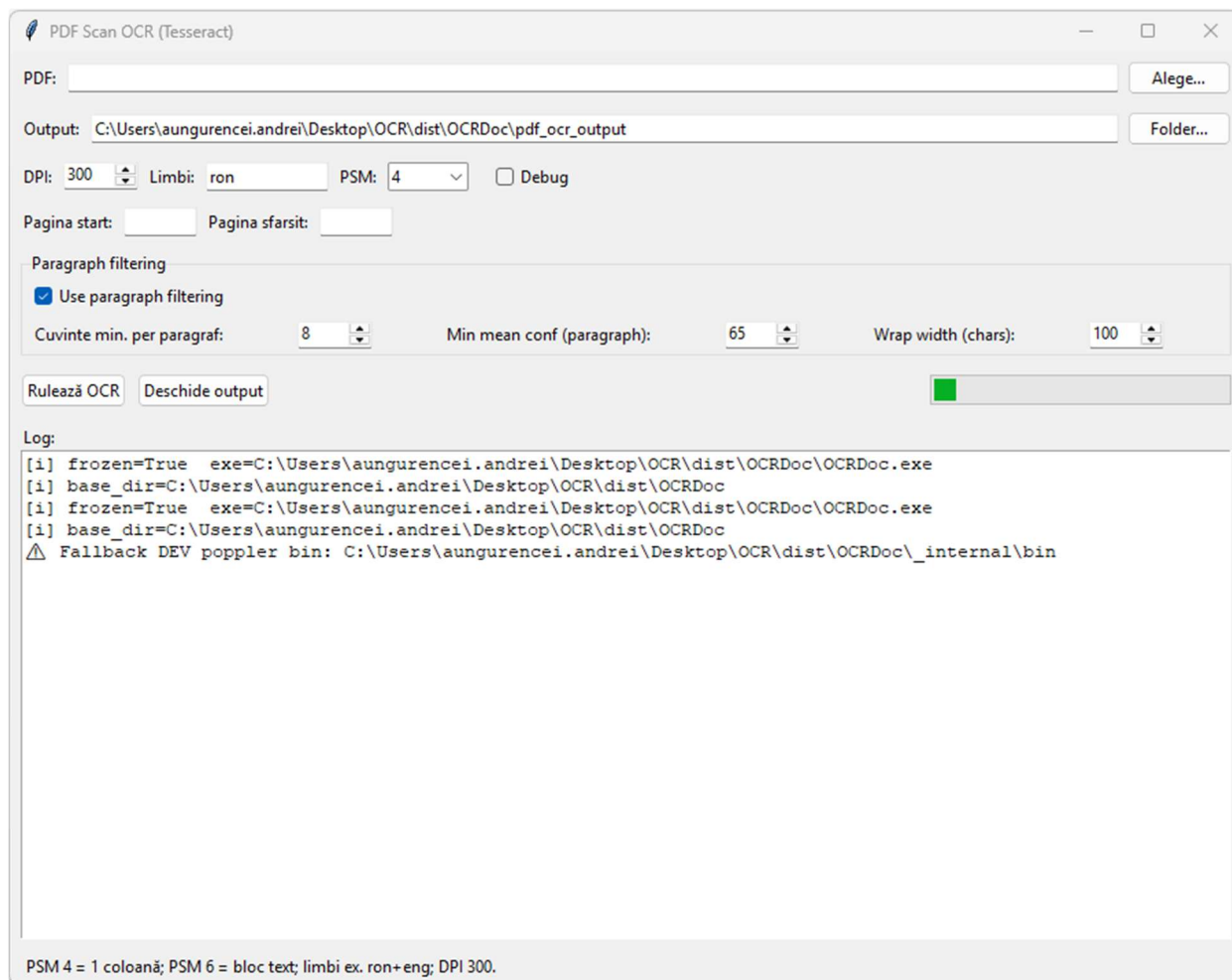
Aplicatie OCR

1. Descriere

Documentele PDF care au fost scanate sunt de fapt niste poze, din acest motiv textul scanat nu se poate refolosi decat daca este rescris de mana. Acesta aplicatie usureaza munca omului de a refolosi acel text intr-un alt fisier de ex. Word. Momentan la versiunea 0.1.0 pe care o lansam aceasta va extrage textul din acel document scanat si le va adauga intr-un fisier text. In urma testelor de catre utilizatori se vor adauga imbunatatiri si la cerere alte functionalitati ale aplicatiei.

