Iterated Document Content Classification

Chang An, Henry S. Baird and Pingping Xiu

Motivation

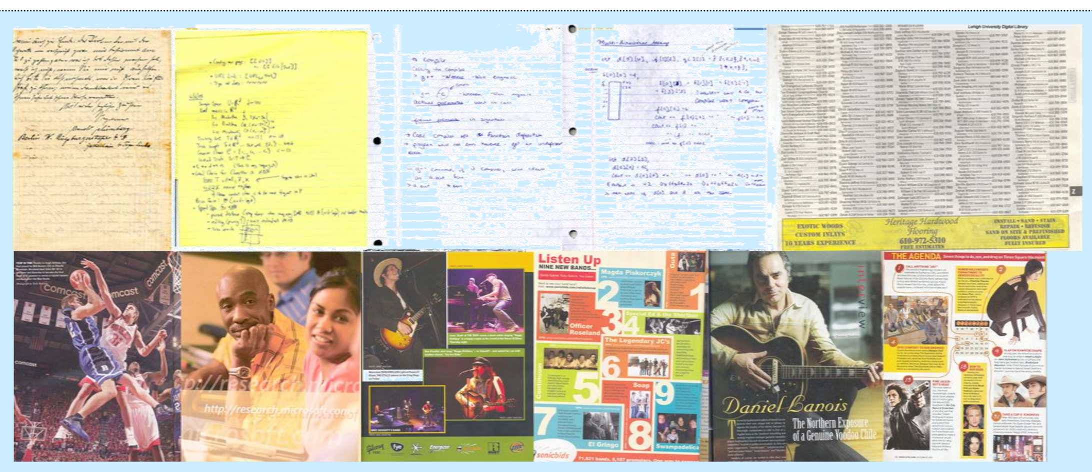
Our previous methods classified each individual pixel separately

This policy allows content classes to vary frequently within small regions

Local uniformity is required for down-stream process

Post-classification method is chosen to enforce local uniformity without imposing arbitrary shapes

