Multi-lingual Videotext Recognition

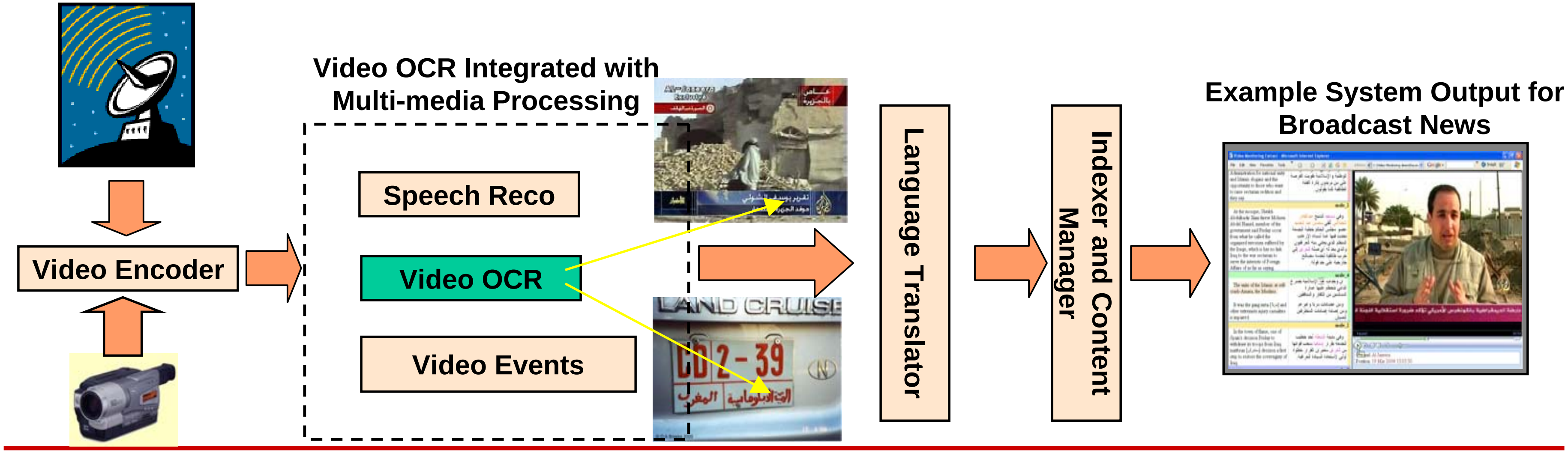
Rohit Prasad, Prem Natarajan, Ehry MacRostie, Michael Decerbo, John Makhoul

Outline

- Goals and Expected Impact
- Challenges in Videotext Recognition
- Description of Videotext Recognition System
- Results on English Broadcast News
- Speed Improvements
- Preliminary results on Arabic Broadcast News
- Conclusions and Future Work

