

Interactive Adaptive Text Localization and Distinction

Samuel James Pinson

Dr. William Barrett

Computer Science Department

Brigham Young University

Text localization, the process of locating text in a document, is performed before optical character recognition (OCR) to reduce computation time. Only portions of a document that contain text need to be sent to an OCR engine. Text distinction, the process of distinguishing between multiple classes of text, improves OCR accuracy by directing text to appropriate OCR engines. Handwritten text is sent to a handwriting OCR engine, and machine print is sent to a machine print OCR engine.

Text localization and distinction have traditionally been approached with static classifiers that rely on large amounts of specially prepared training data and/or empirically determined heuristics. These classifiers are often not interactive because of high computational requirements. Further, these classifiers are not adaptive, because the classifier cannot (easily) be altered to learn new classification problems.

We present an interactive and adaptive classification system for text localization and distinction. Classification problems are both specified and solved interactively as a human user tags class prototypes. A decision tree is induced from the class prototypes provided. The human user then identifies additional class prototypes to correct misclassifications until acceptable performance is achieved. We evaluate the performance of our system on several real world problems.