Text segmentation for MRC Document compression using a Markov Random Field model

Final dissertation by Eri Haneda

Advisory committee:

Prof. Charles Bouman (Chair)

Prof. Jan Allebach

Prof. Peter Doerschuk

Prof. George Chiu

February 25th, 2011

Primary contributions

Primary projects

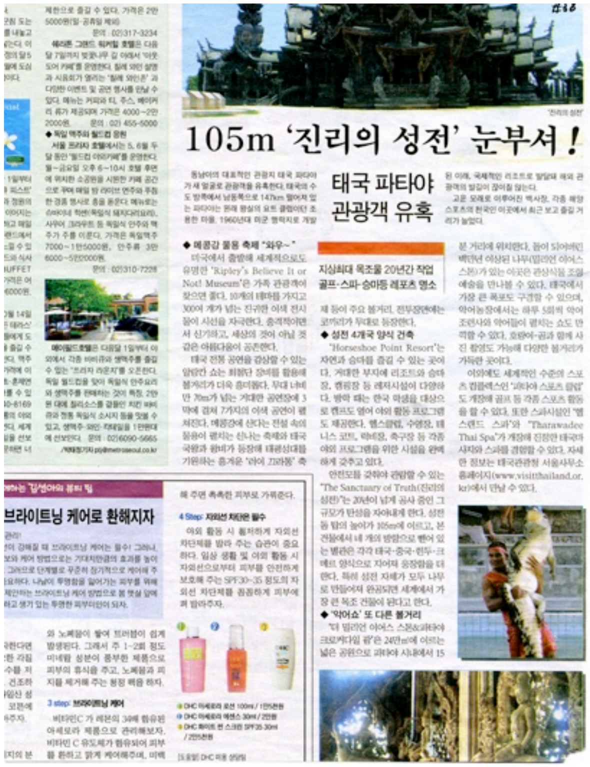
1) Text segmentation for MRC document with Samsung Co. Ltd.
   - □ Multiscale-COS/CCC algorithm

2) Next generation image capture device development with Samsung Co. Ltd.
   - □ Motion/Lamp control
   - □ Scanner image quality bench mark
   - □ Snap-to-White contrast enhancement

3) CT baggage reconstruction (In progress)
   - □ Multislice helical scan CT reconstruction code development
   - □ Image reconstruction using a substitute prior model

Brief summary of snap-to-white project

- The background of scanned images sometimes appears too dark because of the paper material (e.g. newspapers)

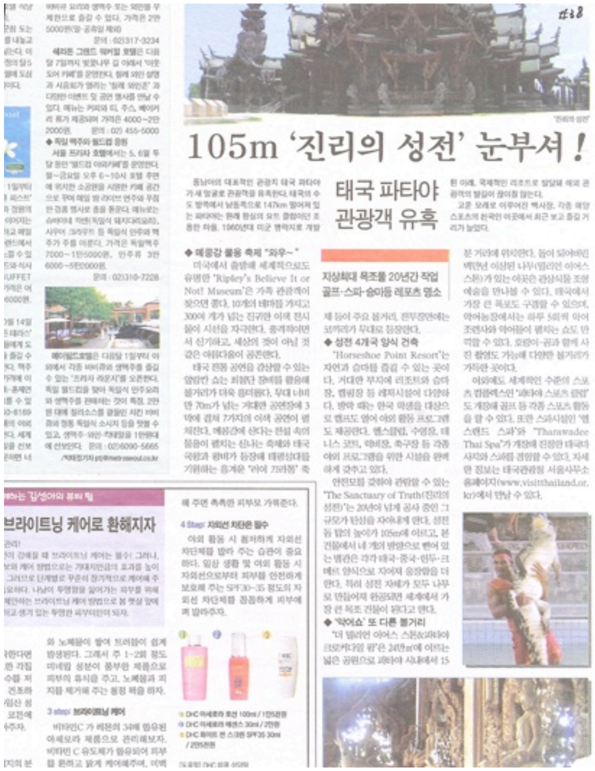


105m '진리의 성전' 눈부셔!

태국 파타야 관광객 유혹

브라이트닝 케어로 환해지자

EPSON

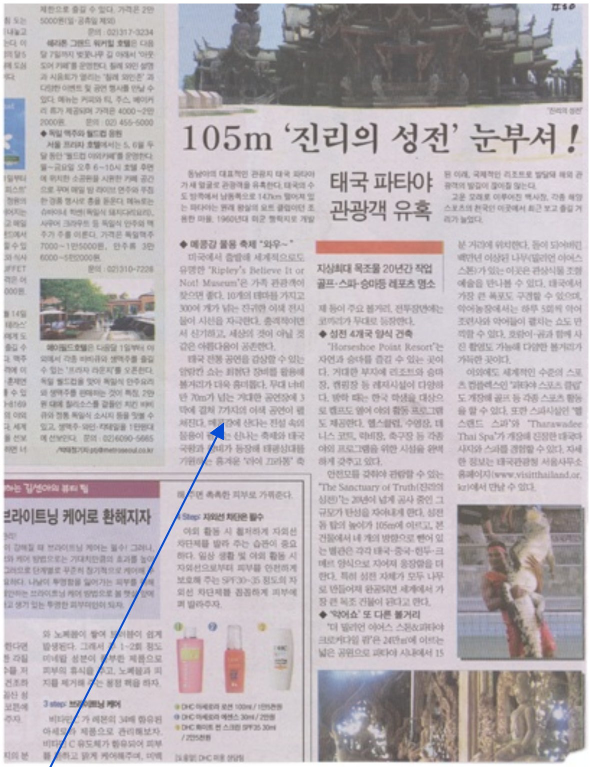


105m '진리의 성전' 눈부셔!

태국 파타야 관광객 유혹

브라이트닝 케어로 환해지자

HP



105m '진리의 성전' 눈부셔!

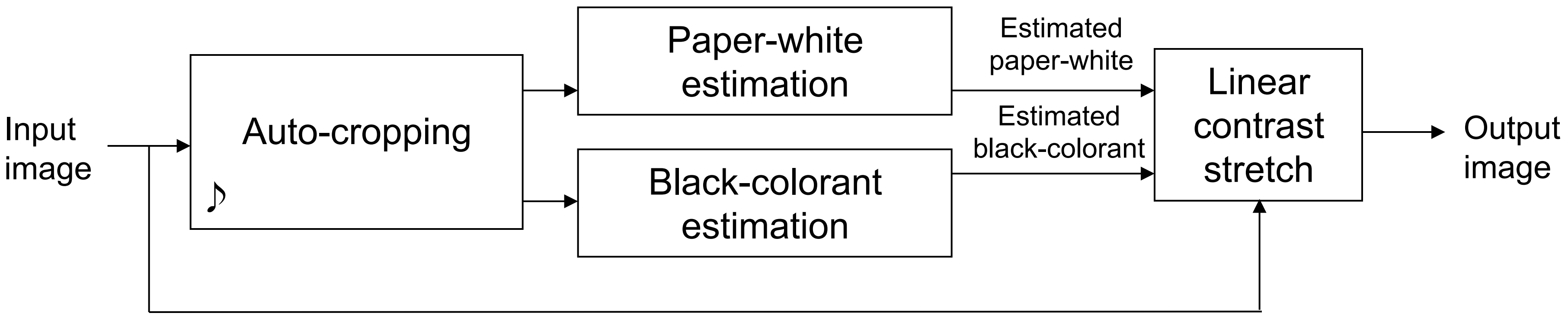
태국 파타야 관광객 유혹

브라이트닝 케어로 환해지자

Samsung

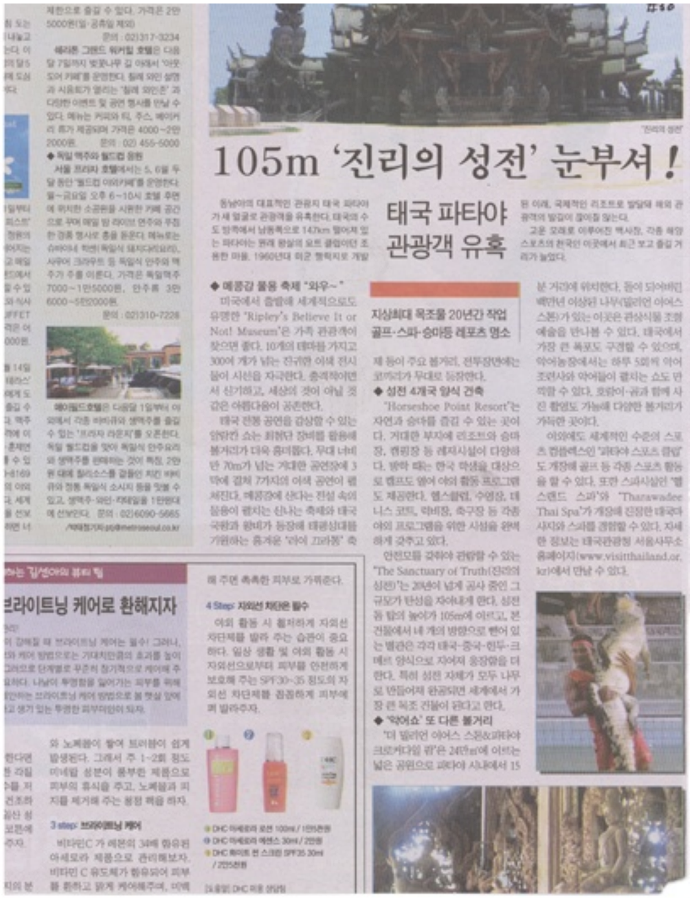
- Low color contrast
- Background is not white

Overview



- Auto-cropping removes extra white or black which is sometimes included around the borders of the image
- Paper-white/black-colorant estimation determines respective RGB values using extrema and region growing
- Linear contrast stretch snaps the estimated paper-white and black-colorant to the largest and smallest encoded value to increase the dynamic range

Original/Adjusted Samsung



105m '진리의 성전' 눈부셔!

태국 파타야 관광객 유혹

◆ 메콩강 물용 축제 "와우~"