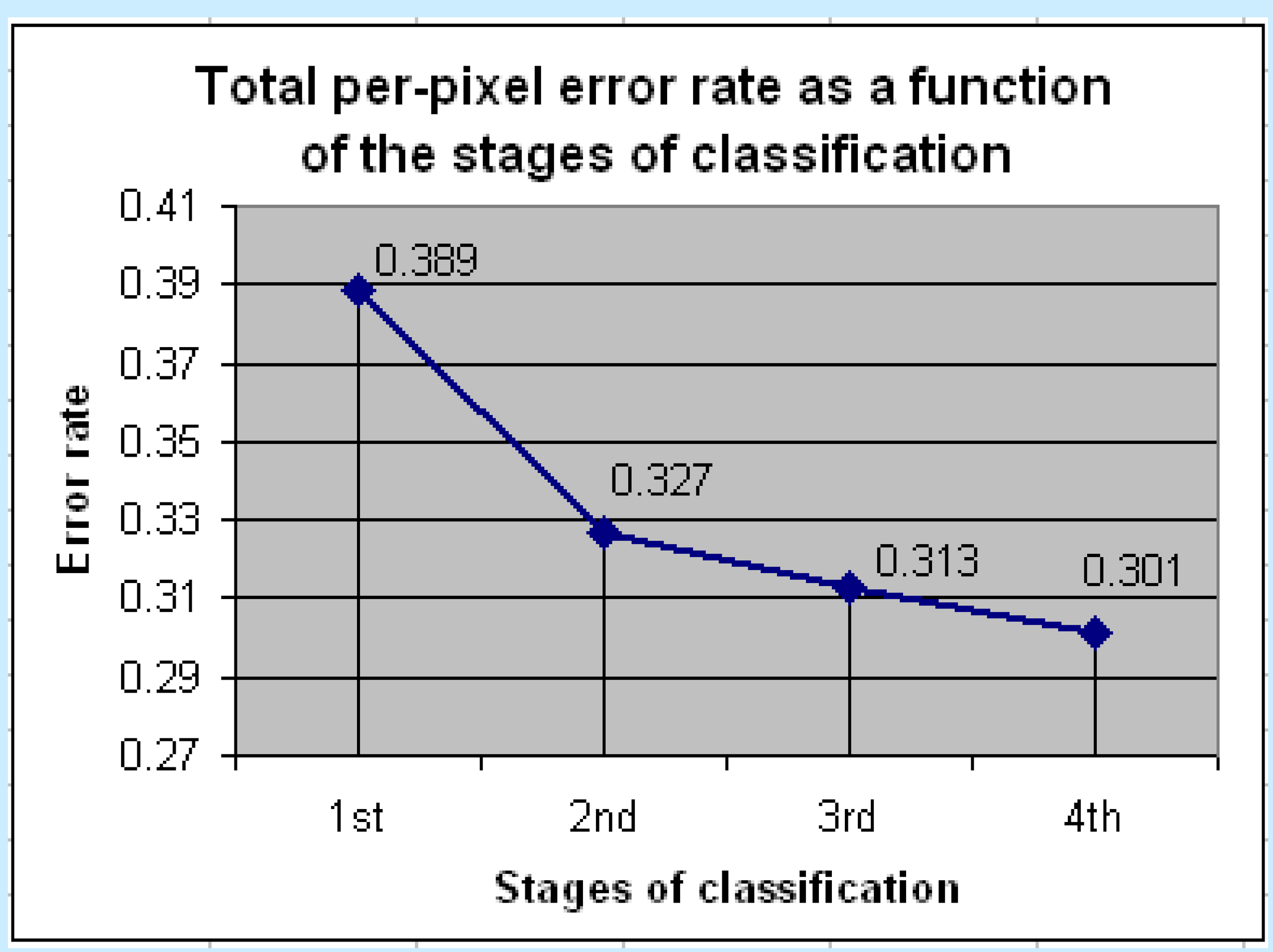
Vast Variety of Images



Training set : 33 images, 87M pixels

Test set : 83 images, 239M pixels

Result: 23% drop in error



Task: Find Uniform Regions of Content in Scanned Document Images

Original image

Output of 1st stage

Color Codes for Content	
Blue: Machine Print	White: Blank
Aqua: Photograph	Purple: Handwriting

Training and test samples are single pixels

Avoids arbitrary, e.g. rectangular, shapes of zones

Classification is by approximate Nearest Neighbors

In output, each content type can be labeled with a color and displayed accordingly

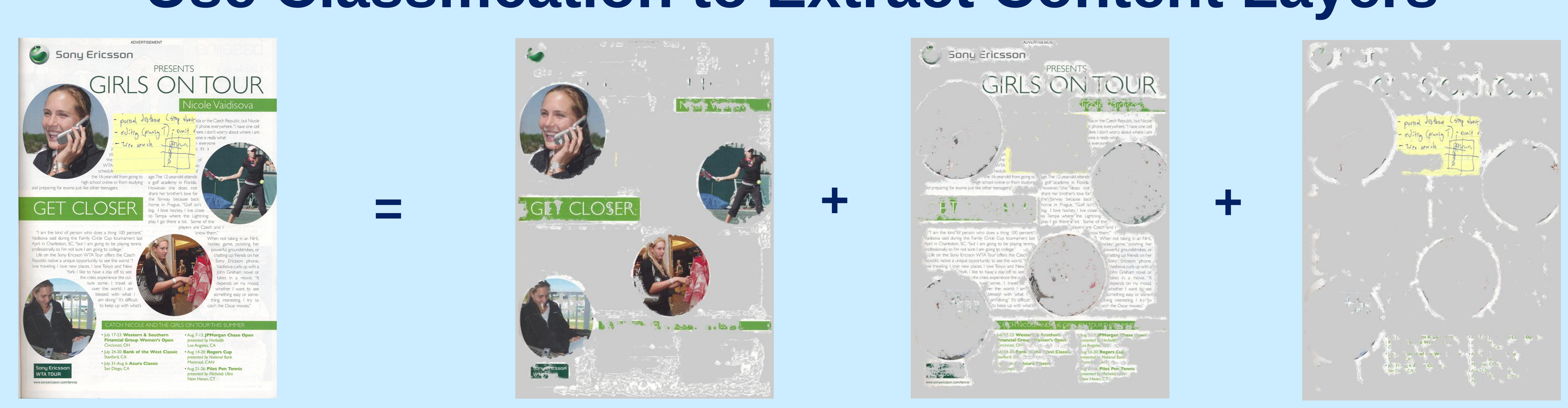
Refinement: Iterated Classification

The 1st-stage classification yields areas where several content classes are mixed together