2. Instructiuni de utilizare

Aplicatia la prima deschidere:



Parametrii pentru
depanare/optimizare

Selectarea paginilor.
Implicit sunt toate
paginile

Buton pentru a rula
aplicatia OCR

Buton pentru a
deschide folderul de
output

PDF Scan OCR (Tesseract)

PDF: Alege...

Output: C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\pdf_ocr_output Folder...

DPI: 300 Limbi: ron PSM: 4 ☐ Debug

Pagina start: Pagina sfarsit:

Paragraph filtering

☒ Use paragraph filtering

Cuvinte min. per paragraf: 8 Min mean conf (paragraph): 65 Wrap width (chars): 100

Rulează OCR Deschide output

Log:

```
[i] frozen=True exe=C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\OCRD.exe
[i] base_dir=C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD
[i] frozen=True exe=C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\OCRD.exe
[i] base_dir=C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD
⚠ Fallback DEV poppler bin: C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin
```

PSM 4 = 1 coloană; PSM 6 = bloc text; limbi ex. ron+eng; DPI 300.

Butonul pentru a alege
calea spre documentul
dorit

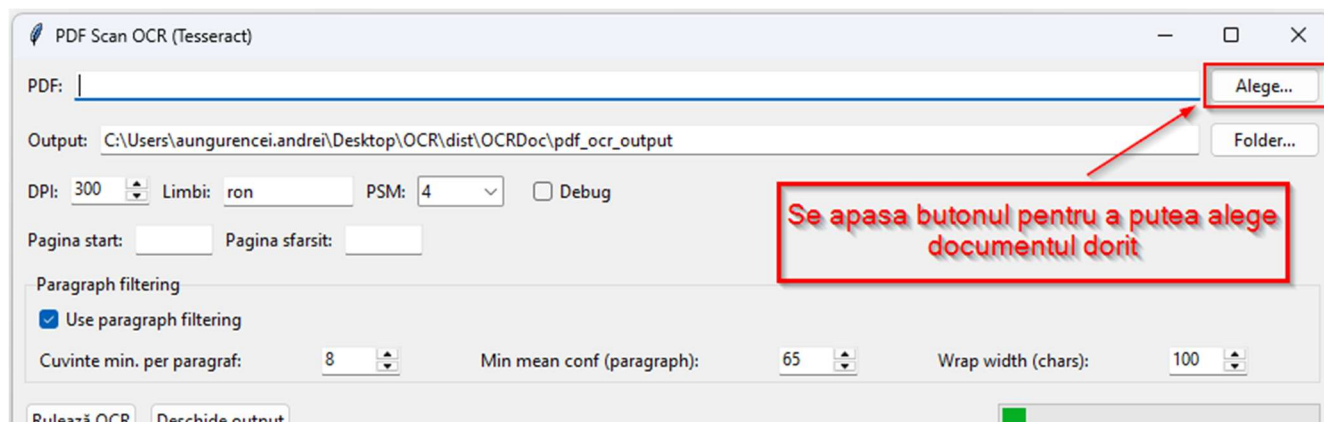
Butonul pentru a alege
calea spre folderul de
output