지상최대 목조물 20년간 작업 골프·스파·승마등 레포츠 명소

◆ 성전 4개국 양식 건축

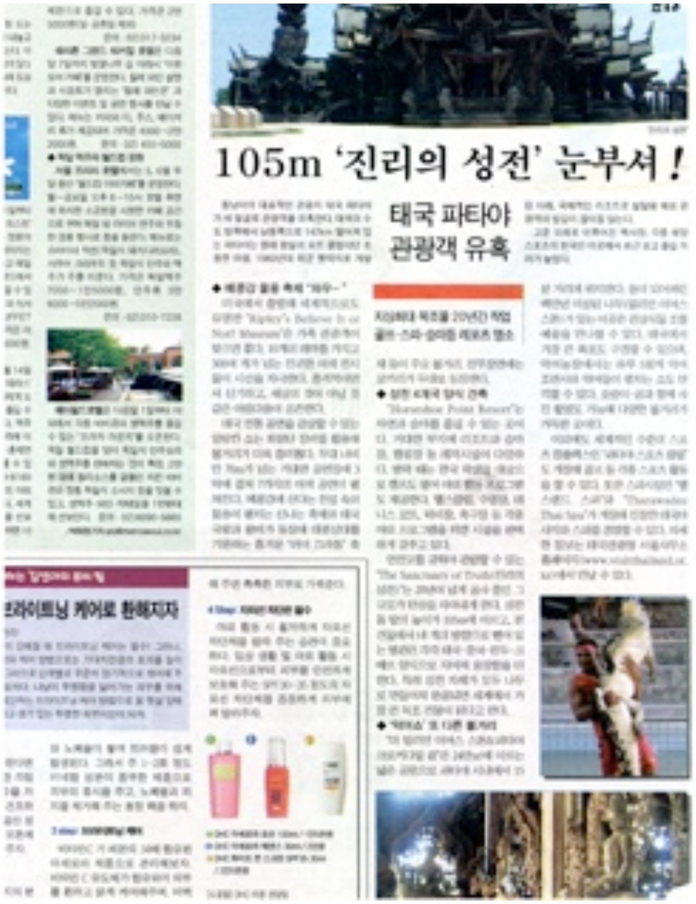
◆ '악어쇼' 또 다른 볼거리

브라이트닝 케어로 환해지자

4 Step: 자외선 차단은 필수

3 step: 브라이트닝 케어

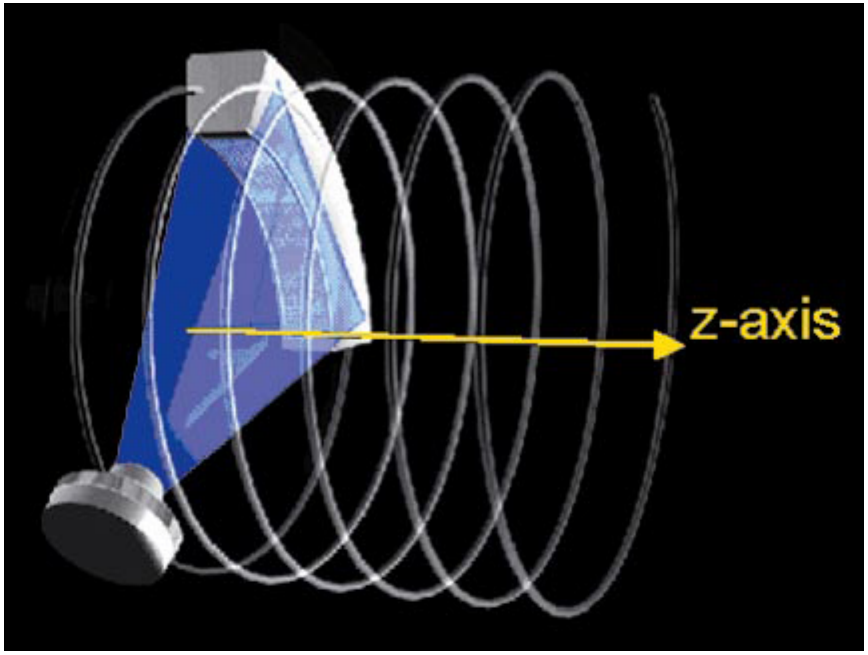
Original Samsung



Adjusted Samsung

Pilot research for CT baggage reconstruction #1

- Short-term goals
  - □ Write preliminary multislice helical CT reconstruction code
  - □ Develop a statistical model for accurate CT baggage reconstruction/segmentation



Multislice helical CT (Source & Detector)
"Multislice CT, M.F. Reiser (2004)"

- Forward projection

$$\underset{\text{Sinogram (Projection)}}{\begin{bmatrix} y \end{bmatrix}} = \begin{bmatrix} \quad A \quad \end{bmatrix} \underset{\text{Image voxels}}{\begin{bmatrix} x \end{bmatrix}}$$

- Reconstruction

$$\hat{x}_{MAP} = \arg\max_{x} \left\{ -\frac{1}{2}(y - Ax)^{T} \underset{\text{Weighting}}{D}(y - Ax) + \underset{\text{Smoothing}}{U(x)} \right\}$$

Pilot research for CT baggage reconstruction #2

- We are trying to develop a new generalized prior model for MAP estimate
- In general, the prior term *p(x)* needs to be modeled under strong assumption

$$\widehat{X}_{MAP} = \arg\max_{x \ge 0} \left\{ -\frac{1}{2}(y - Ax)^T D(y - Ax) + \log p(x) \right\}$$

- Our approach is to generate a second order polynomial approximation to the function log *p(x)* about an initial image x′ with the form

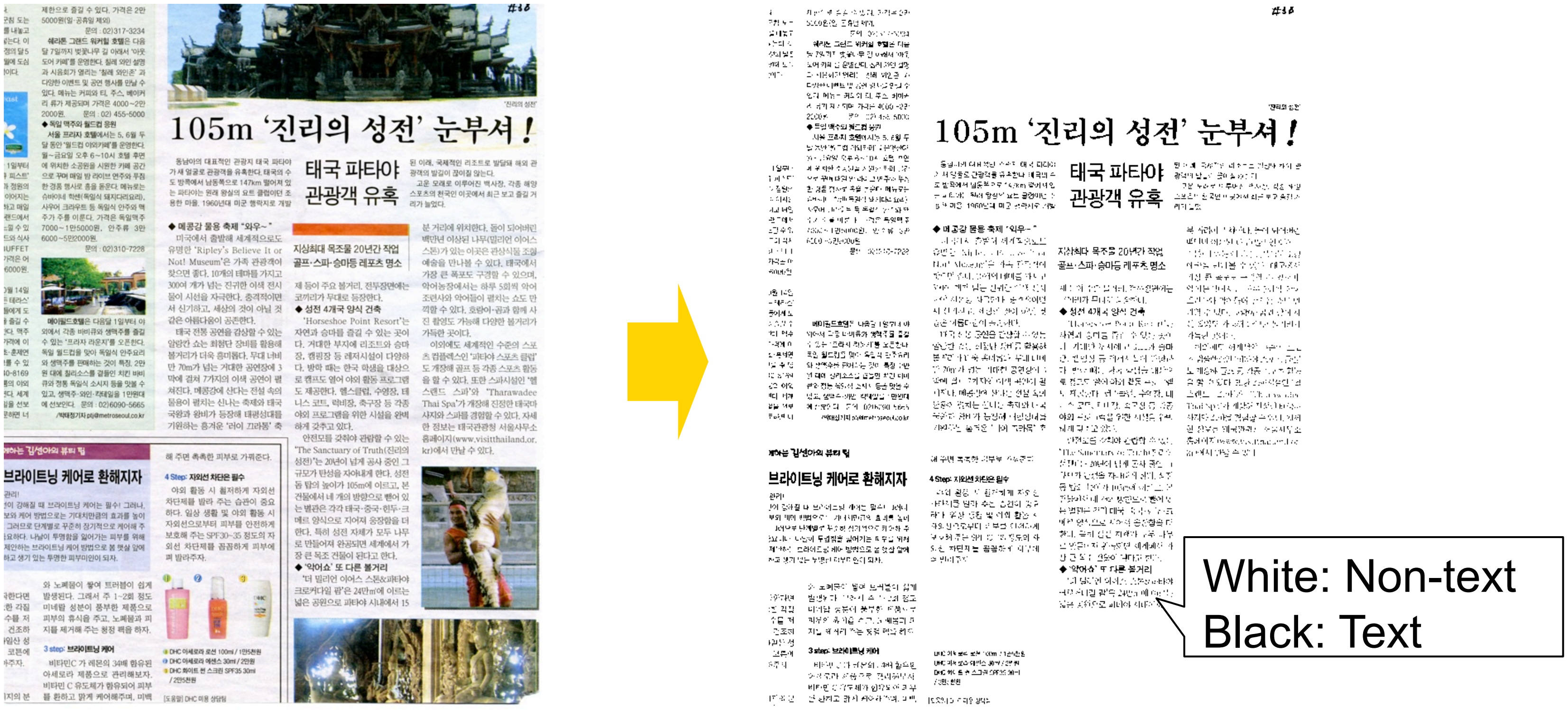
$$\begin{aligned} \log p(x) &\ge f(x; x') \\ &= -\frac{1}{2}(x - x')B(x - x') + d^t(x - x') + c \end{aligned}$$

If we assume dependencies only on neighbors, the expression is dramatically simplified

Text segmentation for MRC Document compression using a Markov Random Field model

What is text segmentation?

- Text segmentation is extracting text components from a document



Input image (Color)

Segmentation result (Binary)

Fig. 1. An example of text segmentation

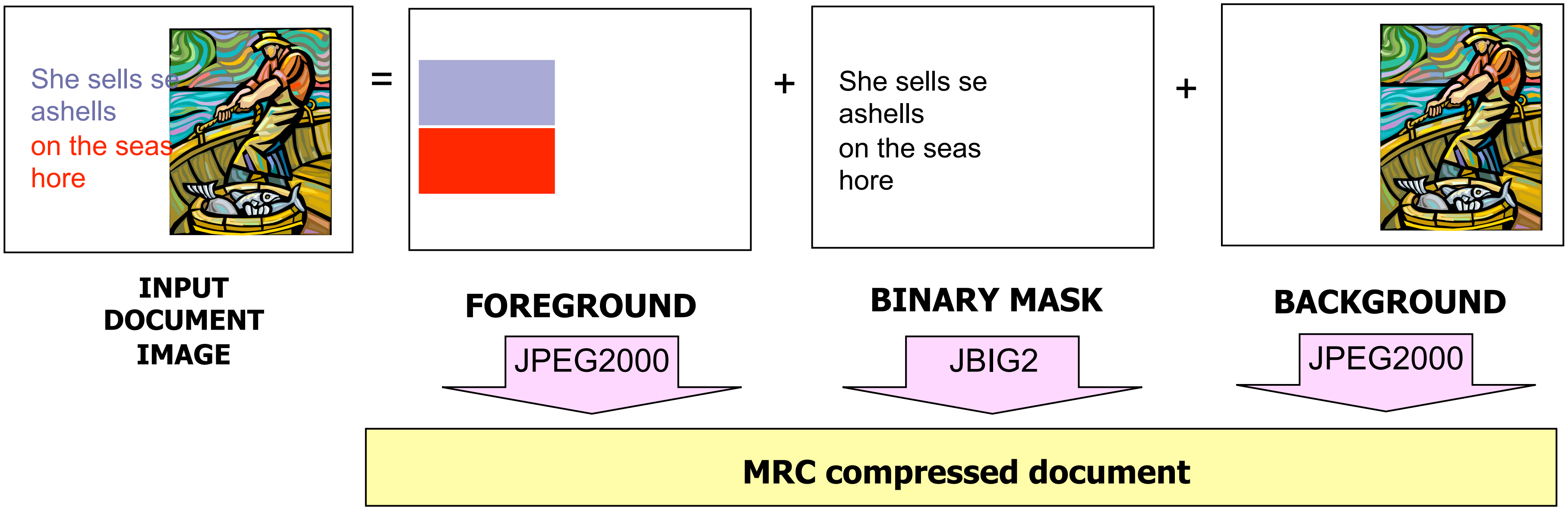
Why is text segmentation useful?

- Useful for layer based document compression
  - □ Layer based document compression is defined in ITU-T. T.44, Mixed Raster Content (MRC) encoding
  - □ Good text segmentation achieves high document compression and preserves high image quality
- Useful for other applications such as OCR etc.

- Our goal is to generate segmentation which is:
  - □ Accurate
  - □ Robust
  - □ Computationally inexpensive

Motivations: Mixed Raster Content

- Mixed Raster Content (MRC) is a standard for layer based document compression defined in ITU-T. T.44



She sells se
ashells
on the seas
hore

=

+

She sells se
ashells
on the seas
hore

+

INPUT DOCUMENT IMAGE

FOREGROUND

JPEG2000

BINARY MASK

JBIG2

BACKGROUND

JPEG2000

MRC compressed document

Fig. 2. Illustration of MRC document compression

Past text segmentation research

- Top-down/Bottom-Up approach
  - X-Y cut algorithm [1], Run Length Smearing Algorithm [2] (RLSA)
- Thresholding approach
  - Otsu, Niblack, Sauvola, Kapur, and Tsai method [3]
- Statistical approach
  - Hidden Markov Model (HMM)
    - The most known commercial software DjVu uses HMM model [4]
  - Markov Random Field (MRF)
    - Zhen et al. used an MRF model to exploit the contextual information for noise removal [5]
    - *Kumar and Kuk also incorporates their proposed prior model to MAP-MRF text segmentation*
  - Conditional Random Field (CRF)
    - New emerging model, originally proposed by Lafferty [6]
    - It directly models the posterior distribution of labels given observations
    - It has been applied to pixel-wise segmentation

Research accomplishment

- □ Three novel algorithms have been developed

  - ■ COS algorithm (Cost Optimized Segmentation)
    Used for initial segmentation. Formulated in a global cost optimization framework.