In real content, almost all small local regions are of uniform class

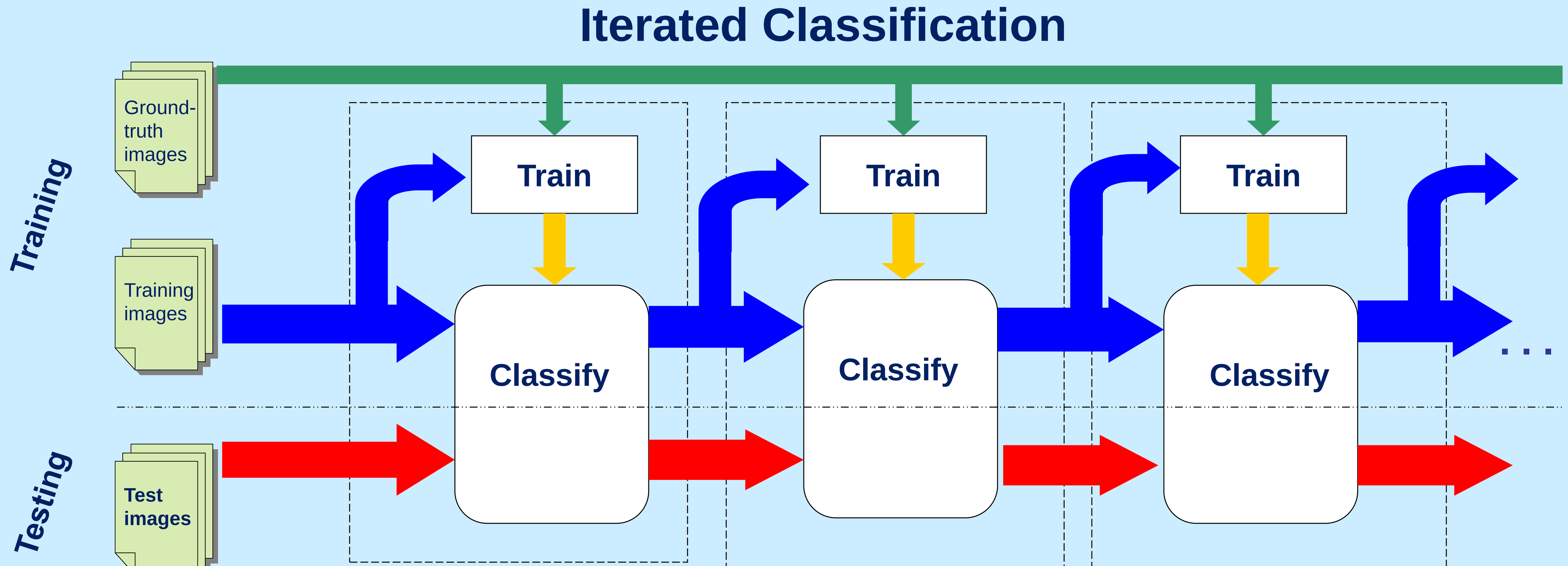
Improve result by post-classification

Use Classification to Extract Content Layers



Photograph (PH) **Machine Print (MP)** **Handwriting (HW)**

Iterated Classification



Instability Issue

In one experiment, at the 9th stage, large solid regions of HW were misclassified as MP.

Promising workarounds:

(1) Drop a training image whenever its error rate rises

(2) Increase the radius of the features

Future Work

Classification with features over a range of scales

Seek guaranteed solutions to the instability issue

Increase the number of iterations

COMPUTER SCIENCE & ENGINEERING

LEHIGH UNIVERSITY

P.C. Rossin College of Engineering and Applied Science

Computer Science and Engineering CSE

Henry S. Baird & Daniel Lopresti
Pattern Recognition Research Lab