Analysis, Processing, Information Retrieval and Storage of Document Images

Habilitation Thesis

Costin-Anton BOIANGIU

University POLITEHNICA of Bucharest

2016

Rezumat

De-a lungul perioadei post-doctorale efortul propriu de cercetare a fost îndreptat îndeosebi spre analiza și prelucrarea documentelor scanate.

În secțiunea 2.1 dedicată cercetării desfășurate si rezultatelor obținute, este dezvoltată tema reconstrucției documentelor, urmărind cursul normal de procesare, de la faza de scanare până la conversia în informație digitală, cu menținerea aspectului original.

Secțiunea este împărțită în șase capitole dedicate temei centrale și unul dedicat altor arii secundare conexe.

Descrierea cercetării temei centrale începe cu achiziția imaginilor prin scanarea documentelor, având în vedere păstrarea cât mai fidelă a culorilor și clarității documentului original. Pentru obținerea acestui obiectiv se descriu si dezvoltă metode de măsurare și calibrare de precizie.

După achiziția lor, imaginile trec printr-o fază de corecție a clarității si alta de conversie la nuanțe de gri cu pierdere minimă de informație utilă, aceasta fiind utilă în fazele ulterioare de binarizare. Același procedeu poate fi aplicat și pentru a tipări un document color utilizând un dispozitiv de tipărire mono-culoare.

Pentru a transmite o imagine spre fazele de analiză, sunt prezentați o serie de algoritmi de binarizare (globali, locali, adaptivi), culminând cu o abordare de procesare ce nu implică intervenția utilizatorului.

Ultima etapă de corecție abordată este cea a detectării și corectării rotației cauzate de procesul de scanare. Pentru că este posibil ca documentul să depășească fizic dimensiunea dispozitivului de scanat, este necesara si aplicarea metodelor de lipire a fragmentelor de documente. Fragmentele pot conține sau nu regiuni comune. Tot în etapa aceasta sunt corectate defectele de scanare de la extremitățile paginilor și sunt decupate paginile urmărind detecția automată a zonei de imprimare.

După etapele de corecție și conversie urmează interpretarea propriu-zisă. In aceasta secțiune sunt discutați mai mulți algoritmi de detectare și clasificare. Capitolul începe cu algoritmi de detecție a spatiilor albe sau a separatorilor generali, de orice formă. Apoi continuă cu o modificare a transformatei Hough în scopul detectării corecte a liniilor frânte, ce formează tabele în documente. Din punctul de vedere al conținutului, urmează algoritmi de detectare a liniilor de text, inclusiv a celor scrise de mană, caracterizate de o curbare puternică. Tot aici sunt tratate și metode de detectare și reparare a literelor fragmentate în urma procesului de binarizare. Procesul următor analizează și măsoară caracteristicile fonturilor ce se găsesc în document: boldness-ul, înclinarea, dimensiunea. Folosind metodele prezentate, se pot clasifica zonele de document în arii ce conțin doar text și care pot fi ordonate pentru a fi trimise la OCR. In aceasta faza se elaborează mai multe metode de segmentare logică a paginii, etichetare și ordonare a regiunilor. Capitolul se finalizează cu un rezultat important din domeniul Geometriei Computaționale, Beta-Shape, folosit la încadrarea optimală în poligoane menite să evite suprapunerile intre clustere.

Capitolul urmator este dedicat interpretării propriu-zise a literelor și cuvintelor prin OCR. Se propune și un pas de postprocesare pentru corecție, folosind dicționare. Tot aici se prezintă diverse formate pentru

salvarea rezultatului după trecerea prin toate fazele. Unul din formatele importante propuse este cel obținut utilizarea tehnologiei MRC, care comprimă fiecare regiune in mod diferit, pentru a optimiza atât calitatea cat și dimensiunea.

Ultimul capitol conține cercetări domenii conexe care ar putea fi continuate în viitorul apropiat. Acestea sunt prelucrarea prin votare, introdusă de segmentarea prin votare și analiza generală a conținutului imaginilor variate, introdusă de clasificarea monedelor.

Prin prisma lucrărilor viitoare, am menționat direcțiile de cercetare ce urmează a fi abordate (localitate/globalitate cu extensii peste domeniul prelucrării imaginilor – anume in predicția seriilor pe pieţele electrică și respectiv financiară).

Cercetările mai avansate vor avea loc prin propuneri de proiecte naționale sau internaționale. Aici sunt exemplificate TRISEMA (predicția pieței electrice folosind algoritmi din prelucrările de imagini și rețele neurale) si FINCRIS (măsurarea crizelor financiare folosind o scara analogica de cutremur financiar - Richter).

Activitățile cu studenții vor evolua spre o învățare competitivă, proiectele echipelor “luptându-se” unul împotriva altuia, într-un mediu virtual. De asemenea, se propune un mediu online pentru desfășurarea de astfel de competiții (Geek Arena).

Din punctul de vedere al procesului de predare a materiei, și acesta se dorește a fi îmbunătățit, în fiecare an considerând comentariile studenților și axându-ne pe utilitatea subiectelor în cercetările actuale. Activitatea cu studenții se va dezvolta prin conectarea la cercetarea modernă și prin integrarea activa a studenților în cercetare.