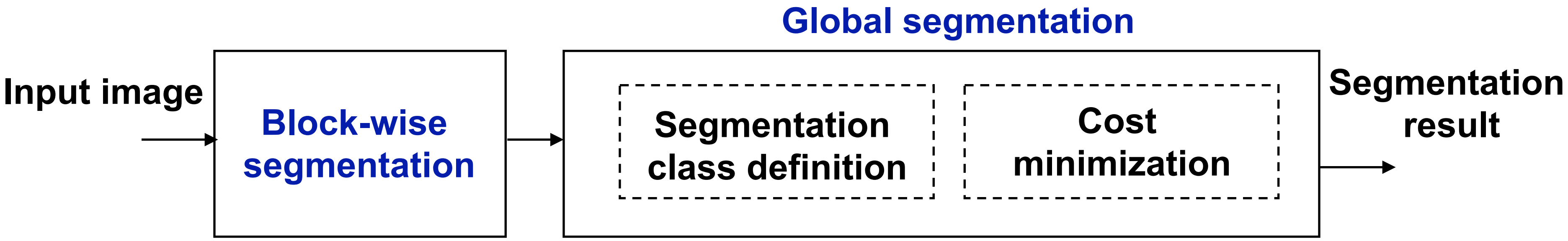
Post-Prelim

  - ■ CCC algorithm (Connected Component Classification)
    Refines segmentation by classifying connected components into text and non-text using a Markov Random Field (MRF) model

Post-Prelim

  - ■ Multiscale-COS/CCC algorithm
    Comprehensive segmentation scheme using multiple resolutions. This improves simultaneous detection of large and small text.

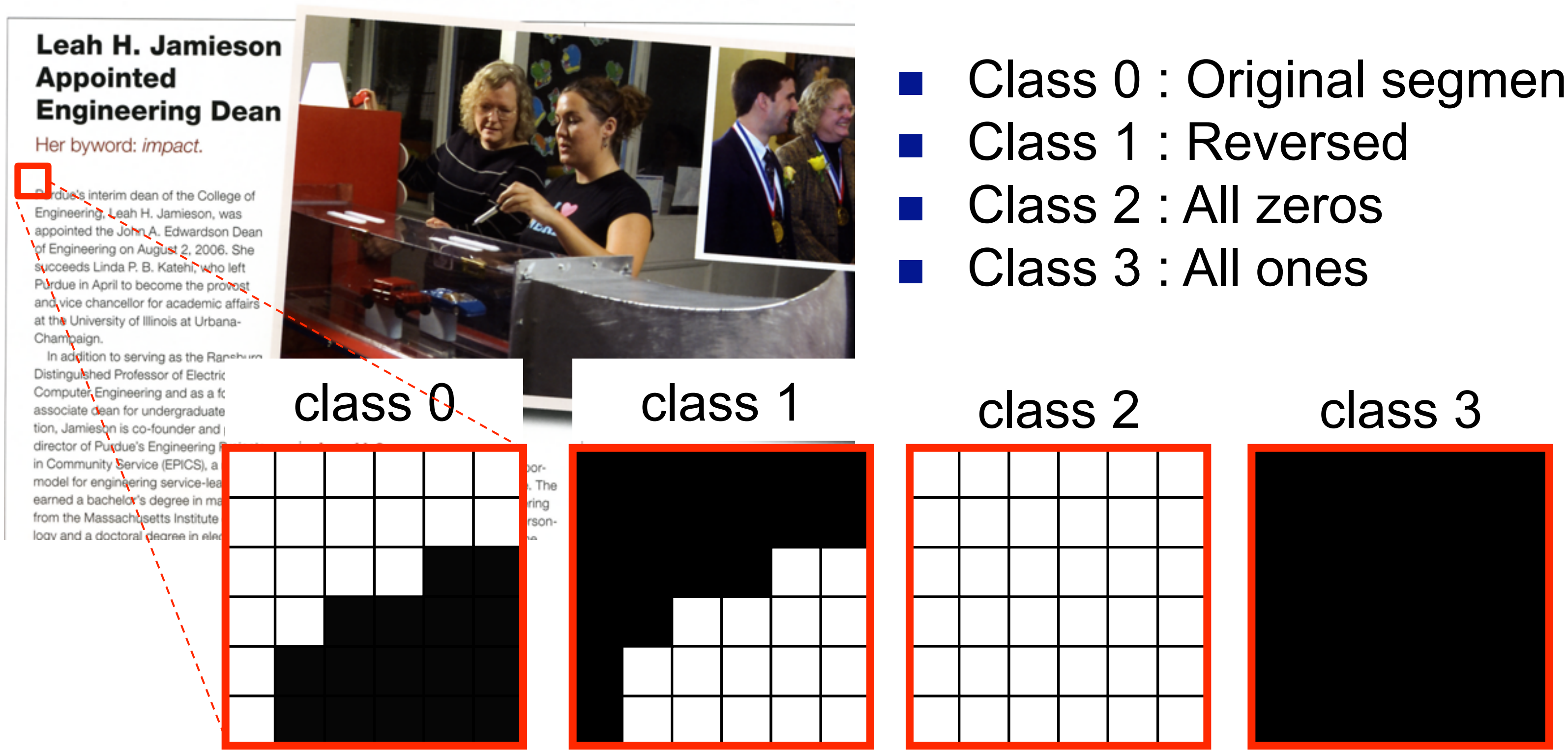
COS algorithm flowchart



- Block-wise segmentation
  - Image is divided into overlapping blocks
  - Each block segmented independently using clustering procedure
- Global segmentation
  - Four possible classes are defined for each block
  - The class of each block is chosen to minimize a global cost

Segmentation class definition

- After initial block segmentation, four possible classes are defined for each block



Leah H. Jamieson Appointed Engineering Dean

Her byword: *impact.*

Purdue's interim dean of the College of
Engineering, Leah H. Jamieson, was
appointed the John A. Edwardson Dean
of Engineering on August 2, 2006. She
succeeds Linda P. B. Katehi, who left
Purdue in April to become the provost
and vice chancellor for academic affairs
at the University of Illinois at Urbana-
Champaign.
In addition to serving as the Ransburg
Distinguished Professor of Electric
Computer Engineering and as a fo
associate dean for undergraduate
tion, Jamieson is co-founder and
director of Purdue's Engineering
in Community Service (EPICS), a
model for engineering service-lea
earned a bachelor's degree in ma
from the Massachusetts Institute
logy and a doctoral degree in elec

por-
. The
ring
rson-

- Class 0 : Original segmentation
- Class 1 : Reversed
- Class 2 : All zeros
- Class 3 : All ones

Black = Foreground **White = Background**

Cost minimization

$$Cost(S)=\sum_{i=0}^{M}\sum_{j=0}^{N}\left\{E(s_{i,j})+\lambda_1 V_1(s_{i,j-1},s_{i,j})+\lambda_2 V_2(s_{i-1,j},s_{i,j})+\lambda_3 V_3(s_{i,j})\right\}$$

$s_{i,j}$: Class of block at location (*i,j*). $S=\{s_{i,j}\}$

E : Total variance of gray levels of each group (0 or 1)

V_1 : Number of mismatches in horizontal overlap region

V_2 : Number of mismatches in vertical overlap region

V_3 : Number of ‘1’ pixels inside block

λ_k : Weight coefficients, *k*=1,2,3

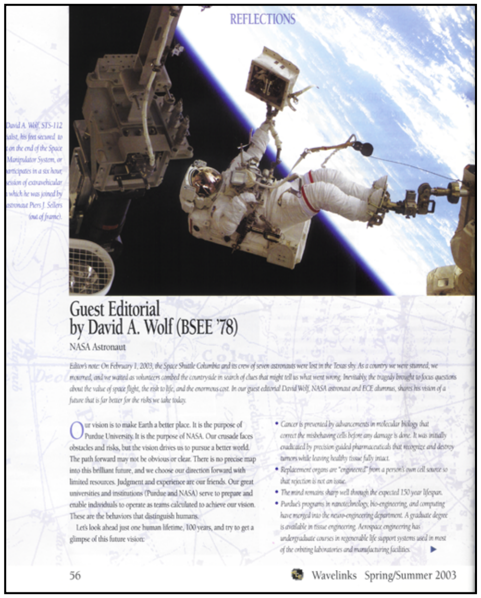
- Cost function may be minimized using dynamic programming

Cost Optimized Segmentation problems

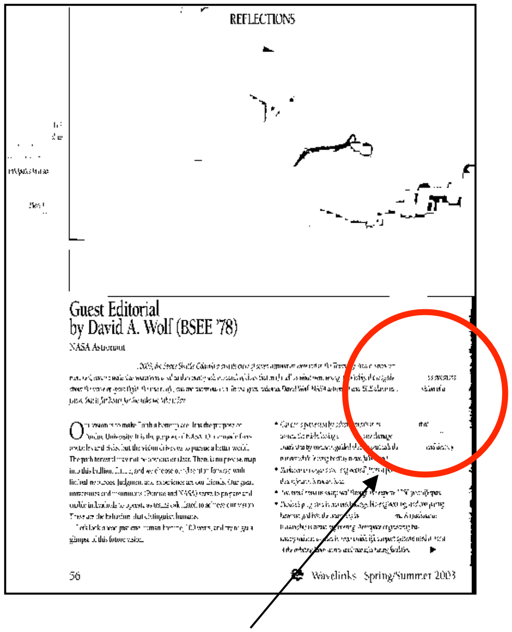
- Optimal parameters $\{\lambda_k\}$ determined by minimizing weighted error

$$error = (1-\omega)N_{missed} + \omega N_{false} \qquad \omega \in [0,1]$$

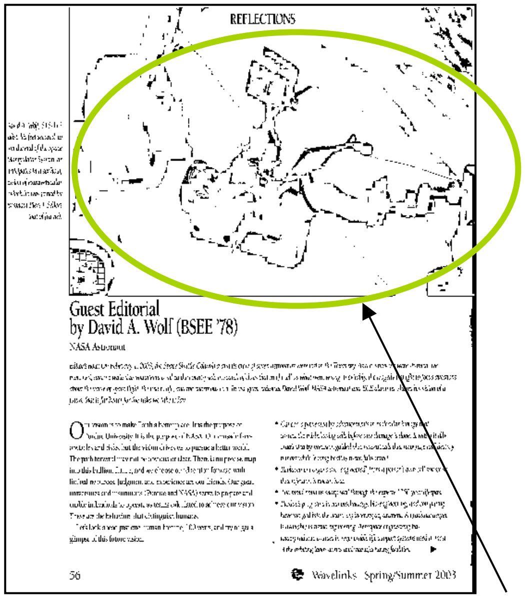
Original



Errors equally weighted w=0.5



Greater weighting on missed detections, w=0.09



Missed detections

False detections

Approach: Minimize missed detections, and eliminate false detections in a later stage. ⇨ Motivation for CCC algorithm

Connected component classification (CCC)

- Refines the COS results by eliminating non-text components using a Markov Random Field (MRF) model

- Classification procedure

  Step1: Extract connected component CC_i

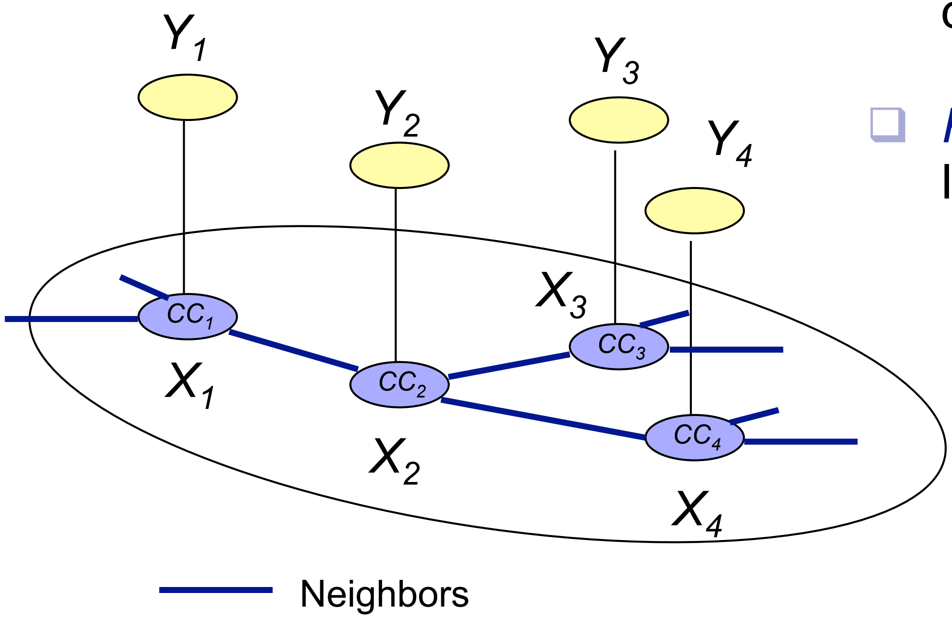
  Step2: Calculate feature vector Y_i

  Step3: Each CC_i is classified as either text (X_i=1) or non-text (X_i=0) using MRF-MAP framework

