Abstract

Mohamed S El-Mahallawy

Omnifont text recognition of printed cursive scripts via HMMs, compact lossless features, and soft data clustering

This paper presents an optical character/text recognition (OCR) system for cursive scripts like those of Arabic, Urdu, Persian, Kurdish, etc. This OCR system is a large-scale one in the sense of architecture, training data size, and state-of-the-art performance. The paper introduces the theoretical derivation and experimental assessment of our two main contributions deployed in this OCR. The first contribution is the design of a new autonomously normalized, lossless, and compact feature vector that enables the production of a truly robust omnifont OCR system for cursive scripts with an ASR-like HMM-based architecture. Half of the components in this feature vector are analogs and the other half are discrete, which obstructs the use of continuous Gaussian mixtures to model an aggregate of such features and mandates the use of discrete HMMs instead, which in turn necessitates the deployment of vector clustering and quantization. The second contribution is a new soft (i.e., probabilistic) vector quantization (VQ) scheme, as opposed to conventional hard-deciding VQ that we analytically derive and then deploy to alleviate overfitting and boost the robustness of our OCR against different kinds of obtrusive variances. We present experimental evidence of these benefits of the presented soft VQ scheme to our OCR. Other machine-learning systems with VQ modules may also deploy soft VQ to obtain the same benefits.