

Quick International Character Recognition (QulChaR)

Amlan Kundu

703-983-6711 • akundu@mitre.org

Linda Van Guilder

703-983-6577 • lcvg@mitre.org

The logo for the MITRE Technology Program, featuring a stylized graphic of stacked blocks in yellow, orange, and blue to the left of the text.

**MITRE
Technology
Program**

MITRE Sponsored Research

Problem

- **Many documents of potential intelligence value are hard copy and contain hand-written foreign languages**
- **Optical handwriting recognition technology for document triage is critical**
- **Handwriting technologies for many script / language pairs will not be commercially developed**
- **Each script-language pair presents a unique set of challenges to character recognition**

Background

সমস্ত মানুষ স্বাধীনভাবে সমান মর্যাদা এবং অধিকার নিয়ে জন্মগ্রহণ করে তাঁদের বিবেক এবং বুদ্ধি আছে ; সুতরাং সকলেরই একে অপরের প্রতি আত্মসুলভ মনোভাব নিয়ে আচরণ করা উচিত ।

<http://www.omniglot.com/writing/bengali.htm>

■ Example: Bengali Script

- Left-to-right
- Cursive, characters connected at top by matra line
- Ligatures required by certain character combinations
- Vowel symbols in context written above matra, below bottom baseline on preceding consonant, or in main string with one of preceding
- Small set of diacritic marks to indicate pronunciation

يولد جميع الناس أحراراً متساوين في الكرامة والحقوق. وقد وهبوا عقلاً وضميراً وعليهم أن يعامل بعضهم بعضاً بروح الإخاء.

<http://www.omniglot.com/writing/arabic.htm>

■ Example: Arabic Script

- Right-to-left, with some left-to-right interspersed
- Cursive, mixed cursive discrete
- Each character has up to four contextual variants
- Number and location of dots distinguish characters
- Diacritic marks around character indicate pronunciation

Objectives

- Identify steps to rapidly modify existing handwriting system to handle new script / language pairs
- Compare algorithms for classification and recognition
- Expand ability to handle meaningful characteristics of other scripts, e.g., dots and diacritics
- Determine contribution of language models to recognition performance
- Create cross-lingual “super inventory” or hierarchy of character features

Activities

- **Experimentation on isolated hand-written Arabic words**
 - Identification of symbol and feature set
 - Modification of dot and diacritic handling
 - Research word-matching algorithms
 - Exploration of language modeling techniques
- **Branching to hybrid statistical – rule-based approach**
- **Testing with alternative machine learning approaches**

Highlight: Feature Research

■ Machine Learning

- Explore contribution of graphical features in isolation from system
 - Word accuracies
 - Transformation-Based Learning, 63.8%
 - J48 decision tree, 35 features (53.4%), 53 features (58.5%)
 - Naïve Bayes with discretization, 35 features (48.5%), 53 features (52.6%)
 - Character accuracies
 - Support Vector Machines, 95.16%
 - Multilayer Perceptrons, 96.21%



Highlight: Algorithm Research

Algorithm	Word Accuracy (250 Word Lexicon)	Word Accuracy (12000 Word Lexicon)
Baseline, MaxLikelihood-Bayes	62%	38%
Group-to-State	55%	25%
Multi-pass Bayes – Characters	58%	40%
Multi-pass Bayes – Groups	62%	
Multi-pass Multilayer Perceptron + SMO – Group	60%	
Multi-pass Multilayer Perceptron + SMO – Group Map	61%	38%

Impacts

- **MITRE will build expertise, be prepared to help address next language processing crisis**
- **Prototypes / modules for isolated word handwritten Arabic recognition for technology transfer**
- **Publicly released source code and MITRE-developed corpora**
- **Publications, technical reports, and journal articles to contribute to the state of the research**

Future Plans

- **FY05** Extend Arabic prototype for configurable features, classification algorithms, string matching; explore multiple machine learning approaches
- **FY06** Experiment with language modeling techniques, including phonotactic and syntactic models; gather data for second language
- **FY07** Generate prototype for second language, perform experiments in extending to third language/script combo, e.g., Bengali