Parametrii pentru filtrarea
de paragrafe pentru un text
cat mai curat

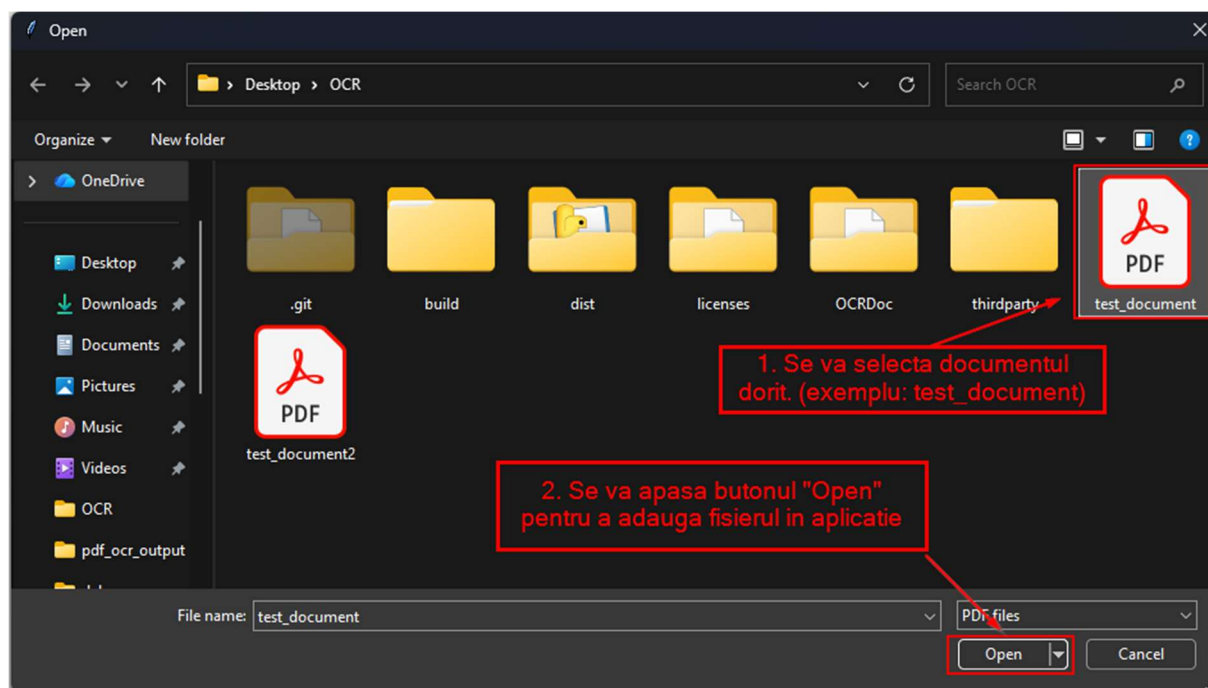
Consola pentru a vizualiza
progresul rularii si daca se
semnaleaza diverse erori.

Rularea Aplicatiei

1. Pentru a incepe lucrul trebuie introdusa calea spre document. Acest lucru se face prin apasarea butonului "Alege...".



2. Se va deschide o fereastră unde apoi utilizatorul isi va selecta documentul dorit. Ex:



3. **Optional:** daca se doreste extragerea de text de pe anumite pagini introduceti numarul paginii de start si de sfarsit in casetele cu acelasi nume (ex. 3 si 7 pentru paginile intre 3 si 7; 4 si 4 pentru a extrage doar pagina 4)

Atentie!!! Pentru extragerea intregului document se va lasa gol deoarece valoarea este implicita.



4. **Optional:** Pentru ca textul rezultat sa fie curat fara alte detectii neconforme este adaugat acest parametru de cuvinte minimum per paragraf care valoarea lui implicita este 8. Aceasta valoare se poate modifica de catre utilizator dupa plac. Se recomanda a se folosi o valoare intre 3 – 8 pentru un text cat se poate de curat. Ceilalti parametrii nu trebuie modificati.

PDF: C:/Users/aungurenci.andrei/Desktop/OCR/test_document.pdf Alege...

Output: C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\pdf_ocr_output Folder...

DPI: 300 Limbi: ron PSM: 4 ☐ Debug

Pagina start: Pagina sfarsit:

Paragraph filtering
☒ Use paragraph filtering

Cuvinte min. per paragraf: 8 Min mean conf (paragraph): 65 Wrap width (chars): 100

Caseta pentru introducerea valorii discutate la punctul 4

5. Pentru rularea programului se va apasa butonul “Ruleaza OCR”. Se asteapta apoi ca in consola sa apara “gata”

PDF Scan OCR (Tesseract)

PDF: C:/Users/aungurenci.andrei/Desktop/OCR/test_document.pdf Alege...