CCC statistical model

- Classification of each $x_i \in \{0,1\}$ determined by maximizing posterior density (MAP)

$$\arg\max_{x\in\{0,1\}^N}\{\log p(x \mid y)\} = \arg\max_{x\in\{0,1\}^N}\{\log p(y \mid x) + \log p(x)\}$$



- ❑ *Data term p(ylx)* assumed to be conditionally independent
- ❑ *Prior term p(x)* for true segmentation labels is modeled by an MRF

$Y=\{Y_1, Y_2, \ldots Y_N\}$
~ Observed data (feature vectors)

$X=\{X_1, X_2, \ldots X_N\}$
~ Classification of CC

Data model, p(y|x)

- Feature vector, Y_i
  - □ Boundary **edge depth** statistics
  - □ **Color uniformity**
- Y_i are conditionally independent given associated X_i

$$p(y \mid x) = \prod_{i=1}^{N} p(y_i \mid x_i)$$

- Feature vector for both text and non-text modeled as a multivariate Gaussian mixture

$$p(y_i \mid x_i = k) = \sum_{m=0}^{M_k - 1} \frac{a_{k,m}}{(2\pi)^{D/2}} \left| R_{k,m} \right|^{-1/2} \exp\left\{ -\frac{1}{2} (y_i - \mu_{k,m})^t R_{k,m}^{-1} (y_i - \mu_{k,m}) \right\}$$

$k \in \{0,1\}$: class label

M_0, M_1 : number of sub-clusters in each Gaussian mixture

Prior model, p(x)

- MRF used to model local interaction between neighboring elements
- An MRF is a density satisfying the Markov property:

$$p(x_s \mid x_r\ for\ r \neq s) = p(x_s \mid x_{\partial s})$$

- MRF may be expressed as a Gibbs distribution (Hammersley-Clifford theorem):

$$p(x) = \frac{1}{Z} \exp\left\{ -\frac{1}{T} \sum_{c \in C} V_c(x) \right\}$$

Z : normalization factor

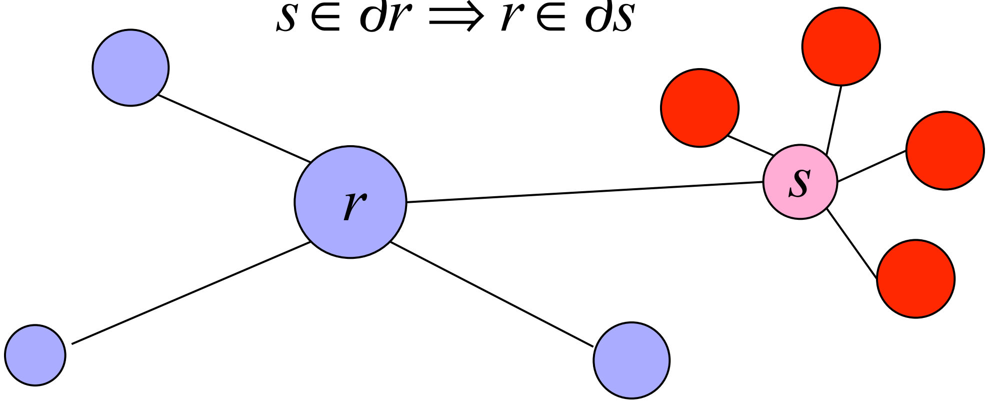
T : "temperature"

V_c : clique potentials

Component-wise MRF

- Neighborhood system
  - □ k-nearest neighbors, based on physical distance
  - □ Enforce neighbors to be mutual

$$s \in \partial r \Rightarrow r \in \partial s$$



- Clique potential
  - □ Dissimilarity measure between neighboring components

Dissimilarity measure, *Di,j*

- Augmented feature vector, *Zi*

  Original feature vector, concatenated with center location of component

- Dissimilarity measure, $D_{i,j}$

  Normalized Mahalanobis distance between feature vectors *Zi* and *Zj*

$$d_{i,j} = \sqrt{(z_i - z_j)^T \Sigma^{-1} (z_i - z_j)}$$

$$D_{i,j} = \frac{d_{i,j}}{\frac{1}{2}(\bar{d}_{i,\partial i} + \bar{d}_{j,\partial j})}$$

S : feature vector covariance

$$\bar{d}_{i,\partial i} = \frac{1}{|\partial i|} \sum_{k \in \partial i} d_{i,k}$$

$$\bar{d}_{j,\partial j} = \frac{1}{|\partial j|} \sum_{k \in \partial j} d_{j,k}$$

Clique potential

- Let *P* denote all neighboring component pairs.
  Then the labels, X, are distributed as

$$p(x) = \frac{1}{Z} \exp\left\{ - \sum_{\{i,j\} \in P} w_{i,j} \delta(x_i \neq x_j) \right\}$$

$$w_{i,j} = \frac{b}{D_{i,j}^{p} + a}$$

a, *b*, and *p* are scalar parameters

- Class probability *p(x)* decreases from neighboring pairs having different class labels
- Decrease is more pronounced when distance $D_{i,j}$ is small

MAP optimization

- Combining data and prior models, compute the MAP estimate for the optimal set of classification labels X

$$\hat{x}_{MAP} = \arg\min_{x \in \{0,1\}^N} \left\{ -\sum_{i \in S} \log p(y_i \mid x_i) + \sum_{\{i,j\} \in P} w_{i,j} \delta(x_i \neq x_j) - c_{txt} \delta(x_i = 1) \right\}$$

- *Ctxt* controls the trade-off between missed and false detections

- Approximate solution using iterative conditional modes (ICM)

Step1 : Initialize each class label x_i with ML estimate

Step2 : For each component, update label

Step3 : If no change occurs to the labels, then stop.
Otherwise go to Step2

Parameter estimation

- Gaussian mixture parameters in data term estimated using *expectation maximization* (EM) algorithm

- Prior model parameters, $f = [p, a, b]$, estimated using pseudo-likelihood maximization (Besag, 1975)

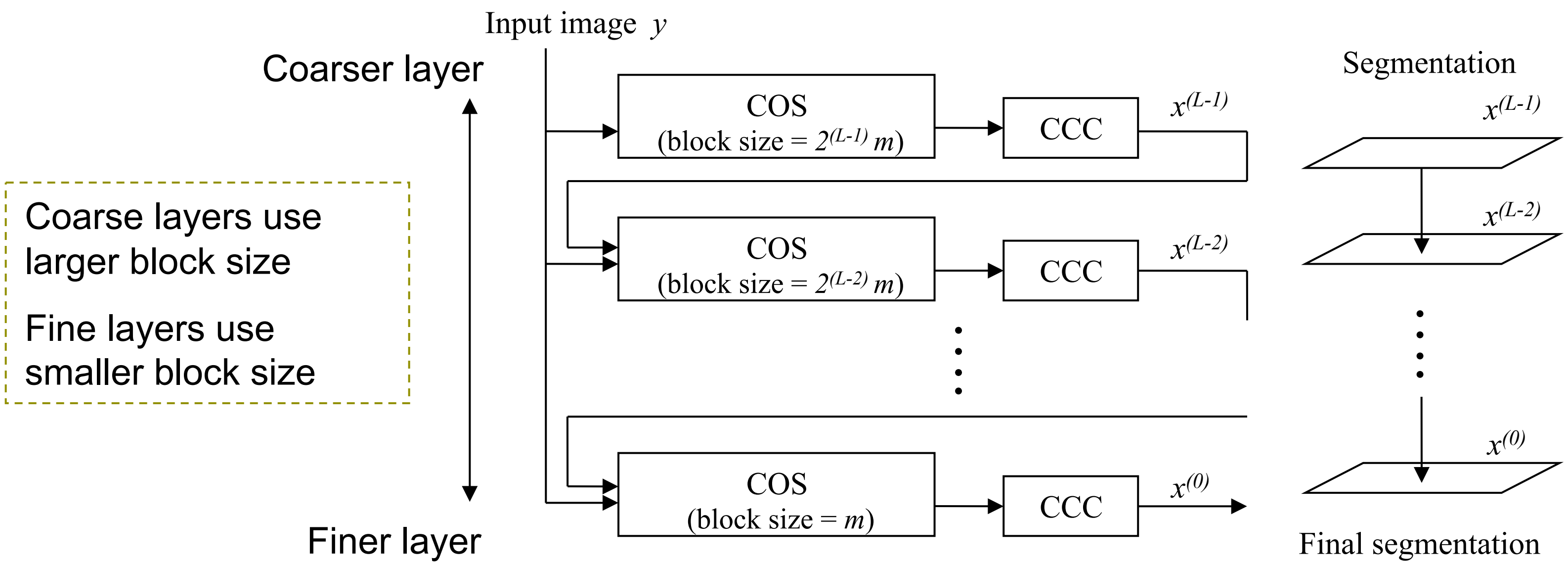
$$\hat{\phi} = \arg\max_{\phi} \prod_{i \in S} p_{\phi}(x_i \mid x_{\partial i})$$

$$= \arg\min_{\phi} \sum_{i \in S} \left\{ \log C_i + \sum_{j \in \partial i} w_{i,j}(\phi)\ \delta(x_i \neq x_j) \right\}$$

$$\textit{where} \quad C_i = \sum_{x_i \in \{0,1\}} \exp\left\{ -\sum_{j \in \partial i} w_{i,j}\ \delta(x_i \neq x_j) \right\}$$

Multiscale-COS/CCC segmentation

- Incorporation of COS/CCC algorithms into a multiscale framework to improve detection of varying size text
- Progress from coarse to fine scales, where coarser scales use larger COS block size
- Segmentation for each layer incorporates result from previous (coarser) layer



Coarse layers use larger block size

Fine layers use smaller block size

Cost function for multiscale-COS/CCC

- New term in the COS cost function represents the number of pixel mismatches between current and previous layers

$$Cost(S^{(n)}) = \sum_{i=0}^{M}\sum_{j=0}^{N}\left\{E(s_{i,j}^{(n)}) + \lambda_1 V_1(s_{i,j-1}^{(n)}, s_{i,j}^{(n)}) + \lambda_2 V_2(s_{i-1,j}^{(n)}, s_{i,j}^{(n)}) + \lambda_3 V_3(s_{i,j}^{(n)}) + \underset{\text{New term}}{\underline{\lambda_4 V_4(s_{i,j}^{(n)}, x_{i,j}^{(n+1)})}}\right\}$$