Goals and Expected Impact

Conceptual View of Video Indexing System



Video Encoder

Video OCR Integrated with Multi-media Processing

Speech Reco

Video OCR

Video Events

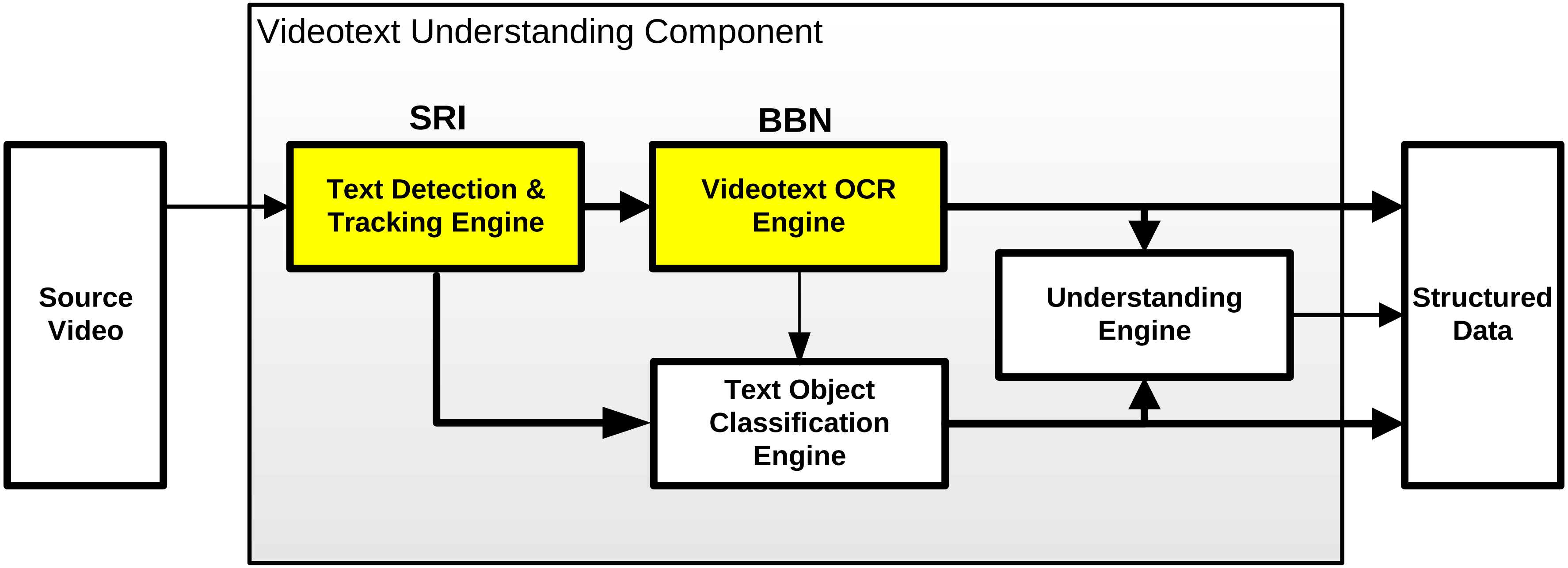
Language Translator

Indexer and Content Manager

Example System Output for Broadcast News

- **Goal: Develop a videotext understanding component for integration into end-to-end video analysis systems**
- **Impact: Enables content-based search and retrieval, real-time alerting, and triage of video in several domains**

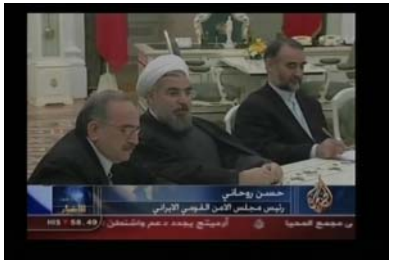
Videotext Understanding: Block Diagram

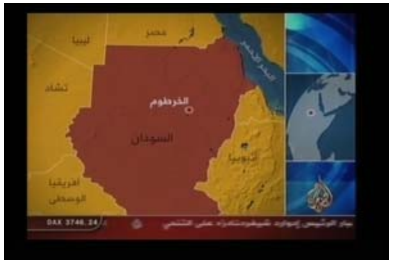

Videotext Understanding Component
SRI
BBN
Source Video
Text Detection & Tracking Engine
Videotext OCR Engine
Understanding Engine
Text Object Classification Engine
Structured Data

