Binarization of Poor-Quality Form Images for Handwriting Recognition

Huaigu Cao

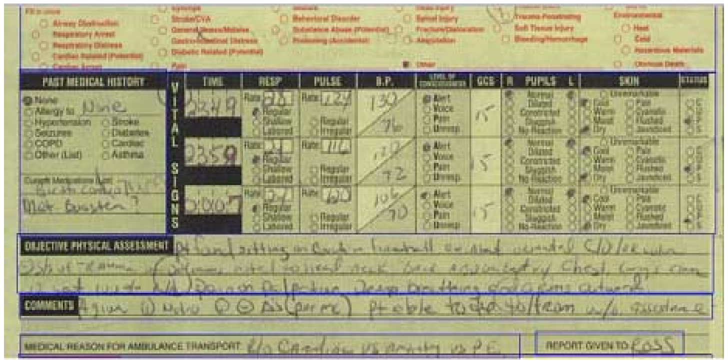
hcao3@cubs.buffalo.edu

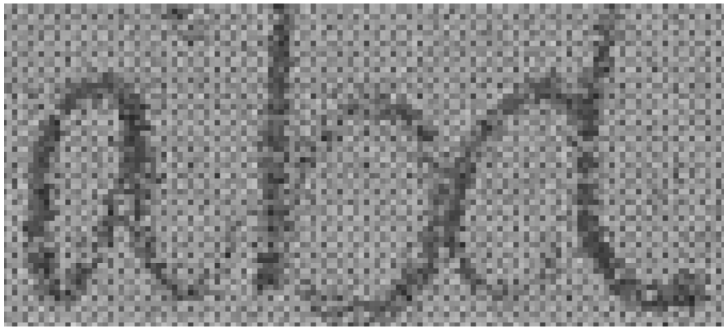
Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works

Data Set

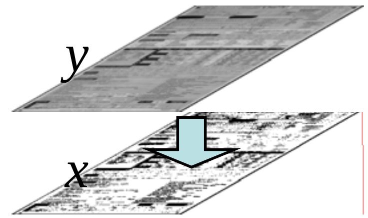
- Low quality medical forms
  - Noisy carbon copies
  - Text crossing form grids
  - Average word recognition accuracy is 20~30%





Formulization of Binarization Problem

x: binarized image; y: grayscale image



Objective:

$$\hat{x} = \arg\max_x \Pr(x \mid y) = \arg\max_x \Pr(x, y) \quad \text{MAP} \qquad \text{or} \qquad \hat{x} = \sum_x x \cdot \Pr(x \mid y) \quad \text{MMSE}$$

Classic Binarization Problems:

In the MAP estimation

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

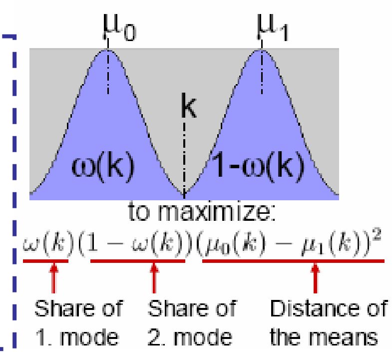
Assuming Pr(x) is constant and the pixels are independent, the binarization problem is converted into the histogram thresholding problem