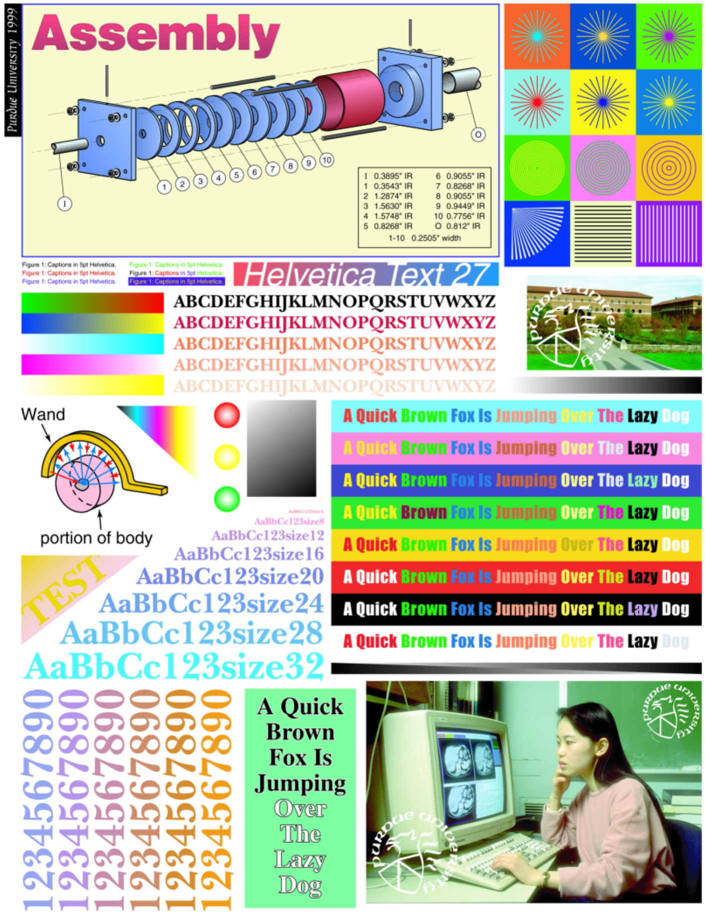
$s_{i,j}^{(n)}$: Class of block at location (*i,j*) on n$_{\text{th}}$ layer.

$$S^{(n)} = \left\{s_{i,j}^{(n)}\right\}$$

- The new term V_4 enforces consistency with coarser segmentation results

Results for complex test image

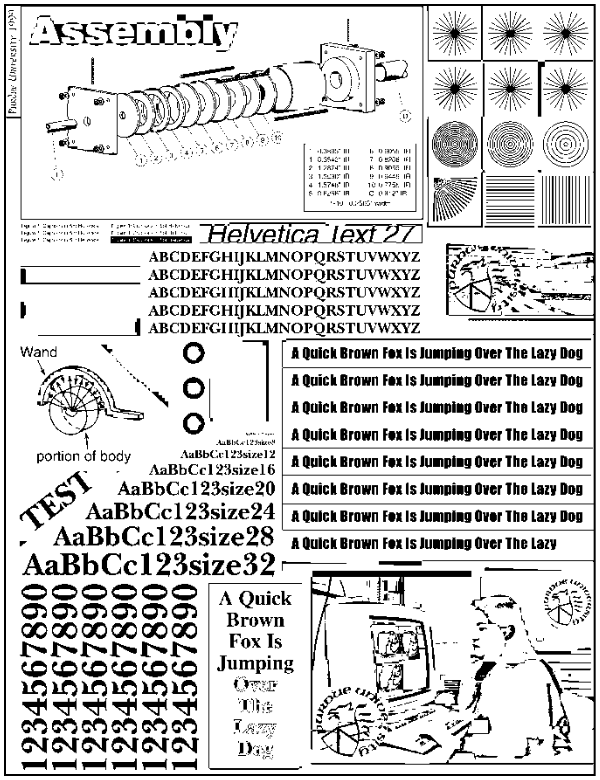
- Comparison of multiscale-COS/CCC and MRC commercial products: *DjVu* (LizardTech) and *LuraDocument* (LuraTech)



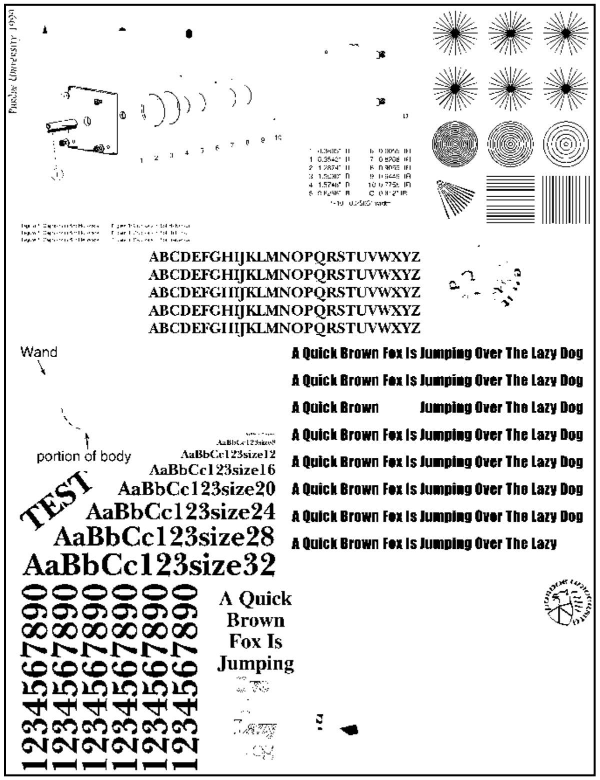
Purdue University 1999
Assembly
I
1
2
3
4
5
6
7
8
9
10
O
I 0.3895" IR
1 0.3543" IR
2 1.2874" IR
3 1.5630" IR
4 1.5748" IR
5 0.8268" IR
6 0.9055" IR
7 0.8268" IR
8 0.9055" IR
9 0.9449" IR
10 0.7756" IR
O 0.812" IR
1-10 0.2505" width
Figure 1: Captions in 5pt Helvetica.
Helvetica Text 27
ABCDEFGHIJKLMNOPQRSTUVWXYZ
ABCDEFGHIJKLMNOPQRSTUVWXYZ
ABCDEFGHIJKLMNOPQRSTUVWXYZ
ABCDEFGHIJKLMNOPQRSTUVWXYZ
ABCDEFGHIJKLMNOPQRSTUVWXYZ
Purdue University
Wand
portion of body
TEST
AaBbCc123size8
AaBbCc123size12
AaBbCc123size16
AaBbCc123size20
AaBbCc123size24
AaBbCc123size28
AaBbCc123size32
A Quick Brown Fox Is Jumping Over The Lazy Dog
A Quick Brown Fox Is Jumping Over The Lazy Dog
A Quick Brown Fox Is Jumping Over The Lazy Dog
A Quick Brown Fox Is Jumping Over The Lazy Dog
A Quick Brown Fox Is Jumping Over The Lazy Dog
A Quick Brown Fox Is Jumping Over The Lazy Dog
A Quick Brown Fox Is Jumping Over The Lazy Dog
A Quick Brown Fox Is Jumping Over The Lazy Dog
1234567890
1234567890
1234567890
1234567890
1234567890
1234567890
A Quick Brown Fox Is Jumping Over The Lazy Dog
Purdue University
Purdue University

Compound test image (400dpi)

