Evaluation Issues in ImageRefiner

Kristen Summers, Luis Hernandez

November 4, 2005
Introduction

- **Goal:** automatically determine how to transform document images, to improve OCR quality

- **Motivation**
  - OCR accuracy depends on image quality
  - Image transforms may improve or degrade images
  - Automatic selection of methods desirable for:
    - Accuracy: improvement of OCR suitability vs. improvement of appearance to human
    - Processing speed: computer-based classification is faster than human-based
  - Joint work with Ilya Zavorin, Eugene Borovikov, Yaguang Yang, Mark Turner
  - Sponsored by Army Research Lab (ARL) for the Language and Speech Exploitation Resources Advanced Concept Technology Demonstration (ACTD)
ImageRefiner

- **Flexible framework for document image enhancement:**
  - Learns what image transformation to apply, given the characteristics of a document image
  - Machine-learning (ML) based
  - Handles bitonal (b/w) or grayscale images
  - Able to incorporate any new or existing image characteristic measures, image transformations, ML methods

- **Includes:**
  - 22 image characterization methods
  - 15 image transformations
  - 5 machine learning methods

- **Strategies:**
  - Transformation based features
  - Adaptive image transformations (e.g. via image segmentation)
  - Multi-step image processing (e.g. iterative processing)

- **Tested with Latin, Arabic and Thai scripts**
Many methods (transformations) are available to clean up document images before OCR.

Applying the wrong transformation(s) can result in lowered OCR accuracy.

Which transformation(s) should be applied to a specific image?
Image Transformation: improvement

Page 2

medical interest in smoking. Smoking has been controversial since the 1600's and anti-smoking pressure groups have probably fueled increased medical interest and inspired at least a portion of the adverse results. Recommend appropriate changes but at least as follows. On Page 2, lines 1-2: Delete "in parallel" through "in smoking."

Page 2, lines 15-16: Change lines 15 and 16 as follows: BAT holds the view that a scientific and medical controversy exists over the issues of smoking and health and the opinions of eminent scientists on the less popular side of the argument must be recognized. Comment: Statement that the tobacco industry cannot make judgments on medical aspects of smoking conflicts with statements on Page 4 that BAT has many scientific employees working on the medical aspects of smoking. Also dangerous for tobacco industry to abandon judgment on smoking and health to others.
Image Transformation: degradation
ImageRefiner Approach

- **Treat selection of improvement methods as machine learning problem:** classify images according to preferred transformations for improving OCR output
  - Measure characteristics of images
  - Consider a set of candidate image transformations
  - Use a training set with text ground truth provided
    - Evaluate OCR output of each transformation on each image, to determine their effects
    - Classify each image according to its best transform
  - Use Machine Learning to generalize
  - Optionally segment the image and apply above process to individual segments
Previous Work

- **McNamara, Casey, Smith, and Bradburn (1993)**
  - Characteristics based on statistics of runs of black pixels, connected components
  - Select from 2 transformations (thinning, modified thickening) or keep original

- **Cannon, Hochberg, and Kelly (1999): QUARC**
  - Designed for very noisy, old, typewritten documents (fixed width) in English
  - Measure characteristics designed to reflect types of additive noise in fixed-width English documents
  - Select from 13 transformations, or keep original. 4 transformations use “typewriter grid”
Script and Language

- **Script-Specificity**
  - Expect to need script-specific choices of image features and possible transformations, as well as learning.
  - New script requires basis for understanding its marks and OCR issues, in addition to training corpus.
    *Is there a kind of ground truth that could help with this?

- **Language-Specificity**
  - Expect that training may be language-specific, but features and transformations should be applicable across languages in a script.
  - New language requires new training corpus.

- **Currently operates in English, Arabic, Thai**
Areas of Evaluation

- **OCR Accuracy:** string edit distance for training and testing
- **Transformation Selection**
  - Goal: Select the best improvement
  - Effects of selected transformation (improve/degrade, magnitude, comparison with effects of best choice)
- **Segmentation**
  - Evaluate choice of regions
  - Effect of transforming a region on OCR of neighboring regions
- **Generality:** how are results affected by inconsistency between training and application data? By heterogeneity of training data?
OCR Accuracy Evaluation

- Based on string edit distance (Levenshtein)
- Alternative measures
  - Line-Based: Calculate at line level for alignment, then character level for character matching (as in Esakov, Lopresti & Sandberg 1994; Chen 1993)
  - Character-Based: Apply string edit distance directly to sequence of characters in document
  - Both rely on reading order matching up correctly
OCR Accuracy Scoring

- Precision, Recall, Combination

- Desirable characteristics in a score measure:
  - Falls in a defined range
  - Single value
  - Reflects effort to correct and/or errors occurring in recognition
  - More script- or language-specific

- Could be much more precise with ground truth that gives location on page
  - Expensive to produce on real data
  - Issue: trade-offs with synthetic data
Classification Evaluation

- **Core question:** did we select the right operation for the image?
  - Binary evaluations of transformation choice:
    - Is this the best?
    - Is the change an improvement?
  - Continuous value:
    - Distribution of OCR improvement over transformation set
    - How much an improvement does the change present?
Classification Evaluation, ctnd.

- **Issues**
  - Target is the best transformation; other transformations may help
  - Magnitude matters
    - Impact on OCR
    - Difference between impact on OCR and that of ideal choice
  - Expanding to sequences of transformations

- **Derived from OCR evaluation**

- **Indirect measure:** Improvement for other purposes (e.g., human readability) would require additional ground truth.
Segmentation Evaluation

- **Currently applies in testing only:** segmentation approach is not trained.

- **Not conventional segmentation:** seek regions that are consistent in their OCR challenges (noise sources and possibly also font size, style, etc.).

- **Core question:** how good is this segmentation?
  - Infinite number of possible segmentations
  - Quality of segmentation should reflect “purity” of regions and also preference for fewer regions
  - How to account for effects of transforming one region on OCR of neighboring region?
Segmentation Examples: Arabic
Evaluating Generality

- **Question:** How are results affected by inconsistency between training and application data?
  
  *Requires measuring similarity/difference between these data sets*
  
  - Distribution of measured features? What if the relevant similarity/difference is in what the system does not measure?
  - Human judgments of type of noise?
  - Known sources of noise?
  - Source (type of document, age, paper quality …)
  - Language/script

- **Question:** How are results affected by heterogeneity of training data?
  
  *Requires same kind of measurement*

- **Wish list:** ground truth corpus annotated with noise sources, document sources, human judgments of image “quality.”
The research reported in this document/presentation was performed in connection with Contract No. DAAD19-03-C-0059 with the U.S. Army Research Laboratory. The views and conclusions contained in this document/presentation are those of the authors and should not be interpreted as presenting the official policies or position, whether expressed or implied, of the U.S. Army Research Laboratory or the U.S. Government unless so designated by other authorized documents. Citation of manufacturers or trade names does not constitute an official endorsement or approval of the use thereof. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.