

Quick International Character Recognition (QUICR)

Amlan Kundu

703-983-6711 • akundu@mitre.org

Linda Van Guilder

703-983-6577 • lcvg@mitre.org



MITRE Sponsored Research

Problem

- Many documents of potential intelligence value are hard copy and contain handwritten foreign languages.
- Handwriting recognition technology for document triage is critical.
- Handwriting technologies for most low-density languages will not be commercially developed.

Background

- **Each script and language pair presents a unique set of challenges to recognition algorithms.**
- **Example: Arabic Script**
 - Cursive
 - Right-to-left, with some left-to-right interspersed
 - 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 dots and diacritics**
- **Determine contribution of language models to recognition performance**
- **Create cross-lingual “super inventory” or hierarchy of character features**

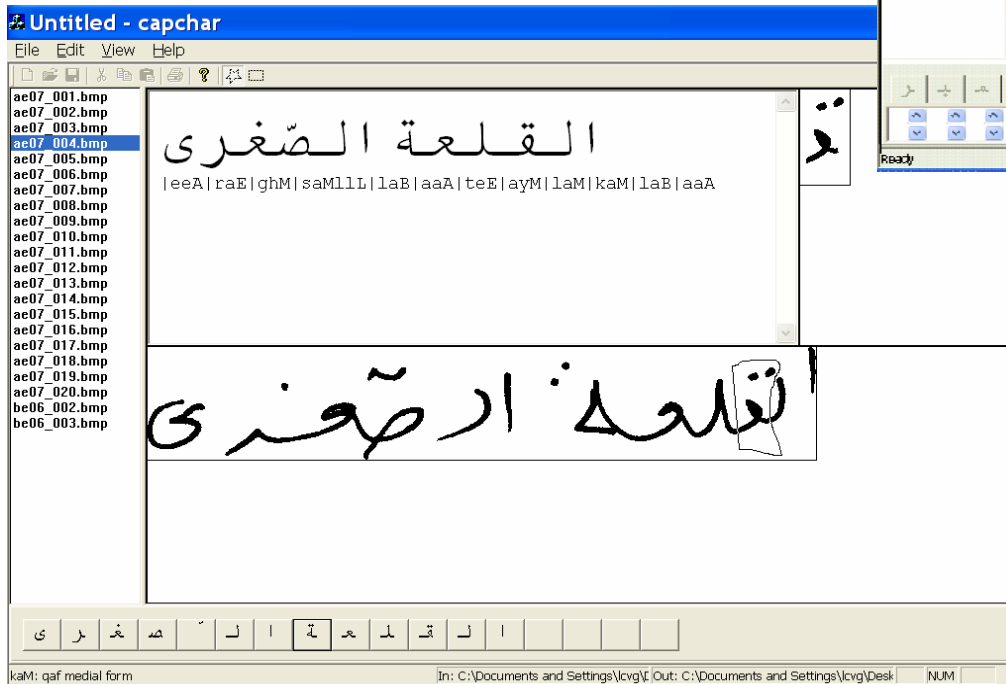
Activities

- **Development of handwritten corpora and ground-truthing tools for public release**
- **Experimentation on handwritten Arabic**
 - Identification of symbol and feature set
 - Modification of dot and diacritic handling
 - Research on word matching algorithms
 - Exploration of language modeling techniques
- **Conversion to hybrid statistical – rule based approach**
- **Implementation of alternative machine learning approaches**

