



cse@buffalo

Binarization of Poor-Quality Form Images for Handwriting Recognition

Huaigu Cao

hcao3@cubs.buffalo.edu



.cse@buffalo

Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works



Data Set

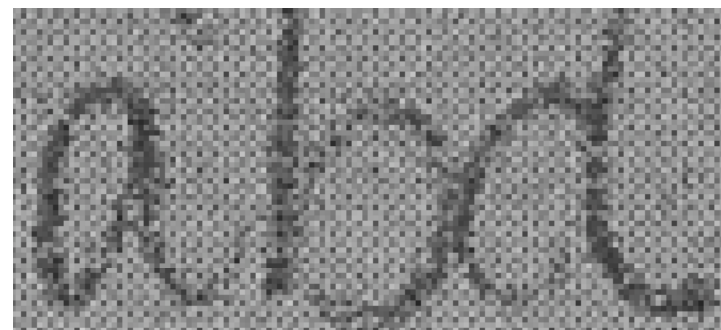
- Low quality medical forms
 - Noisy carbon copies
 - Text crossing form grids
 - Average word recognition accuracy is 20~30%

TIME	RESP	PULSE	B.P.	TEMP	GCE	R	PUPILS	L	SKIN	STATUS
0249	Rate 22 Regular Labs	Rate 122 Regular Labs	130 76		15		Alert Hazy Constricted Slightly No Reaction		Alert Hazy Constricted Slightly No Reaction	Unconscious Pup Constricted Hazy Anisocoria
0359	Rate 11 Regular Labs	Rate 117 Regular Labs	112 72		5		Alert Hazy Constricted Slightly No Reaction		Alert Hazy Constricted Slightly No Reaction	Unconscious Pup Constricted Hazy Anisocoria
0407	Rate 11 Regular Labs	Rate 106 Regular Labs	106 70		5		Alert Hazy Constricted Slightly No Reaction		Alert Hazy Constricted Slightly No Reaction	Unconscious Pup Constricted Hazy Anisocoria

OBJECTIVE PHYSICAL ASSESSMENT: Pt lying on back in hospital bed. Head turned to the left. Pupils are equal and reactive to light. Chest is clear. Heart is normal. Abdomen is soft and nontender. Deep tendon reflexes are normal.

COMMENTS: Pt is lying on back in hospital bed. Head turned to the left. Pupils are equal and reactive to light. Chest is clear. Heart is normal. Abdomen is soft and nontender. Deep tendon reflexes are normal.

MEDICAL REASON FOR AMBULANCE TRANSPORT: Pt. Cardiac vs. anxiety vs. PE. REPORT GIVEN TO: BASS

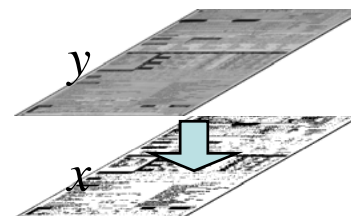




Formulization of Binarization Problem

cse@buffalo

x : binarized image; y : grayscale image



Objective:

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

MAP

or

$$\hat{x} = \sum_x x \cdot \Pr(x | y)$$
 MMSE

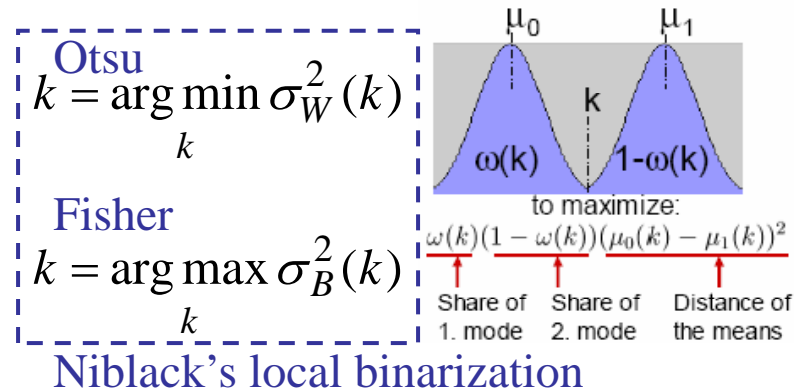
$$= \arg \max_x \Pr(x, y)$$

Classic Binarization Problems:

In the MAP estimation

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

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

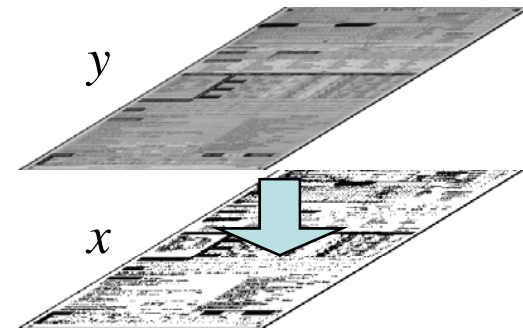




Motivation – Using the Markov Random Fields (MRF) for Binarization

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

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



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)



cse@buffalo

Motivation – Ruling Line Removal

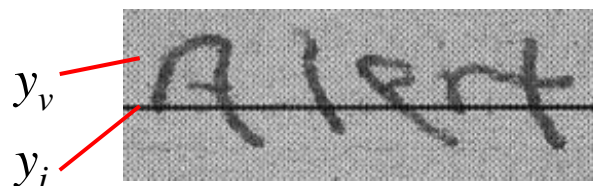
x : binarized image (the MRF)

y : grayscale image (the observation)

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

y_v : visible observation;

y_i : invisible observation



Objective:

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

or

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



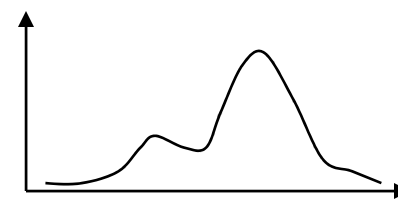
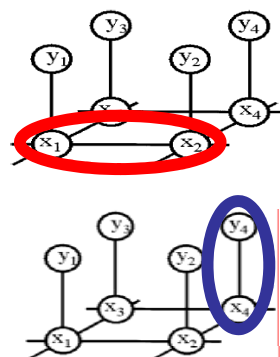
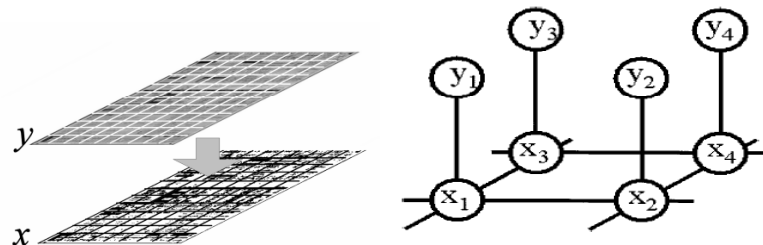
.cse@buffalo

Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works

Topology of the MRF

- Patch-based topology
 - x and y are divided into 5×5 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





cse@buffalo

Outline

- Problem Statement
- Proposed Approach
- Results
- Future Works



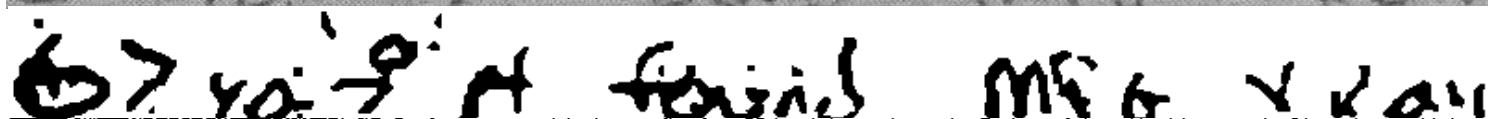
cse@buffalo

Results of MRF Binarization and Ruling-Line Removal

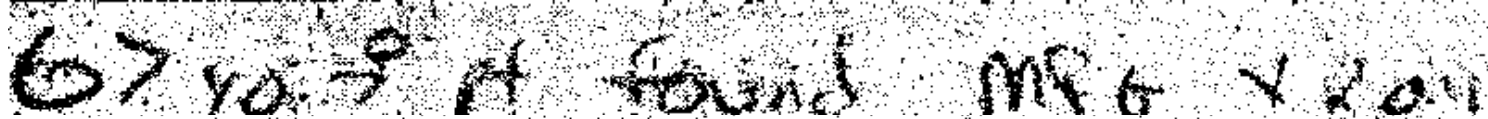
Binarization



Input



MRF

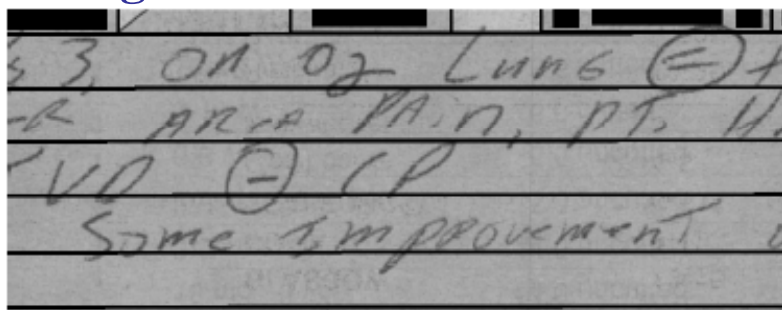


Niblack

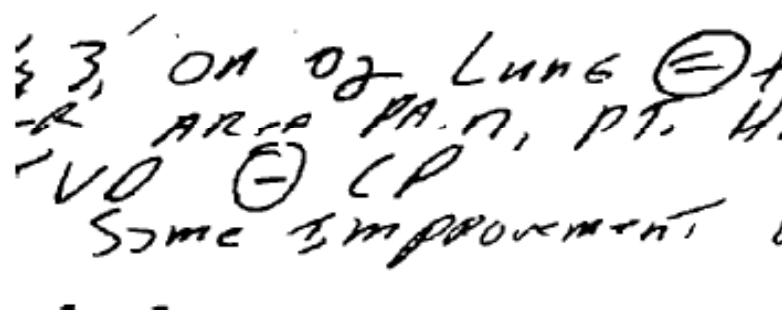


Otsu

Ruling-line Removal



Input

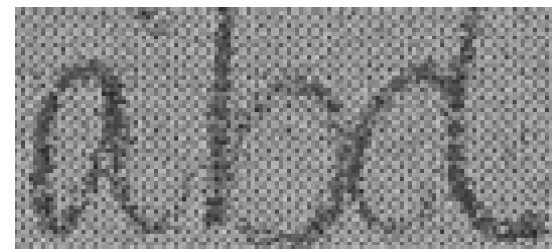
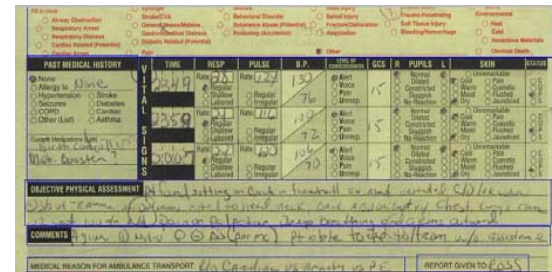


MRF



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



cse@buffalo

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