Videotext: Examples from Different Domains

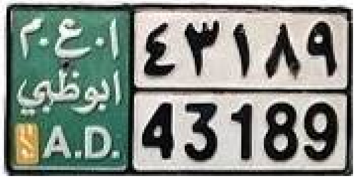
Meeting Videos

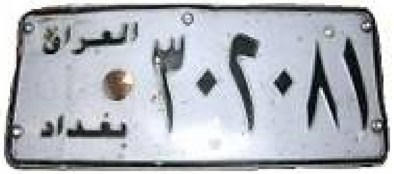
Surveillance Videos





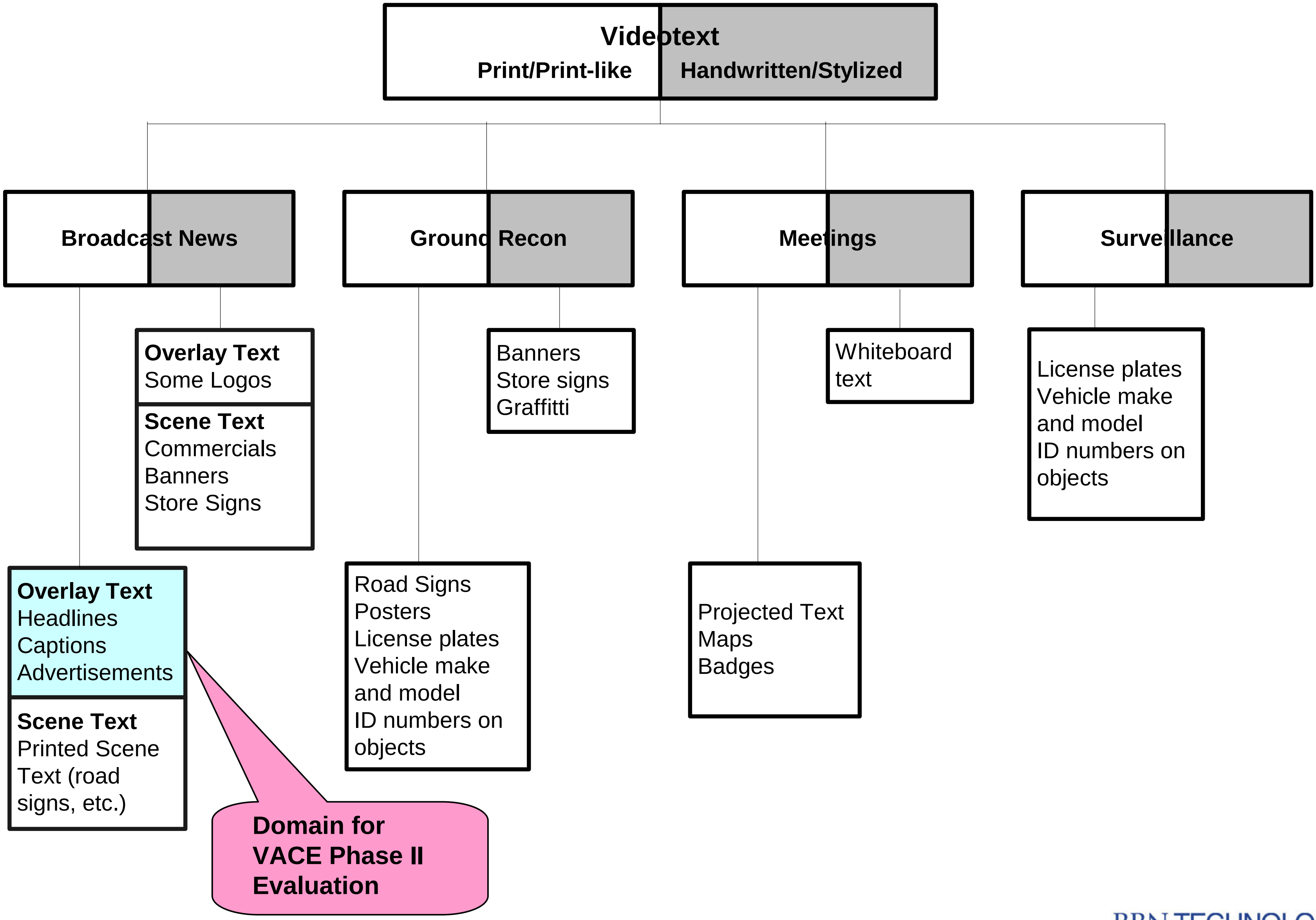
Broadcast News (BN) Videos





Vehicle License Plates

Taxonomy of Text in Video



Videotext
Print/Print-like
Handwritten/Stylized

Broadcast News

