STYLE QUANTIFICATION FOR FIELD CLASSIFICATION

By

Xiaoli Zhang

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: Electrical, Computer and Systems Engineering

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

George Nagy, Thesis Adviser

Dr. W. Randolph Franklin, Member

Dr. Mukkai Krishnamoorthy, Member

Dr. Kenneth S. Vastola, Member

Rensselaer Polytechnic Institute
Troy, New York

December 2006
(For Graduation May 2007)

ABSTRACT

The co-occurring patterns in a group carrying the traits of a common origin are statistically dependent via an underlying style context. Exploiting style consistency in groups of patterns from multiple sources has been demonstrated to yield higher accuracies in OCR applications. The accuracy gains obtained by a style consistent classifier depend on the amount of style in a dataset in addition to the classifier itself. The computational complexity of style-constrained classifiers precludes their applicability in situations where datasets have small amount of style. We formally define two kinds of style information involved in OCR systems, intra-class style and inter-class style. We present a system to quantify the amount of intra-class and inter-class styles using entropy, correlation and mutual information. To validate our style measures, we propose style homogenization to eliminate the styles in a dataset. We demonstrate the efficacy of the three proposed metrics by comparing the amount of styles between the datasets before and after style homogenization, and through their correspondence with the number of errors obtained by field classification. We also use unsupervised learning to exploit broad inter-class style context which is still beneficial to field classification. We present a clustering scheme to group sources into fewer broad styles to reduce the small sample effects caused by finite sample size within a source. This allows applying the discrete style classifier, which performs on such a dataset better than the style-conscious quadratic classifier. We demonstrate the effectiveness of our approaches on real machine-printed and handwritten data. On the NIST handwritten digit data, the discrete style classifier obtains a field error rate up to 2% lower than an already accurate style-conscious quadratic classifier by operating on triples of patterns.