

Research on feature extraction applied to word search in document images

Le Xuan Long

Abstract: This study presents a feature extraction framework tailored for word spotting in document images. Addressing the challenges of complex layouts, varied fonts, and degraded image quality, the proposed method integrates robust local descriptors with a learned representation to enhance retrieval accuracy. Key contributions include the design of a hybrid feature set combining gradient-based and structural information, optimized for both printed and handwritten text. An efficient indexing scheme is employed to enable scalable search within large document collections. Experimental evaluations on benchmark datasets demonstrate significant improvements in mean average precision compared to existing approaches, particularly for noisy and historical documents. The framework effectively balances discriminative power and computational efficiency, offering a practical solution for digitized archival retrieval. This work advances the field by providing a versatile feature extraction pipeline that adapts to diverse document conditions, facilitating reliable keyword spotting without full optical character recognition.

Keywords: Feature extraction, Word search, Document image, Image processing