Ground Recon

Meetings

Surveillance

Overlay Text
Some Logos

Scene Text
Commercials
Banners
Store Signs

Banners
Store signs
Graffitti

Whiteboard text

License plates
Vehicle make and model
ID numbers on objects

Overlay Text
Headlines
Captions
Advertisements

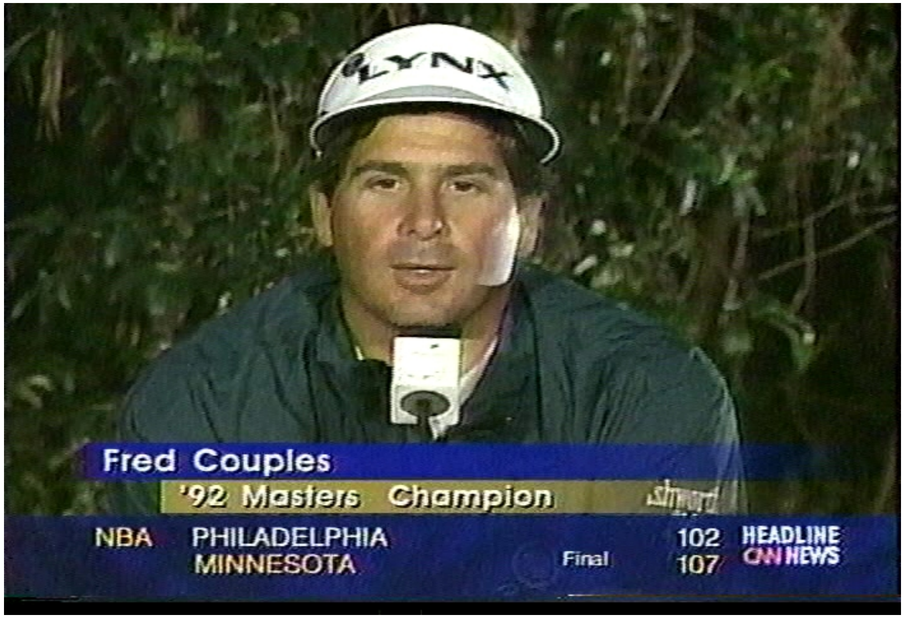
Scene Text
Printed Scene Text (road signs, etc.)

