

Logical Entity Recognition in Multi-Style Document Page Images

Mao S ^a, Xu Z ^b, Tjahjadi T ^b, Thoma GR ^a

Proc. 18th International Conference on Pattern Recognition (ICPR), pages 876-879, Hong Kong, China, August. 2006

^a U.S. National Library of Medicine, Bethesda, MD 20894, USA

^b School of Engineering, University of Warwick, Coventry, UK

^a {smao, gthoma}@mail.nih.gov , ^b {Zheng.Xu, T.Tjahjadi}@warwick.ac.uk

Abstract:

Logical entity recognition in document page images is the essential part of a document image analysis system. A heterogeneous collection of document pages usually has many layout styles. Features extracted from same logical entities in different styles may have very different values and vice versa. Therefore, logical entity classifiers learned from a training set of multi-style document pages may not be reliable due to possible feature overlap of different logical entities in different styles. In this paper, we propose a novel method in which style information is used in both logical entity classifier training and recognition phases. In the training phase, training data are first classified into distinct styles, and a dedicated Support Vector Machine (SVM) is then learned for each style. In the recognition phase, the style of a new document page image is first identified and its logical entities are then recognized using corresponding SVM. We show in our experiments that the use of the style information significantly improves the accuracy of logical entity recognition in multi-style document page images.