Localizing Text and Symbols in Images from Biomedical Journal Articles

J. Li, D. Demner-Fushman, S. Antani, G. Thoma

Abstract:

Automatic localization and recognition of text and symbols in images found in biomedical journal articles could significantly improve indexing and retrieval of biomedical literature, thus contributing to clinical decision support. Main difficulties in automatic localization of text and symbols in medical images are in the irregularity of their occurrence and in the variety of font features. The difficulties are compounded by image quality, image background interference, arbitrary location, and variability in the text block size. We present results of automatic localization and annotation of text and symbols in medical images. Our methods take advantage of gross image features and automatically identified image modality (classification of images into 4 broad types: color, illustration, radiographic and other.) 2D adaptive noise removal Wiener filtering is used as preprocessing step to reduce the image noise. Automatic histogram thresholding, morphological method, Quadtree technique, DCT, and connected component analysis are selectively used on different image types for extracting text and symbol locations. Text area merging and region growth techniques are used as post-processing methods to improve the precision of the bounding box locations. Initial experiments on 100 images achieve precision and recall of 78.42% and 89.38%, respectively, with an average accuracy of 72.02%.