Road Signs
Posters
License plates
Vehicle make and model
ID numbers on objects

Projected Text
Maps
Badges

Domain for VACE Phase II Evaluation

Sample BN Video Frames

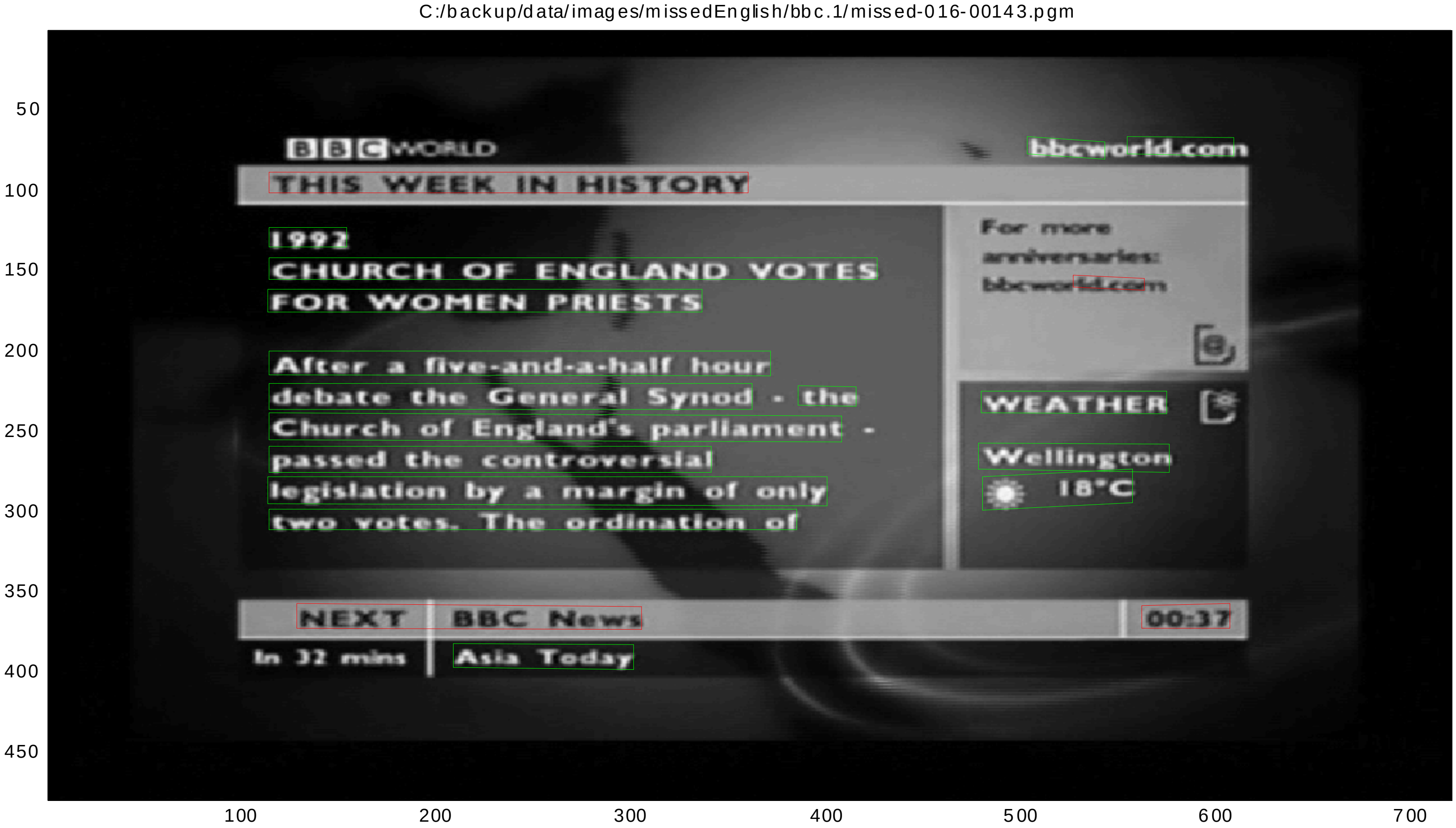


مباشر
العقبة
عبدالله: كثيرون ينظرون إلى قمة العقبة
في تشكك ويجب أن نستعيد ثقتهم
لفلسطيني في ختام قمة العقبة بوقف الانتفاضة
مكتب رئي

Key Challenges in BN Videotext Recognition

- **Low Resolution**
  - Resolution of videotext is much lower than the resolution of scanned document images
- **Moving overlay text**
  - Causes text to exhibit jagged edges and smear
- **Compression**
  - Causes artifacts that add to recognition challenge
- **Perspective distortion in scene text**

Text Detection


C:/backup/data/images/missedEnglish/bbc.1/missed-016-00143.pgm
50
100
150
200
250
300
350
400
450
100
200
300
400
500
600
700

BBC WORLD

bbcworld.com

THIS WEEK IN HISTORY

1992

CHURCH OF ENGLAND VOTES
FOR WOMEN PRIESTS

After a five-and-a-half hour debate the General Synod - the Church of England's parliament - passed the controversial legislation by a margin of only two votes. The ordination of

