Kompleksowy system identyfikacji autora rękopisu za pomocą cech pisma

mgr inż. Jakub L. Pach, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska

System ma na celu automatyczna identyfikację autora rękopisów łacińskich za pomocą indywidualnych cech charakteru pisma. Obecnie orzekanie autentyczności dokumentu jest wykonywane przez specjalistę grafologia, co stanowi finansową barierę dla instytucji w Polce i na świecie do weryfikacji autentyczności cennych manuskryptów. Wymagania wobec takiego systemu są trudne ze względu na konieczność oddzielenia pisma od szumu na fotografii cyfrowej. W dokumentach maszynowych pismo jest równoległe względem linii tekstu. Litery są powtarzalne i nie nachodzą na siebie, co przeważnie nie jest spełnione w piśmie odręcznym. System składa się z trzech części – przetwarzania wstępnego, ekstrakcji cech pisma i klasyfikacji. Pierwszy etap jest istotny, ponieważ błędy popełnione tutaj obniżają skuteczność całego systemu. Celem przetwarzania wstępnego jest oddzielenie tła od teksu właściwego, a także podzielenie bloków tekstu na linie pisma. Obejmuje ono binaryzację, wykrywanie obszaru tekstu oraz linie pisma. Binaryzacja została opracowana na podstawie metody gaussowskiej, dostosowanej do skomplikowanego charakteru danych i zoptymalizowana obliczeniowo. Wykrywanie obszaru pisma zaimplementowano w postaci kombinacji algorytmów RLE (ang. Run Length Encoding). Moduł odpowiedzialny za wykrywanie linii został oparty na transformacie Hougha. Dzięki wprowadzeniu do niej wspomagających spójnych składowych grafu SCC (ang. supporting CC) skuteczność wykrywania linii dla manuskryptów łacińskich została istotnie zwiększona. Po nich następuje ekstrakcja cech pisma wykorzystująca transformatę SIFT (ang. Scale Invariant Feature Transform ), zapewniającą skaloniezmiennicze przekształcenie cech. Charakter pisma został przedstawiony za pomocą dwóch cech, tj. skali i orientacji. Do budowy słownika pisarzy wykorzystano zmodyfikowaną metodę histogramową. Uproszczono ją i zoptymalizowano. Moduł odpowiedzialny za prawidłową klasyfikację wykorzystuje dwa niezależne względem siebie klasyfikatory. Do podziału danych na trenujące i testujące wykorzystano walidację krzyżową Leave One Out. Podejmowanie decyzji opisano stosując klasyfikator najbliższego sąsiada i drzewa decyzyjnego.