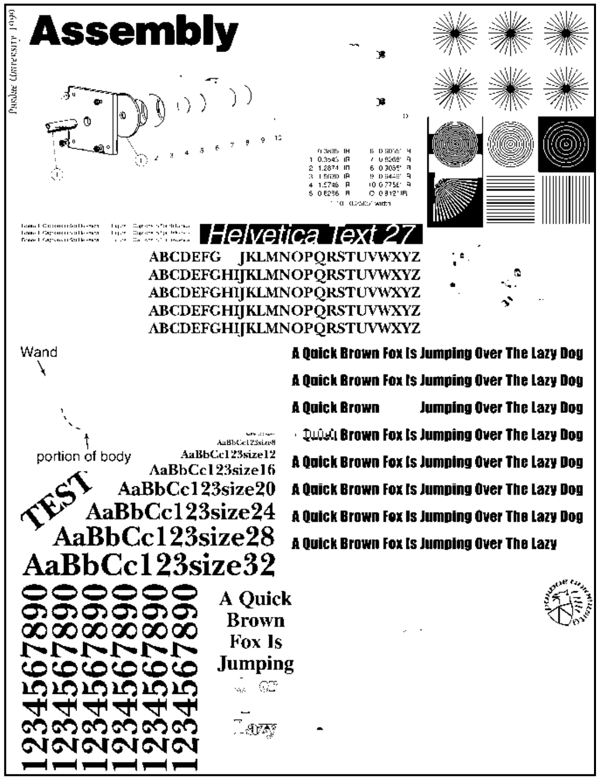
Segmentation comparison



COS only

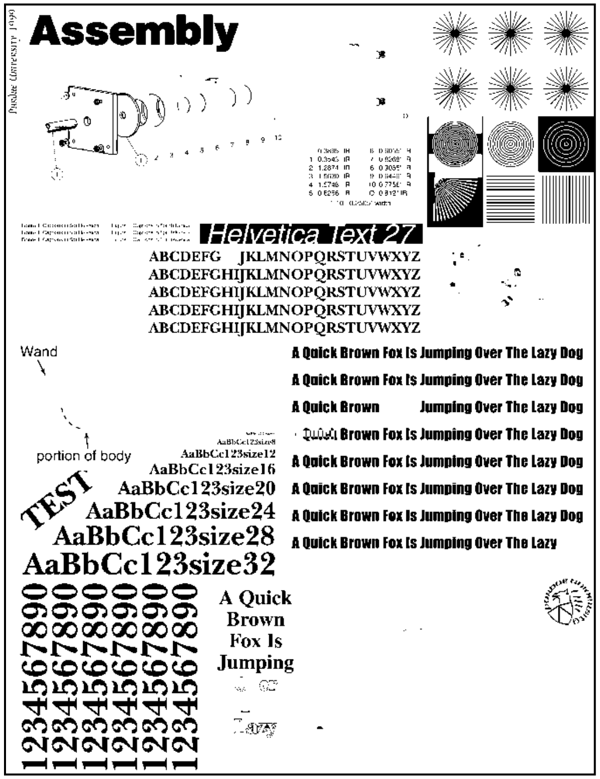


COS/CCC

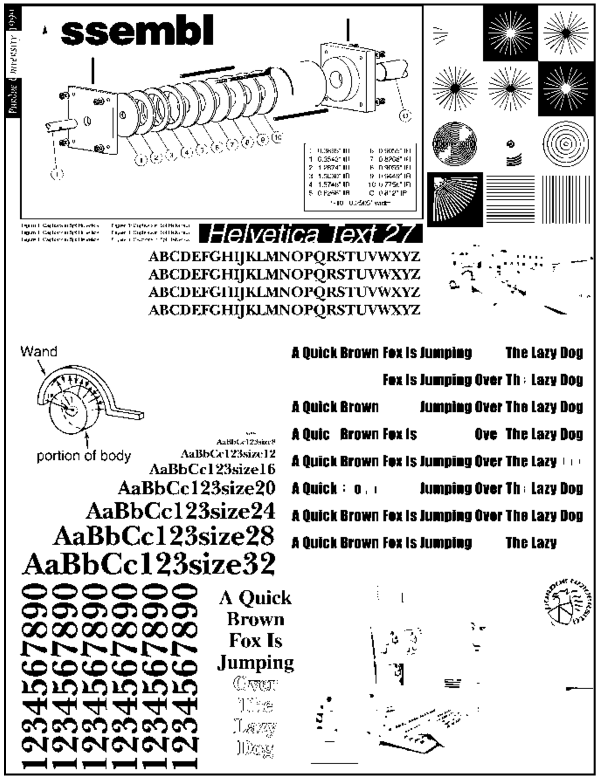


Multiscale-COS/CCC

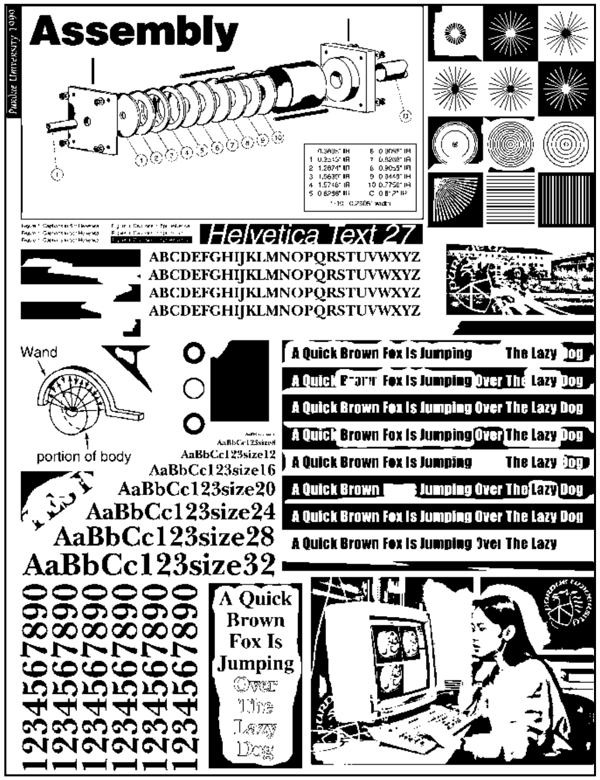
Comparison with commercial products



Multiscale-COS/CCC

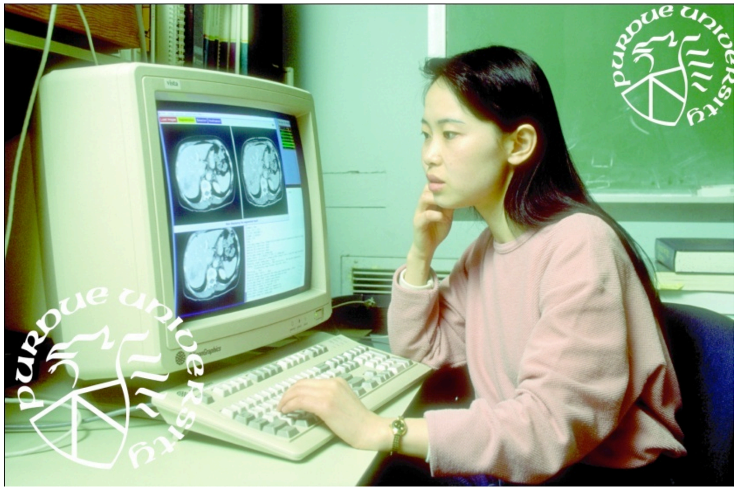


DjVu

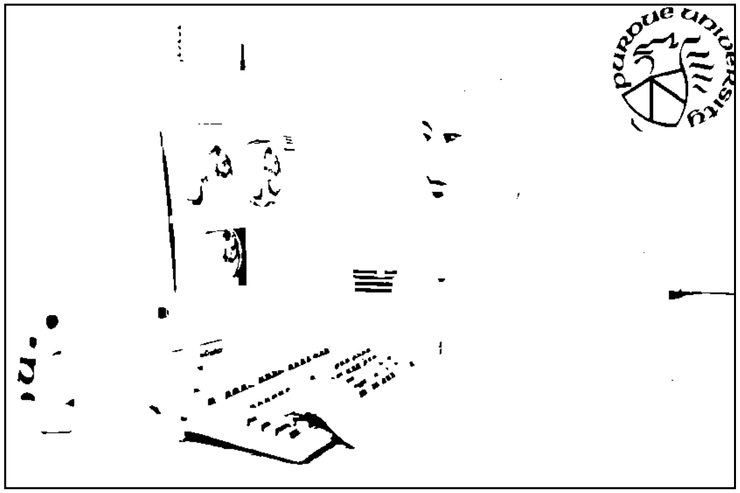


LuraDocument

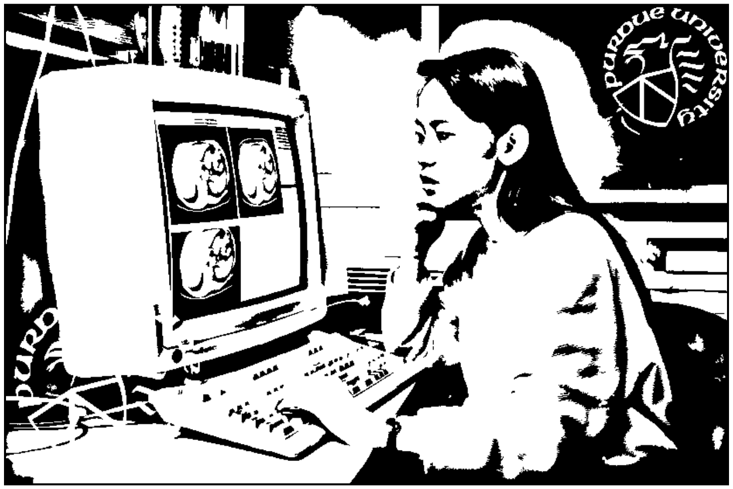
Closer look (Picture regions)



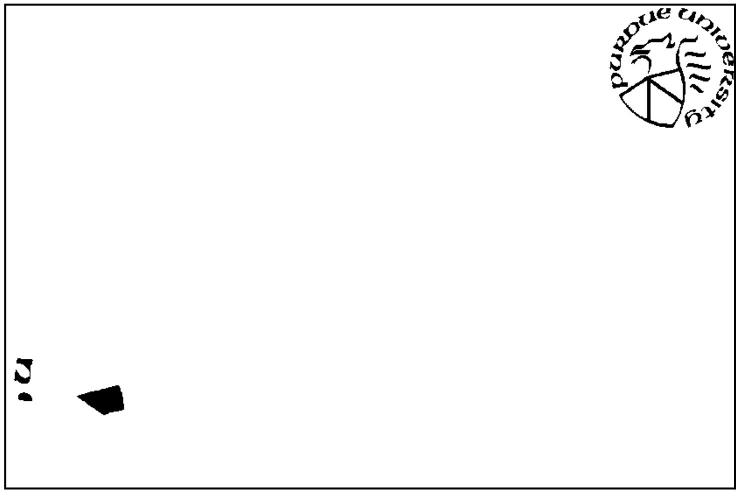
Original



DjVu

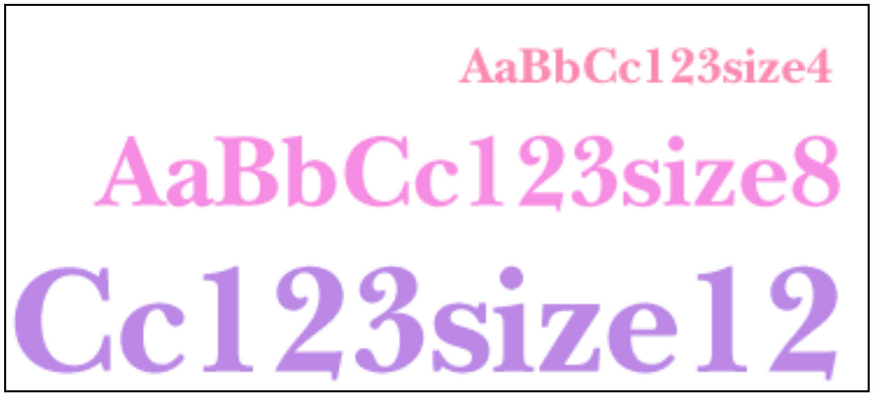


Luratech

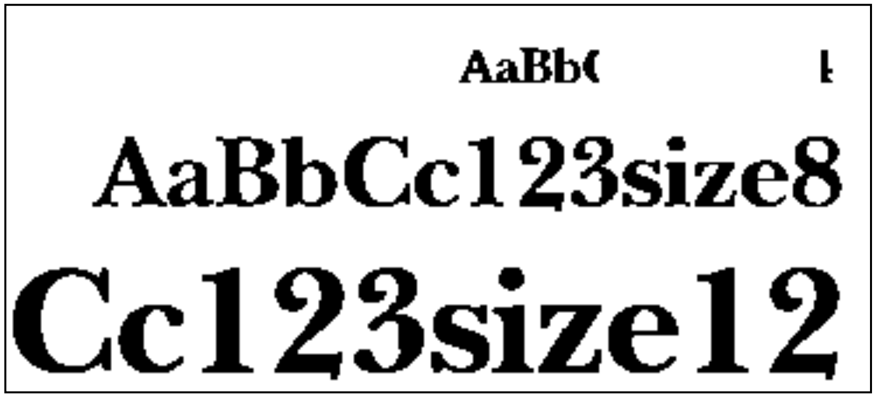


Multiscale-COS/CCC

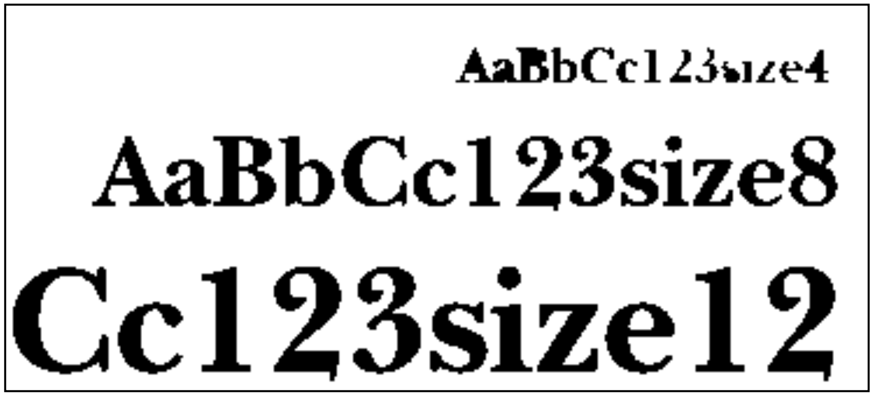
Closer look (Text regions)