For more anniversaries: bbcworld.com

WEATHER

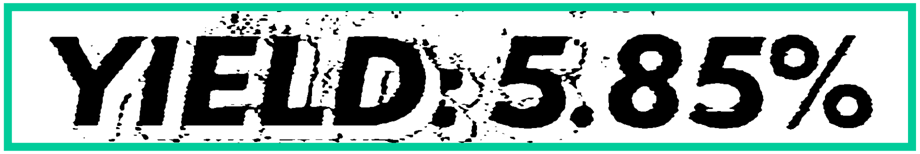
Wellington

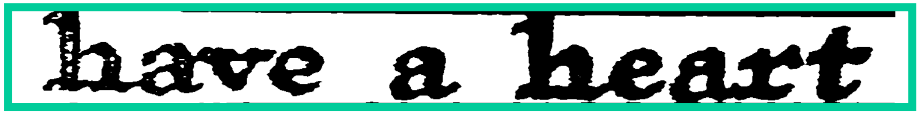
18°C

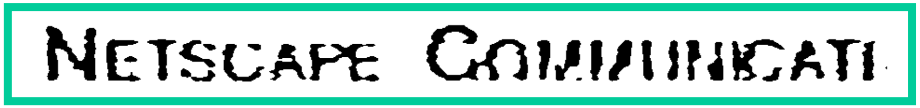
NEXT | BBC News | 00:37

In 32 mins | Asia Today

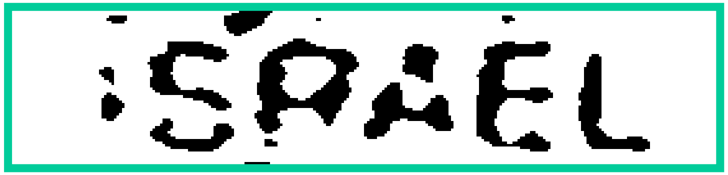
Sample Detected and Binarized BN Images