Highlight: Corpus Collection

Page 1 of 16

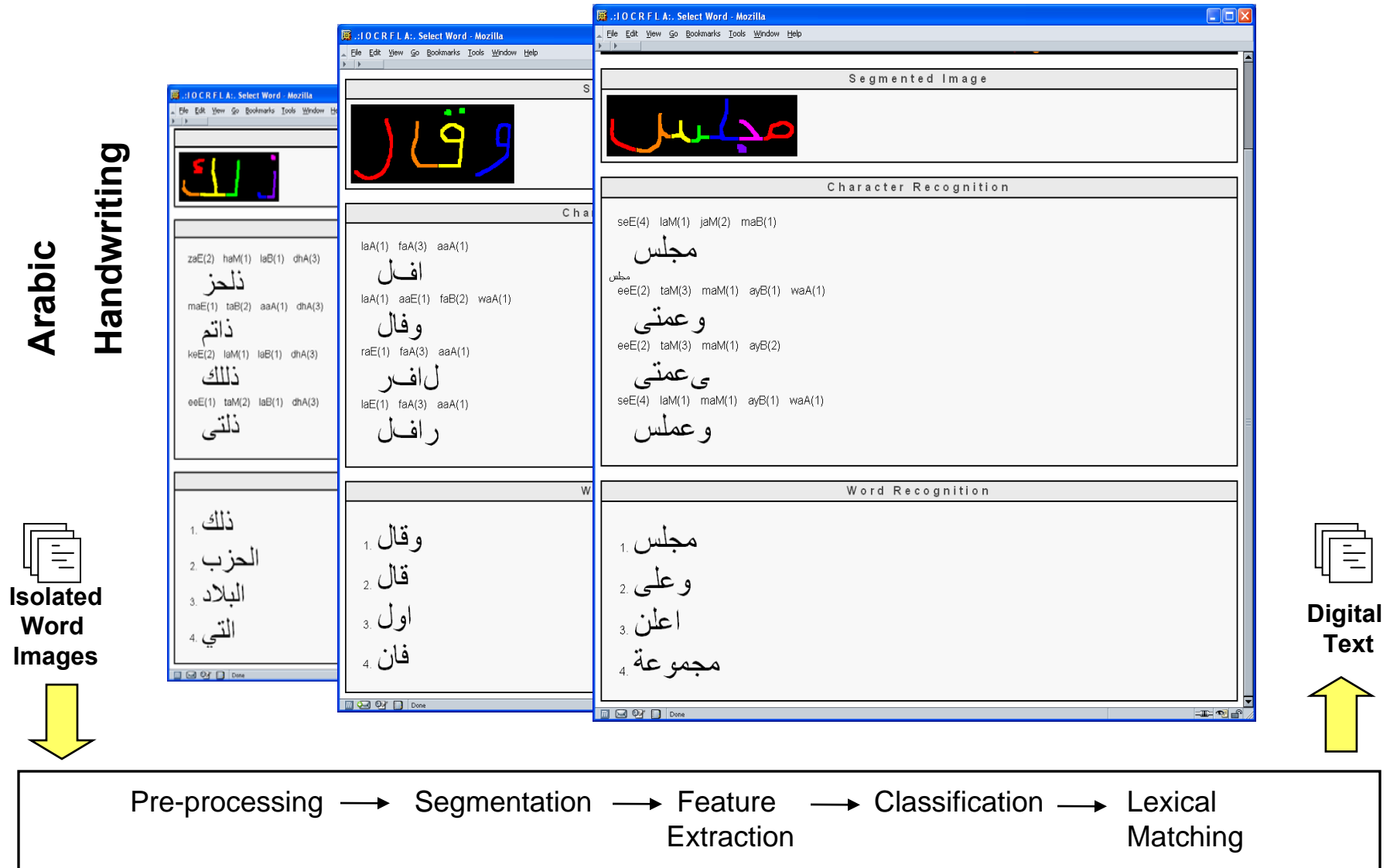
| Repetition | Repetition | Repetition | Word | # |
|------------|------------|------------|--------|---|
| | | | شغيباً | 1 |
| | | | | |
| | | | مادرات | 2 |
| | | | | |



MITRE

© 2005, The MITRE Corporation

Demonstration



Impacts

- **MITRE prepared to help address the next language processing crisis**
- **Prototypes for recognition of isolated word handwritten Arabic 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., Korean