Binarization of Poor-Quality Form Images for Handwriting Recognition

Huaigu Cao

hcao3@cubs.buffalo.edu
Outline

• Problem Statement

• Proposed Approach

• Results

• Future Works
Data Set

- Low quality medical forms
  - Noisy carbon copies
  - Text crossing form grids
  - Average word recognition accuracy is 20~30%
Formulation of Binarization Problem

\( x \): binarized image; \( y \): grayscale image

**Objective:**

\[
\hat{x} = \arg \max_x \Pr(x \mid y)
\]

\[
= \arg \max_x \Pr(x, y)
\]

\[
\hat{x} = \sum_x x \cdot \Pr(x \mid y)
\]

**Classic Binarization Problems:**

In the MAP estimation

\[
\hat{x} = \arg \max_x \Pr(y \mid x) \Pr(x)
\]

Assuming \( \Pr(x) \) is constant and the pixels are independent, the binarization problem is converted into the histogram thresholding problem

\[
k = \arg \min_k \sigma^2_W(k)
\]

\[
k = \arg \max_k \sigma^2_B(k)
\]

Niblack’s local binarization

\[
Thr = \text{mean} + \lambda \cdot \text{dev}
\]
Motivation – Using the Markov Random Fields (MRF) for Binarization

$$\hat{x} = \arg \max_x \Pr(y | x) \Pr(x) \quad \text{MAP}$$

$$\hat{x} = \sum_x x \Pr(y | x) \Pr(x) / \Pr(y) \quad \text{MMSE}$$

In addition to binarization using histogram thresholding, $\Pr(x)$ provides constraints of **connectivity and smoothness**

$\Pr(x)$ can be represented by a Markov Random Field under local dependence assumption

Computational Complexity is reduced by the **Belief Propagation (BP)** algorithm (linear time in terms of the size of the image)
Motivation – Ruling Line Removal

$x$: binarized image (the MRF)

$y$: grayscale image (the observation)

\[ y = [y_v, y_i], \]

$y_v$: visible observation;

$y_i$: invisible observation

**Objective:**

\[ \hat{x} = \arg \max_x \Pr(x \mid y_v) \quad \text{MAP} \]

or

\[ \hat{x} = \sum_x x \cdot \Pr(x \mid y_v) \quad \text{MMSE} \]
Outline

• Problem Statement

• Proposed Approach

• Results

• Future Works
Topology of the MRF

• Patch-based topology
  – x and y are divided into 5x5 non-overlapping blocks
  – Each patch has $2^{25}$ possible states

• Computational issue
  – Computational Complexity is reduced by the **Belief Propagation (BP)** algorithm (linear time in the size of the image; but quadratic time in the number of states)
  – VQ and pruning are used for reducing the set of states

• Pair-wise prior probability
  – learned from clean samples of handwriting

• Observation density
  – Represented by local grayscale histogram
Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works
Results of MRF Binarization and Ruling-Line Removal

Binarization

Ruling-line Removal

Input

MRF

Niblack

Otsu

May 08, 2008

Center for Unified Biometrics and Sensors

University at Buffalo The State University of New York
Performance of MRF Binarization and Ruling-Line Removal

<table>
<thead>
<tr>
<th>Method</th>
<th>Milewski</th>
<th>MRF</th>
<th>Niblack</th>
<th>Otsu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set #1</td>
<td>Top 1 rate</td>
<td>17.5%</td>
<td>25.9%</td>
<td>19.4%</td>
</tr>
<tr>
<td></td>
<td>Top 2 rate</td>
<td>24.4%</td>
<td>36.6%</td>
<td>26.9%</td>
</tr>
<tr>
<td></td>
<td>Top 5 rate</td>
<td>33.4%</td>
<td>44.9%</td>
<td>35.9%</td>
</tr>
<tr>
<td></td>
<td>Top 10 rate</td>
<td>39.6%</td>
<td>51.7%</td>
<td>42.3%</td>
</tr>
<tr>
<td>Set #2</td>
<td>Top 1 rate</td>
<td>19.5%</td>
<td>30.3%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Top 2 rate</td>
<td>28.1%</td>
<td>40.7%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Top 5 rate</td>
<td>37.6%</td>
<td>52.7%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Top 10 rate</td>
<td>45.0%</td>
<td>60.0%</td>
<td>NA</td>
</tr>
<tr>
<td>Overall</td>
<td>Top 1 rate</td>
<td>18.7%</td>
<td>28.6%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Top 2 rate</td>
<td>26.7%</td>
<td>39.1%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Top 5 rate</td>
<td>36.0%</td>
<td>49.7%</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>Top 10 rate</td>
<td>42.9%</td>
<td>56.8%</td>
<td>NA</td>
</tr>
</tbody>
</table>

Data: carbon copies of PCR handwritten forms

* Set #1 does not require line removal
** Set #2 requires line removal
Outline

• Problem Statement
• Proposed Approach
• Results
• Future Works
Future Works

• Practical Issues
  – Adaptive selection of model according to the size of text
  – Automatic ruling line detection