Otsu

$$k = \arg\min_k \sigma_W^2(k)$$

Fisher

$$k = \arg\max_k \sigma_B^2(k)$$



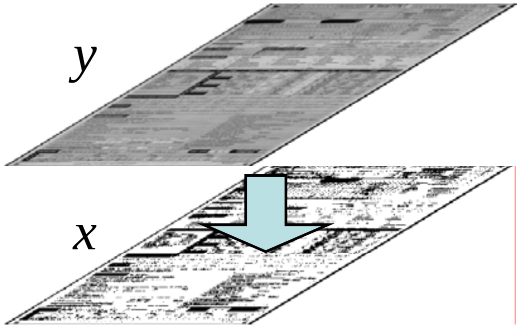
Niblack's local binarization

$$Thr = mean + \lambda \cdot dev$$

Motivation – Using the Markov Random Fields (MRF) for Binarization

$$\hat{x} = \arg\max_{x} \Pr(y \mid x)\Pr(x) \qquad \text{MAP}$$

$$\hat{x} = \sum_{x} x \Pr(y \mid x)\Pr(x)/\Pr(y) \qquad \text{MMSE}$$



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

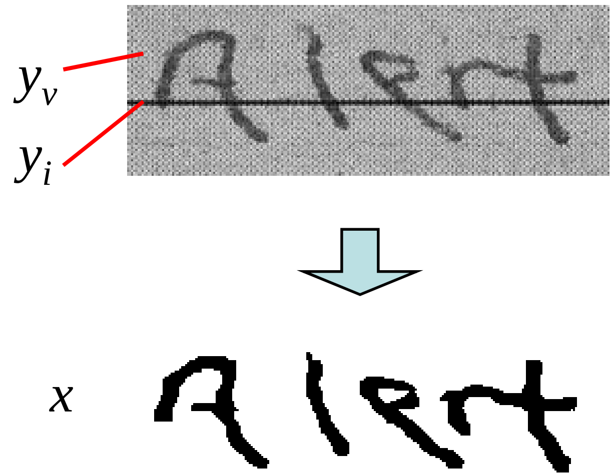
x: binarized image (the MRF)

y: grayscale image (the observation)

$$y = [y_v, y_i],$$

$$y_v : \text{visible observation;}$$

$$y_i : \text{invisible observation}$$



Objective:

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

or

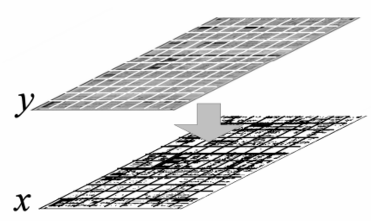
$$\hat{x} = \sum_x x \cdot \Pr(x \mid y_v) \quad \text{MMSE}$$

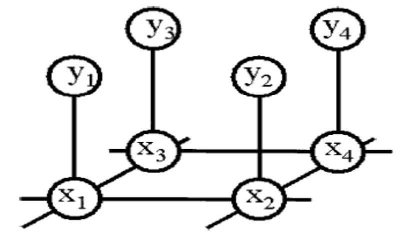
Outline

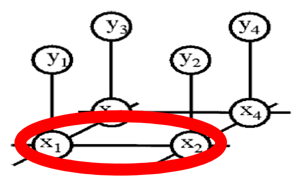
- Problem Statement
- Proposed Approach
- Results
- Future Works

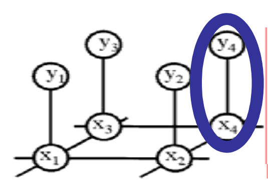
Topology of the MRF

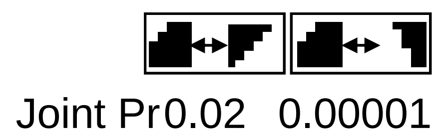
- Patch-based topology
  - x and y are divided into 5x5 non-overlapping blocks
  - Each patch has 2^{25} possible states
- Computational issue
  - Computational Complexity is reduced by the **Belief Propagation (BP)** algorithm (linear time in the size of the image; but quadratic time in the number of 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