Output: C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\pdf_ocr_output Folder...

DPI: 300 Limbi: ron PSM: 4 ☐ Debug

Pagina start: Pagina sfarsit:

Paragraph filtering
☒ Use paragraph filtering

Cuvinte min. per paragraf: 8 Min mean conf (paragraph): 65 Wrap width (chars): 100

Rulează OCR Deschide output

Log:

Rulează OCR Deschide output

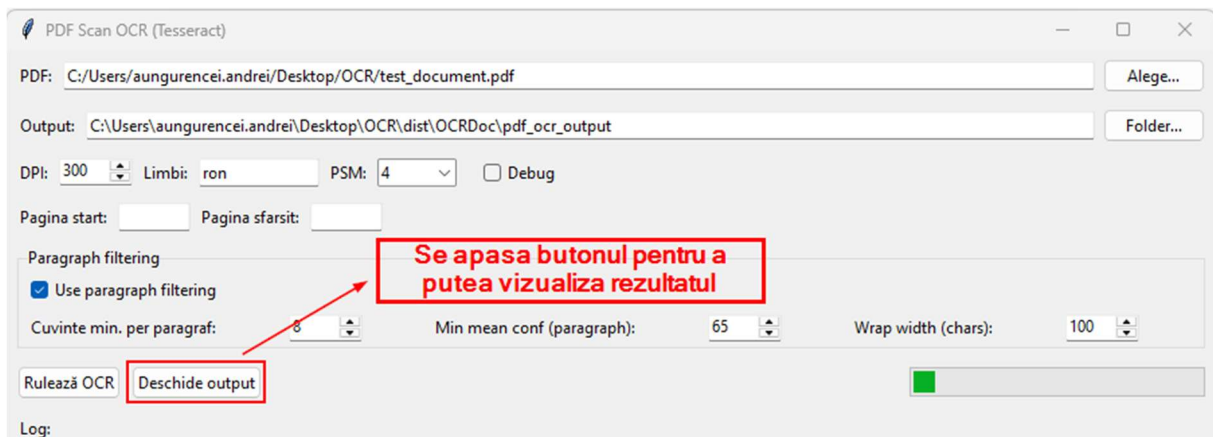
Log:

```
⚠ Fallback DEV poppler bin: C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin
[info] BIN PATH - C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin
[check] tesseract.exe: True ->
C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin\tesseract.exe
[check] tessdata/: True -> C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin\tessdata
[check] osd.traineddata: True
[check] eng.traineddata: True
[check] ron.traineddata: True
[check] pdftoppm.exe: True -> C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin\pdftoppm.exe
[check] pdftocairo.exe: True ->
C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin\pdftocairo.exe
[set] pytesseract.tesseract_cmd =
C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\_internal\bin\tesseract.exe
- PDF: C:/Users/aungurenci.andrei/Desktop/OCR/test_document.pdf
- Output: C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\pdf_ocr_output
- DPI=300, PSM=4, Lang=ron, Debug=False, Start=None, End=None
- Paragraph filter: use=True, min_words=8, min_conf=65, wrap=100
[pagina 1] conf_global=86.67 (raw_mean=86.67)
🔪 Text combinat (clean):
C:\Users\augurenci.andrei\Desktop\OCR\dist\OCRD\pdf_ocr_output\test_document.pdf_20251113_113415.txt
☒ Gata.
```

PSM 4 = 1 coloană; PSM 6 = bloc text; limbi ex. ron+eng; DPI 300.

6. Pentru vizualizarea textului extras se va apasa pe butonul “Deschide output”. Se va deschide o fereastră unde veti avea un fisier text cu numele documentului ales si data rularii aplicatiei. In acesta se va gasi textul din toate paginile documentului combinate.

Pe langa acesta va mai exista un folder cu textul pe pagini separate. Acesta folder se va suprascrie de fiecare data cand aplicatia va rula.



Se va deschide aceasta fereastră unde puteti gasi textul rezultat din OCR