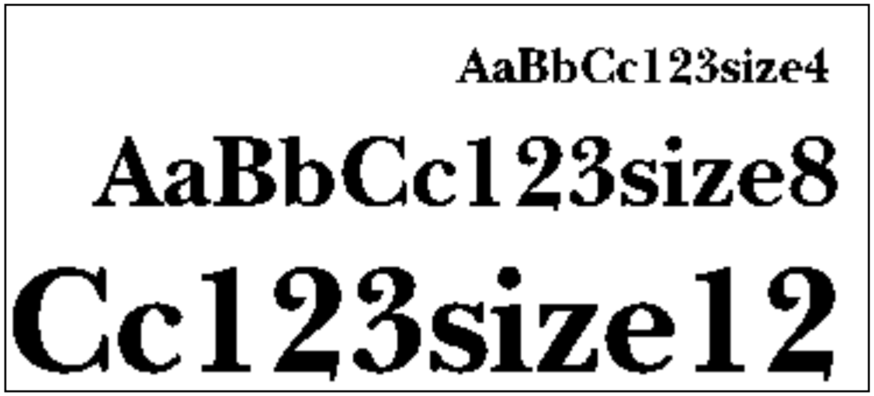
Original



DjVu



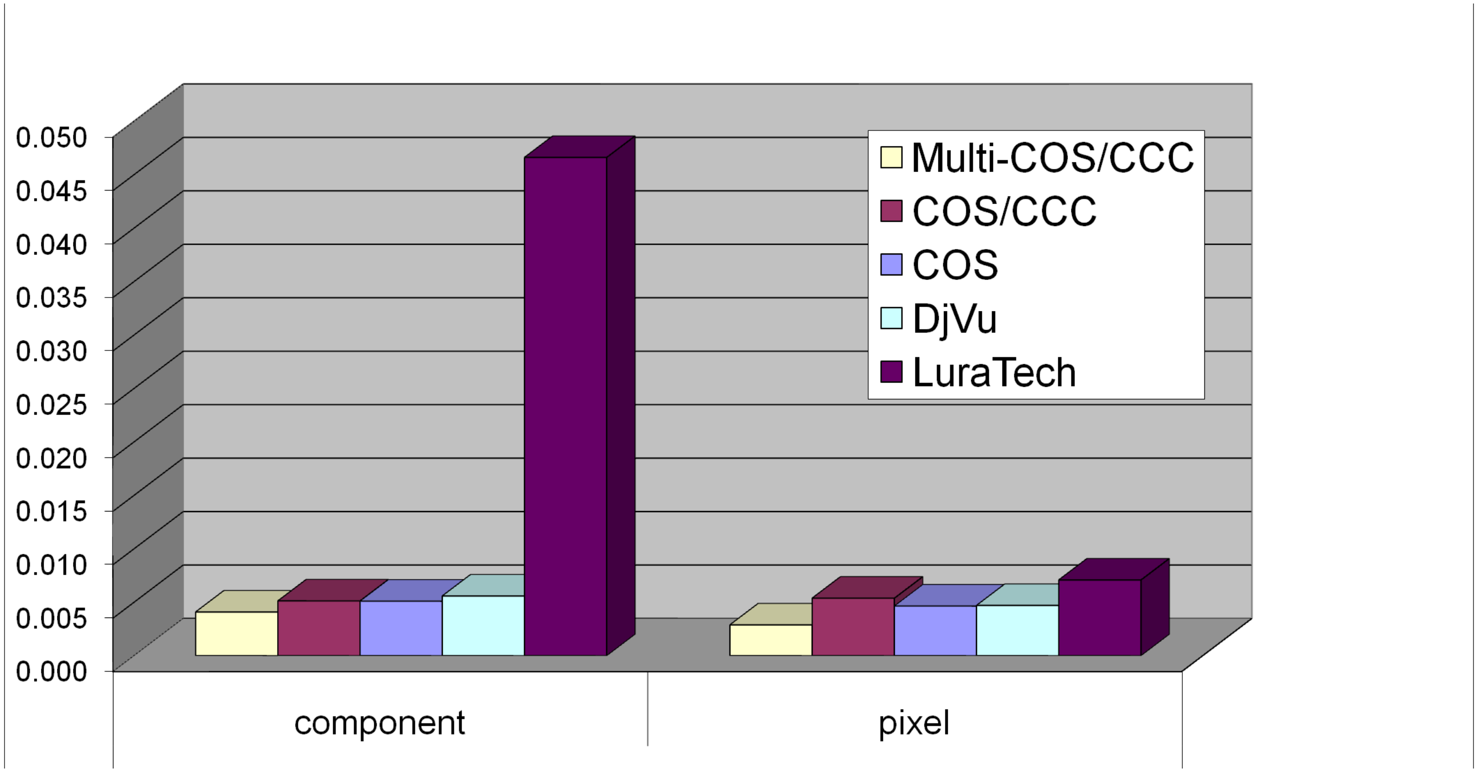
Luratech



Multiscale-COS/CCC

Segmentation error comparison

Missed text detection % (Averaged over EPSON, HP, Samsung scanners)

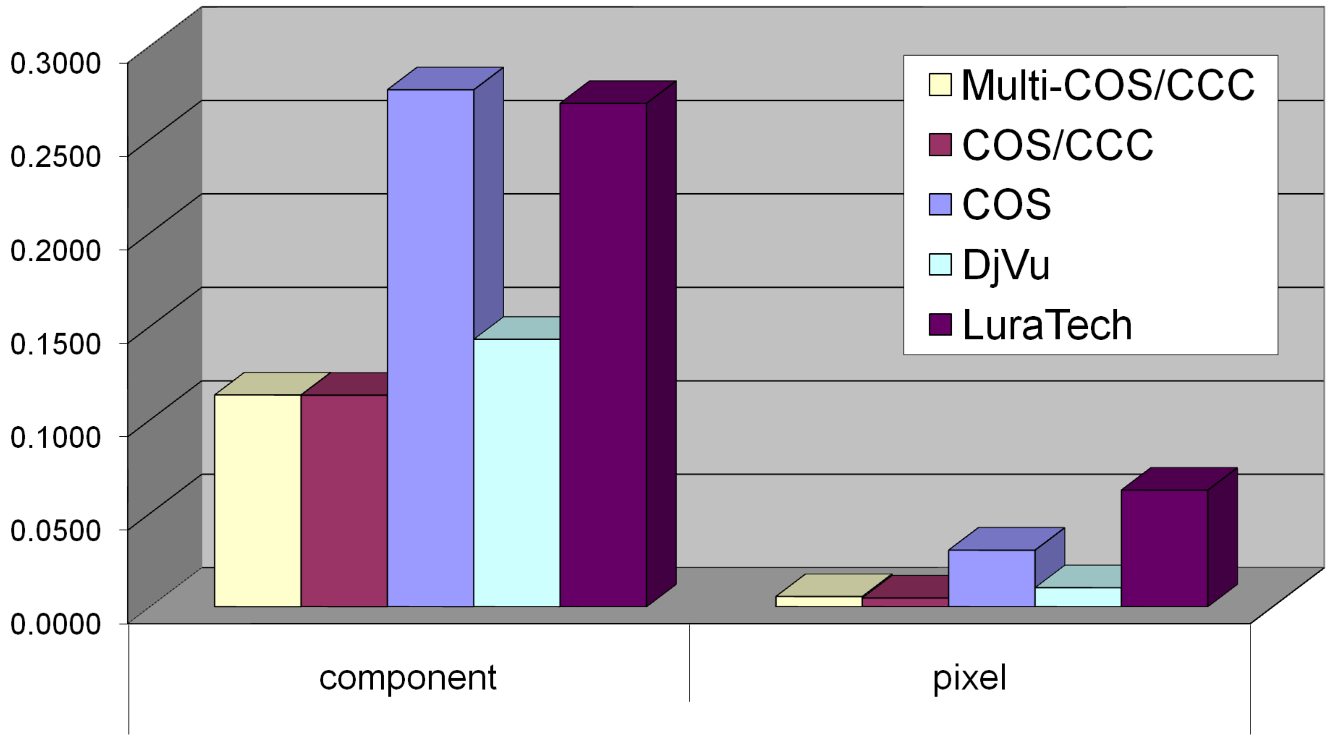


Component = (# missed components) / (# components in ground truth)
Pixel = (# pixels of missed components) / (image size)

- Multiscale-COS/CCC has fewer missed detection than the other algorithms

Segmentation error comparison

False detection % (Averaged over EPSON, HP, Samsung scanners)

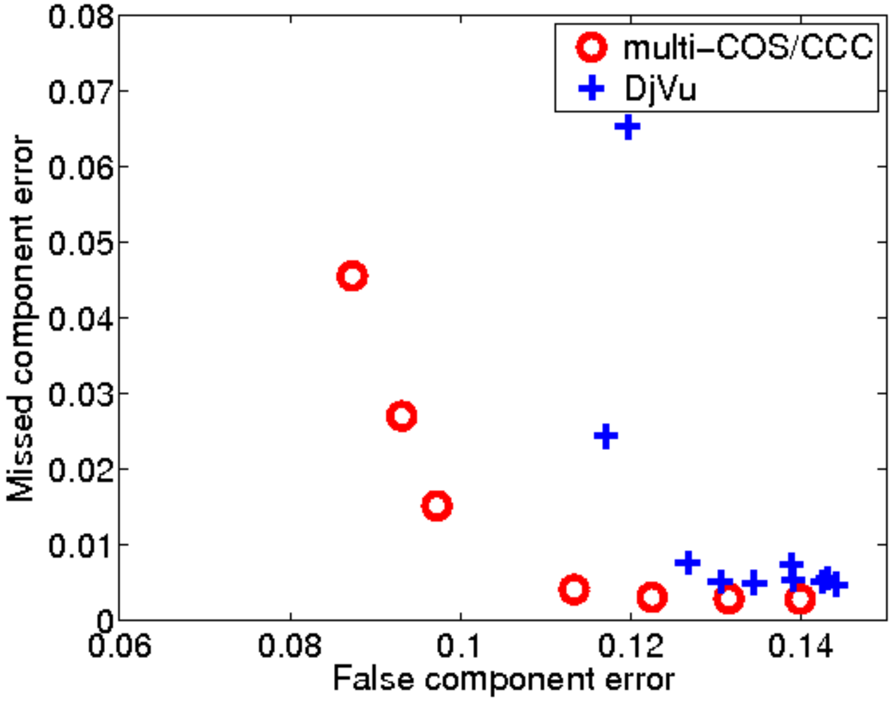


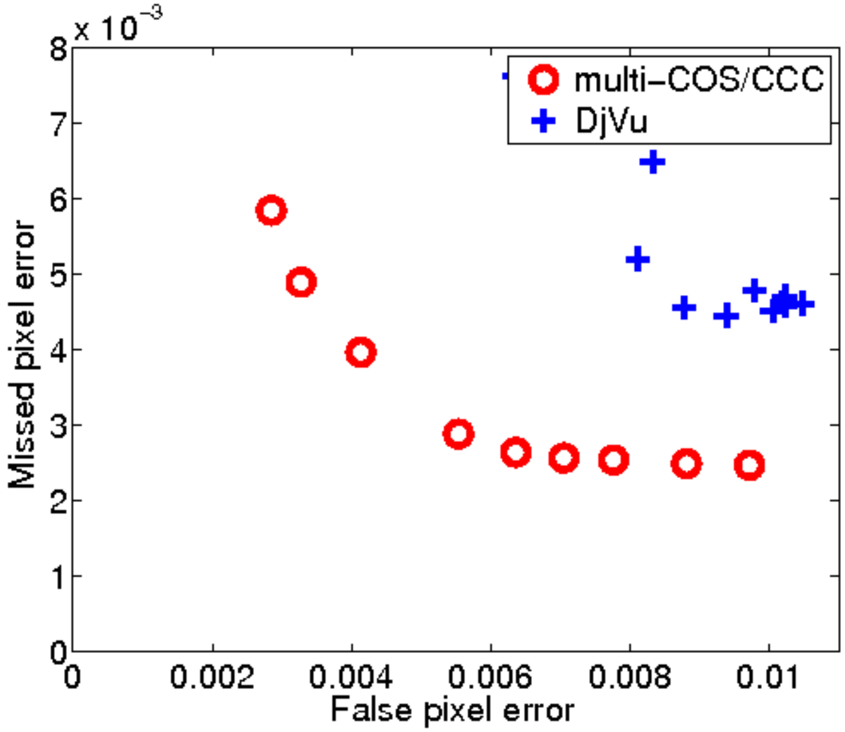
Component = (# false detection) / (# components in ground truth)
Pixel = (# pixels of false detection) / (image size)

- Multiscale-COS/CCC has fewer missed detection than the other algorithms

Trade-off between missed and false detections

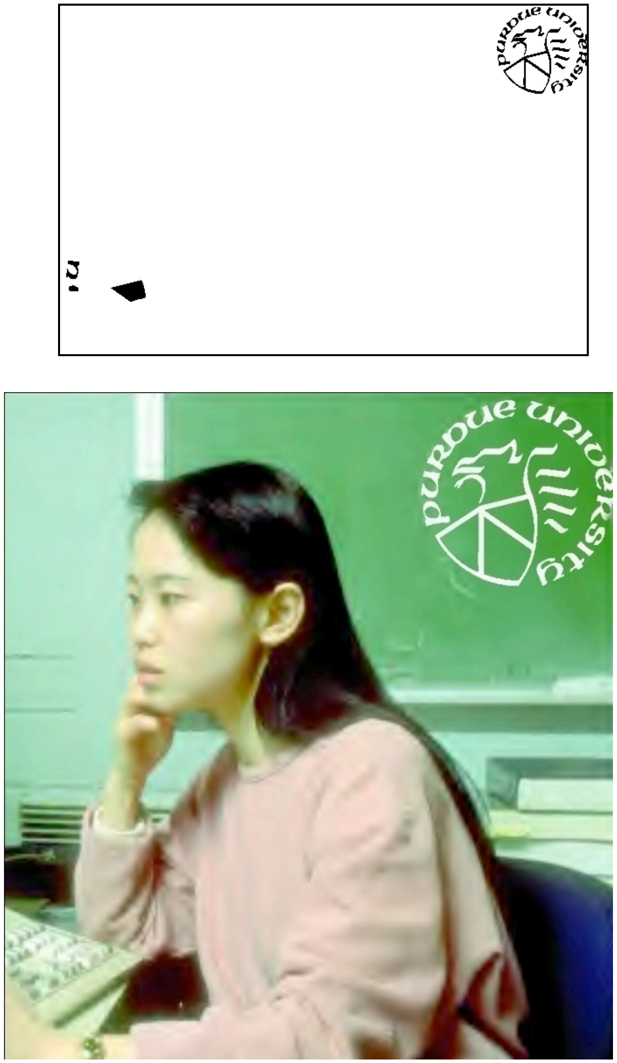
- Multiscale-COS/CCC vs. DjVu (Best commercial product)



