A Discriminative Semi-Markov Model for Robust Scene Text Recognition

Jerod J. Weinman, Erik Learned-Miller, and Allen R. Hanson

Computer Vision Laboratory, Dept. of Computer Science, University of Massachusetts Amherst

With support from:

Problem and Motivation

Our goal: Read text from images of the everyday world.

Applications include:
- Aid to visually impaired
- Photo annotation
- Portable translation

Many new challenges are atypical of OCR page readers:

- *Small sample (<50 chars)*

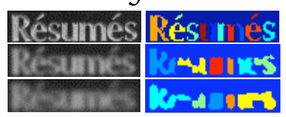
- *Perspective distortion*

- *Uneven lighting*

- *Unusual fonts*

- *Word segmentation*

- *Character segmentation*



Previous work has assumed word boundaries, character segmentations, and/or known lexicon words. The conditions above make these assumptions problematic.

We propose a model that integrates lexical decision and both word and character segmentation with recognition.

Semi-Markov Model

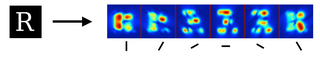
We use a discriminatively trained semi-Markov model for recognition. Like a CRF, it models dependencies between states. It has the additional property of modeling the duration of a state, i.e., character.

$$p(\mathbf{y}|\mathbf{x};\vec{\theta}) = \frac{1}{Z(\mathbf{x})}\exp\{U(\mathbf{y},\mathbf{x};\vec{\theta})\}$$

Model parameters $\vec{\theta}$ can be learned from labeled training data.

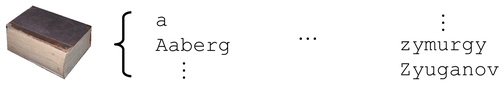
The exponent (U) is composed of functions that score a segmentation and labeling ($\mathbf{y}$) of a text image ($\mathbf{x}$) using:

- *Appearance*

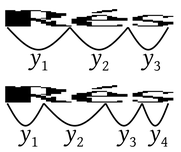


- *Bigrams* $P(\text{TH}\mid\text{English}) = \frac{39}{1000}$ $P(\text{QU}\mid\text{English}) = \frac{1.4}{1000}$ $P(\text{QA}\mid\text{English}) = \frac{.0001}{1000}$

- *Lexicon*



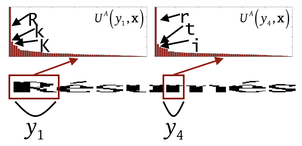
Parse Scoring



A text image is parsed, or divided into labeled segments. Parses may differ in the number of segments, and the states y_i may differ in width.

Appearance $U^A(y_i,\mathbf{x})$

Every possible parse segment is scored by a discriminant U^A for character appearance and width.



Bigrams $U^B(y_i,y_{i+1})$

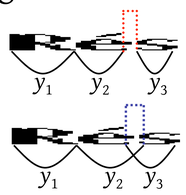
Each pair of neighboring segments get a bigram score U^B.

Lexicon U^L

A character invariant U^L replaces U^B in sequences forming lexicon words to promote known word recognition.

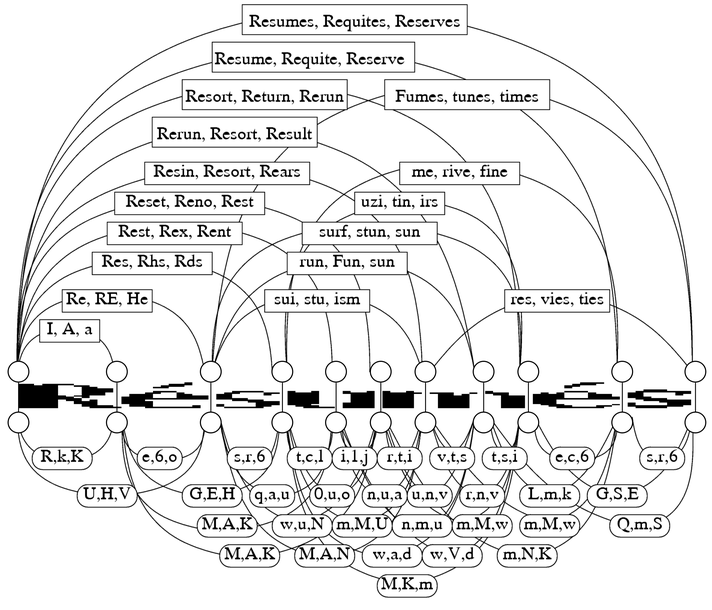
Parse Gap/Overlap $U^P(y_i,y_{i+1},\mathbf{x})$

Neighboring states may have a gap or an overlap, as in the case of ligatures. These are also scored using image features.



Interpretation Graph

All possible parses can be represented by a weighted graph. Dynamic programming can be used to find the optimal path, or the best interpretation of the image.



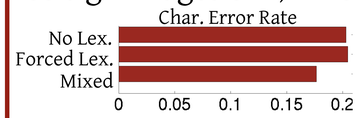
Experiments

Training

- *Appearance:* Synthetic images generated from 934 fonts
- *Bigrams:* 82 books from Project Gutenberg
- *Lexicon:* 50th frequency percentile words from SCOWL

Evaluation

85 sign images of 1,144 characters at ~12 pixel x-height.



Allowing words both in and out of the lexicon reduces error by 13%.

Original Image	Semi-Markov Model Output	Binary Image	OmniPage Output
First	First	First	Ella
COFFEE HOUSE	COFFEE HOUSE	COFFEE HOUSE	COFFEE Maus
DOUGLASS	Douglass	DOUGLASS	uucL,ass
Free checking	Free checking	Free checking	Free ehe in
LIBRARY	LIBRARY	LIBRARY	LIBRFIRY
AMHERST	AMHERST	AMHERST	A),1HERbT
TAVERN	TAVERN	TAVERN	TAVF
Fleet	Fleet	Fleet	
CHURCH	CHURCH	CHURCH	EMI
MONEYFLO	MONEYFLO	MONEYFLO	JUONEYFLO
MONKEY	MONKEY	MONKEY	.010NIK EY

System	Char Err.
OmniPage	23.5%
OmniPage + Binarized	16.6%
Semi-Markov	15.0%

Above: Our model reduces error by 10%.

Left: Examples.

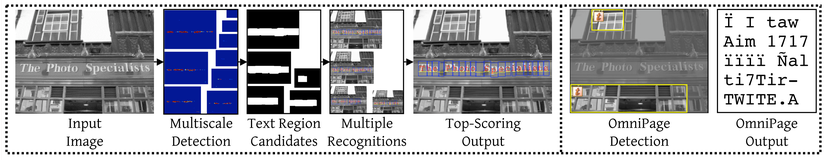
Low Resolution

Integrated segmentation and recognition allow our model to recognize images at lower resolutions than trained on.

Image	Semi-Markov	OmniPage
Résumés	Resumes	Resumes
Résumés	Resumes	Resumes
Résumés	Resumes	WM/ I WS
	Resumes	
	Resumes	krauts
	tttA wms	11.-4her

Image	Semi-Markov	OmniPage
Jeffery Amherst	Jeffery Amherst	Je eryAmherst
Jeffery Amherst	Jeffery Amherst	JefferyAmherst
Jeffery Amherst	Jeffery Amherst	JONryAmhent
Jeffery Amherst	Lit wryA An memo	GkilaryAmhent
Jeffery Amherst	An wrrA Am wme	.WlervAmMnt

End-to-End



With a text detector, we have a complete reading system.