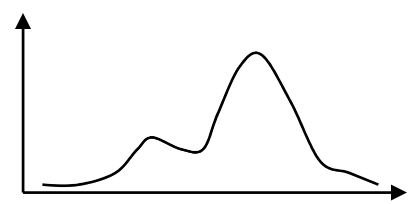










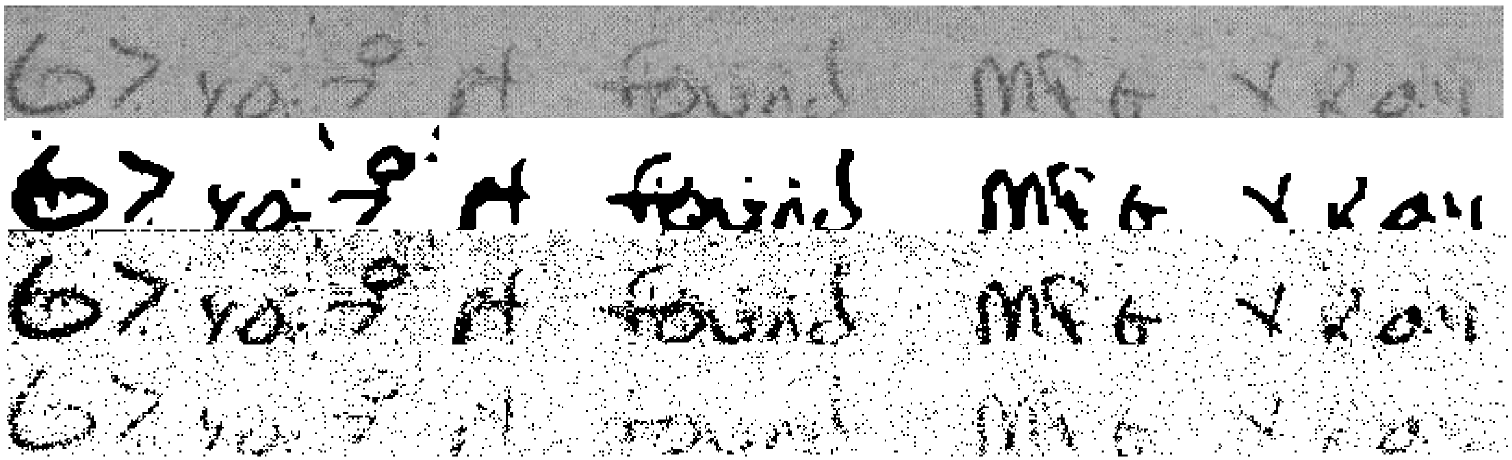


Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works

Results of MRF Binarization and Ruling-Line Removal

Binarization



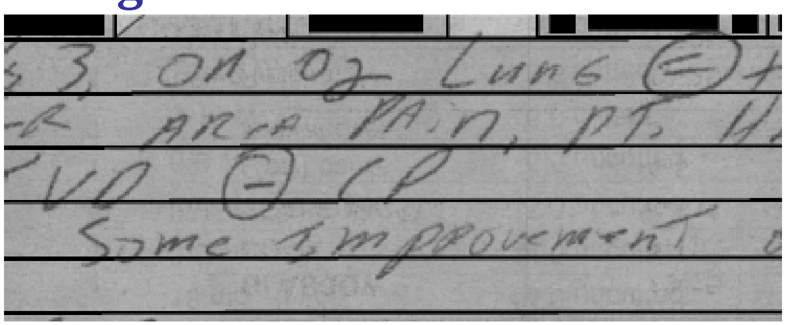
Input

MRF

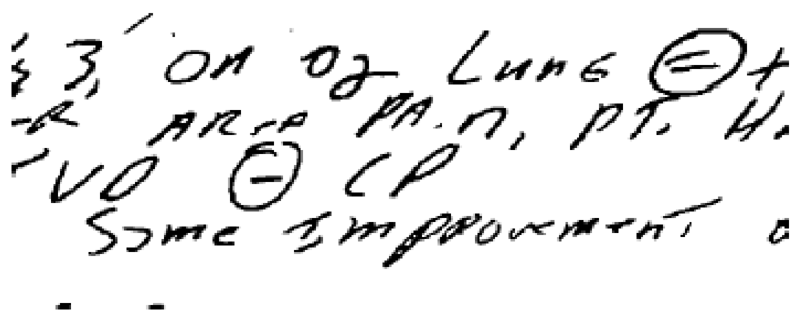
Niblack

Otsu

Ruling-line Removal



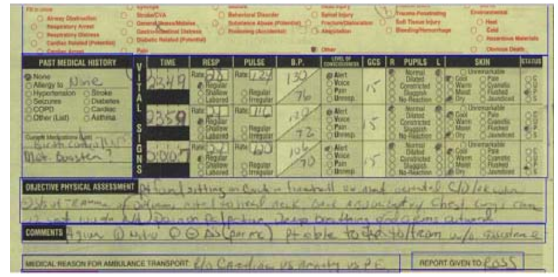
Input

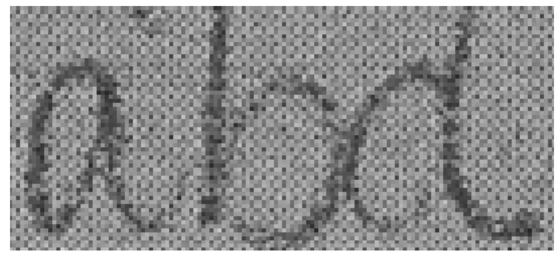


MRF

Performance of MRF Binarization and Ruling-Line Removal

Method		Milewski	**MRF**	Niblack	Otsu
Set #1	Top 1 rate	17.5%	**25.9%**	19.4%	11.6%
	Top 2 rate	24.4%	**36.6%**	26.9%	16.0%
	Top 5 rate	33.4%	**44.9%**	35.9%	23.3%
	Top 10 rate	39.6%	**51.7%**	42.3%	28.8%
Set #2	Top 1 rate	19.5%	**30.3%**	NA	NA
	Top 2 rate	28.1%	**40.7%**	NA	NA
	Top 5 rate	37.6%	**52.7%**	NA	NA
	Top 10 rate	45.0%	**60.0%**	NA	NA
Overall	Top 1 rate	18.7%	**28.6%**	NA	NA
	Top 2 rate	26.7%	**39.1%**	NA	NA
	Top 5 rate	36.0%	**49.7%**	NA	NA
	Top 10 rate	42.9%	**56.8%**	NA	NA





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