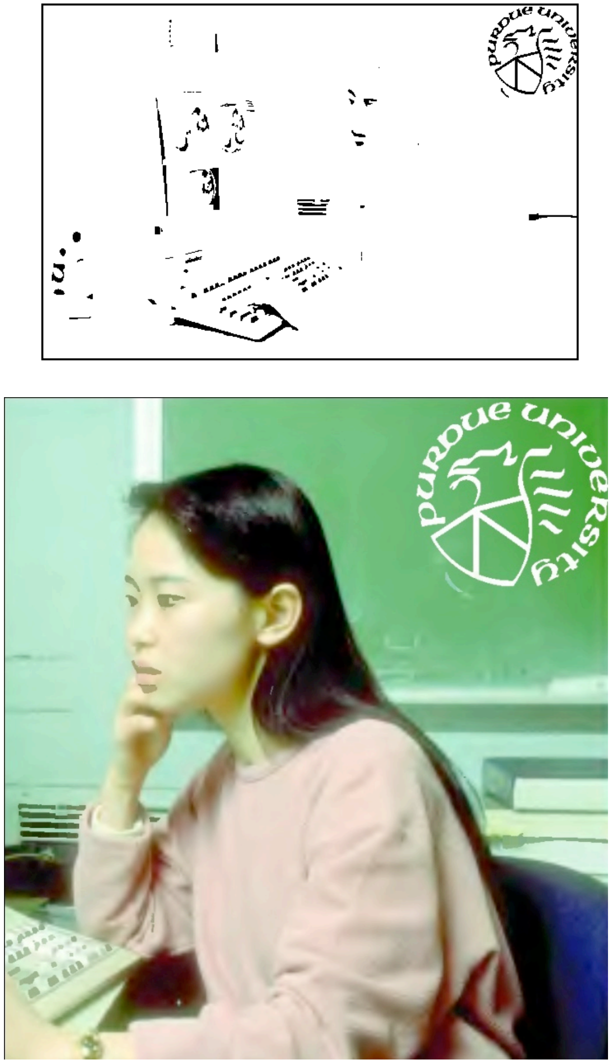


- Averaged over EPSON, HP, and Samsung scanner data
- Multiscale-COS/CCC has superior error rates over DjVu

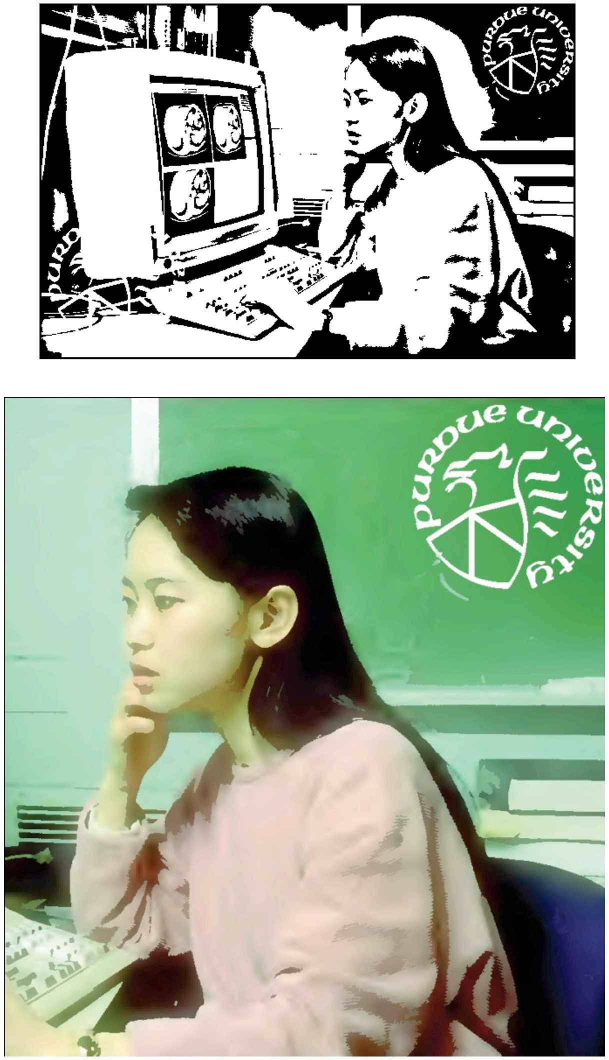
Decoded MRC image comparison #1



Multiscale-COS/CCC (289:1)



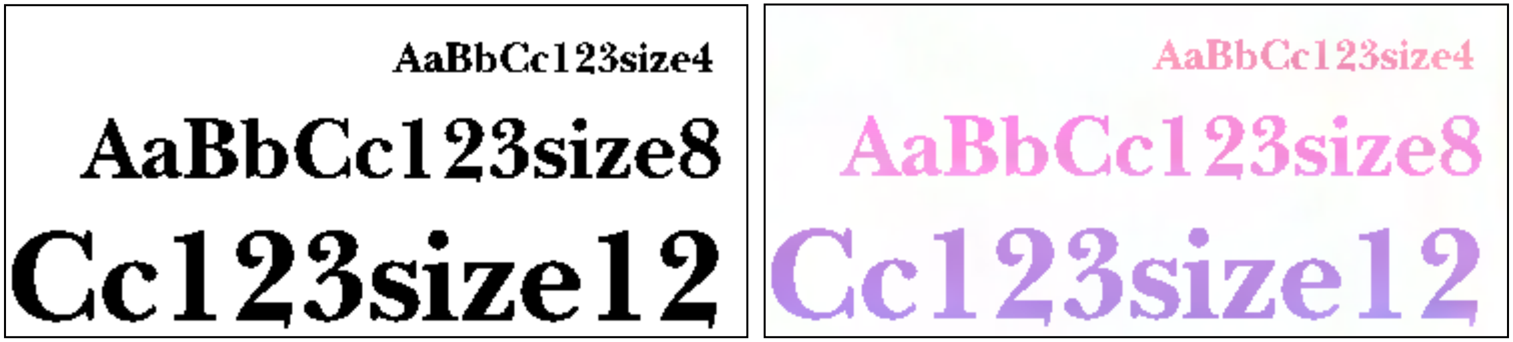
DjVu (281:1)



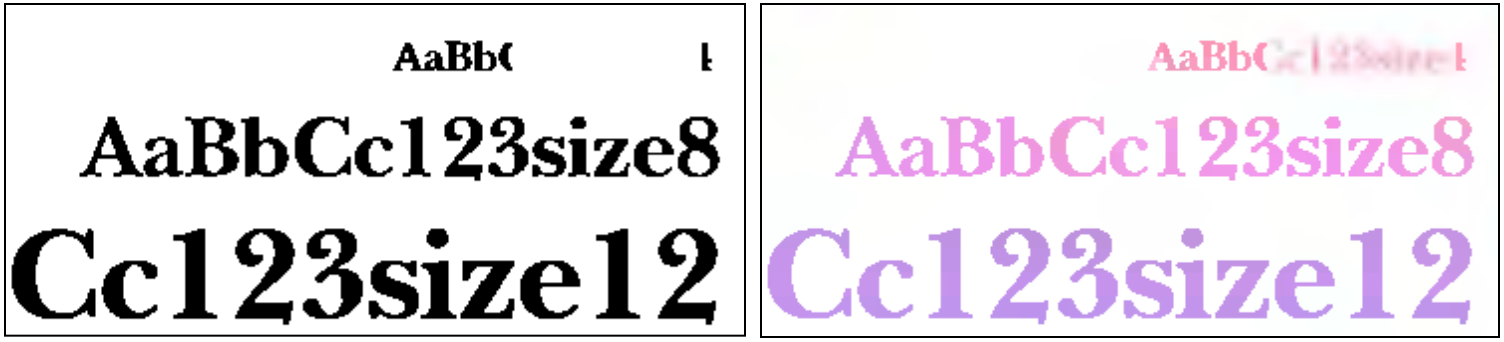
LuraDocument (242:1)

Decoded MRC image comparison #2

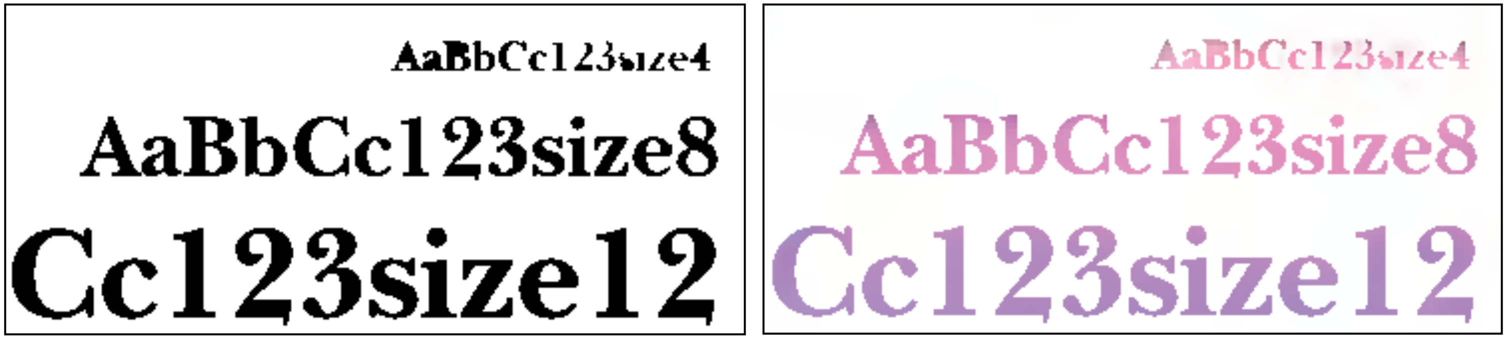
Multiscale-COS/CCC
(289:1)



DjVu
(281:1)



LuraDocument
(242:1)



Summary

- Developed three novel algorithms for text segmentation: COS, COS/CCC, Multiscale-COS/CCC
  - Accurate text extraction compared to commercial products
  - Flexible for future developments
  - Robust over various paper materials, scanner types, and various image backgrounds

- Can extend segmentation to other applications such as Optical Character Recognition (OCR)

Publications/Patents

- Multiscale text segmentation for MRC document
  (Two conference papers, One journal paper, One patent)
  - □ "Segmentation for MRC compression," in Proc. of SPIE Conf. on Color Imaging XII, 2007
  - □ "Multiscale segmentation for MRC compression using a Markov Random Field (MRF) model" in IEEE ICASSP, March 2010
  - □ "Text segmentation for MRC document compression" accepted by IEEE Trans. on Image Processing on Oct 2010
  - □ Patents: combined declaration by Samsung Co. Ltd and Purdue University, United States

- Next generation image capture device development
  (Two patents for snap-to-white algorithm)
  - □ Apparatus and method of segmenting an image in an image coding and/or decoding system
    Application 20080175477
  - □ Auto-cropping method for image capture device
    Application 20090323129

Reference

1. G. Nagy, S. Seth, and M. Viswanathan, “A prototype document image analysis system for technical journals,” *Computer,* vol. 25, no. 7, pp. 10–22, 1992.
2. J. Fisher, “A rule-based system for document image segmentation,” in *Pattern Recognition, 10th international conference,* 1990, pp. 567–572.
3. P. Stathis, E. Kavallieratou, and N. Papamarkos, “An evaluation survey of binarization algorithms on historical documents,” in *19th International Conference on Pattern Recognition, 2008, pp. 1–4.*
4. P. Haffner, L. Bottou, and Y. Lecun, “A general segmentation scheme for DjVu document compression,” in *Proc. of ISMM 2002, Sydney, Australia, April 2002.*
5. Y. Zheng and D. Doermann, “Machine printed text and handwriting identification in noisy document images,” vol. 26, no. 3, pp. 337–353, 2004.
6. J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proc. 18th International Conf. on Machine Learning. organ Kaufmann, 2001, pp. 282–289.*