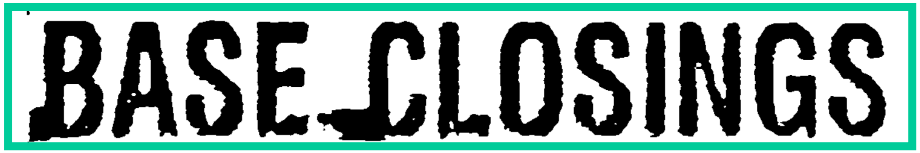


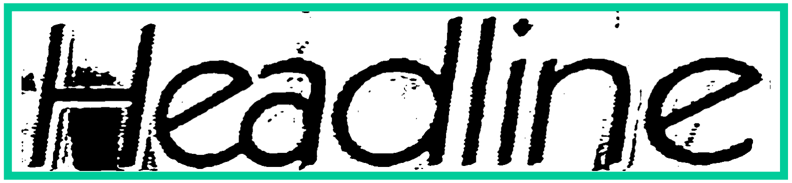


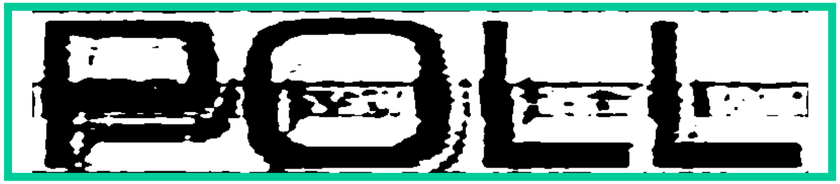


Text detection misses part of the text object



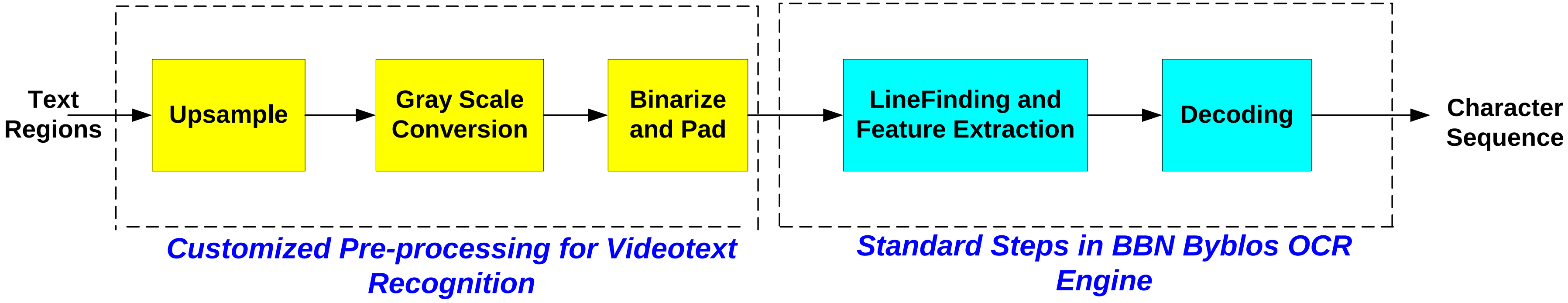




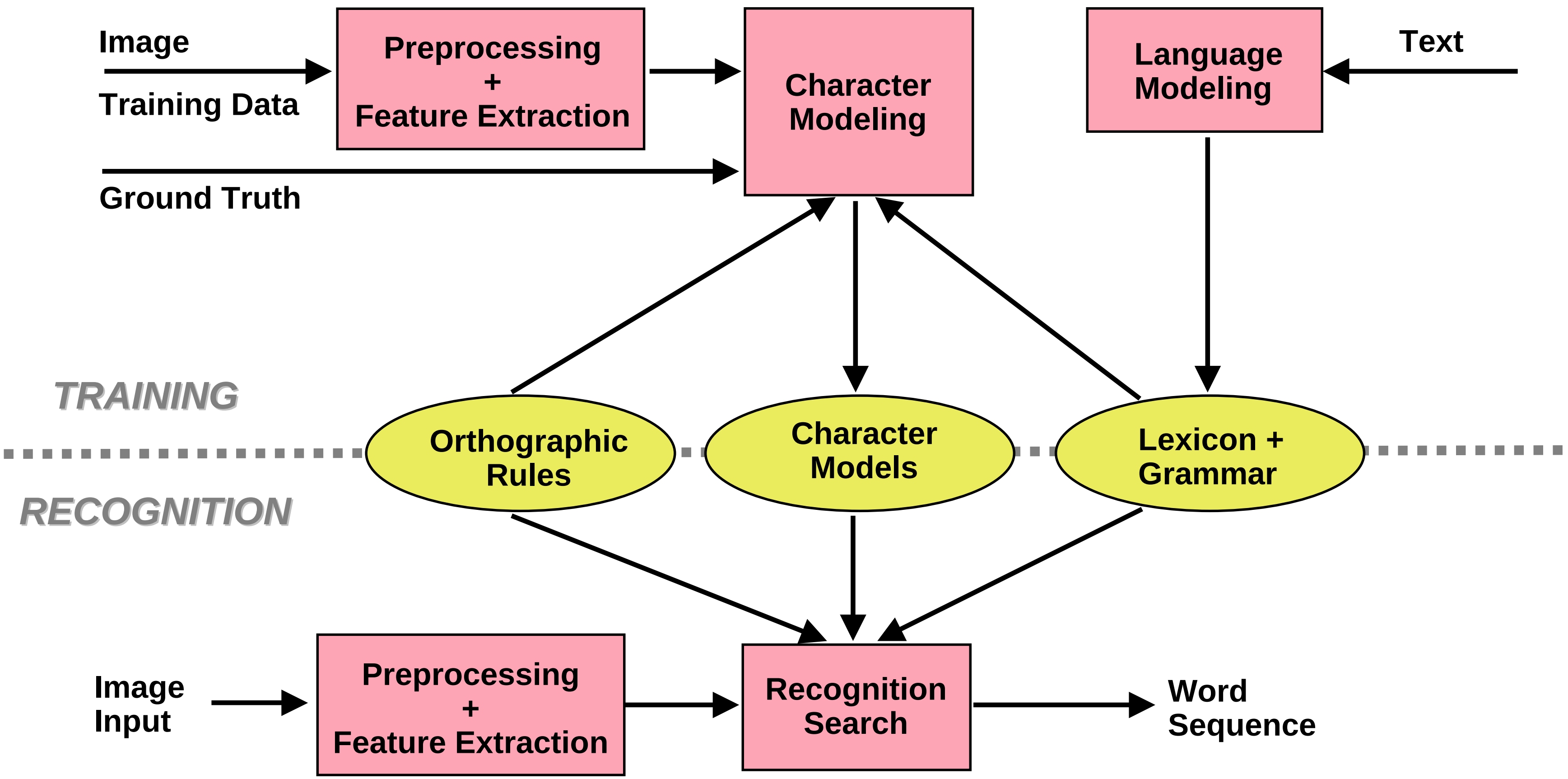


BBN's Videotext Recognition Methodology

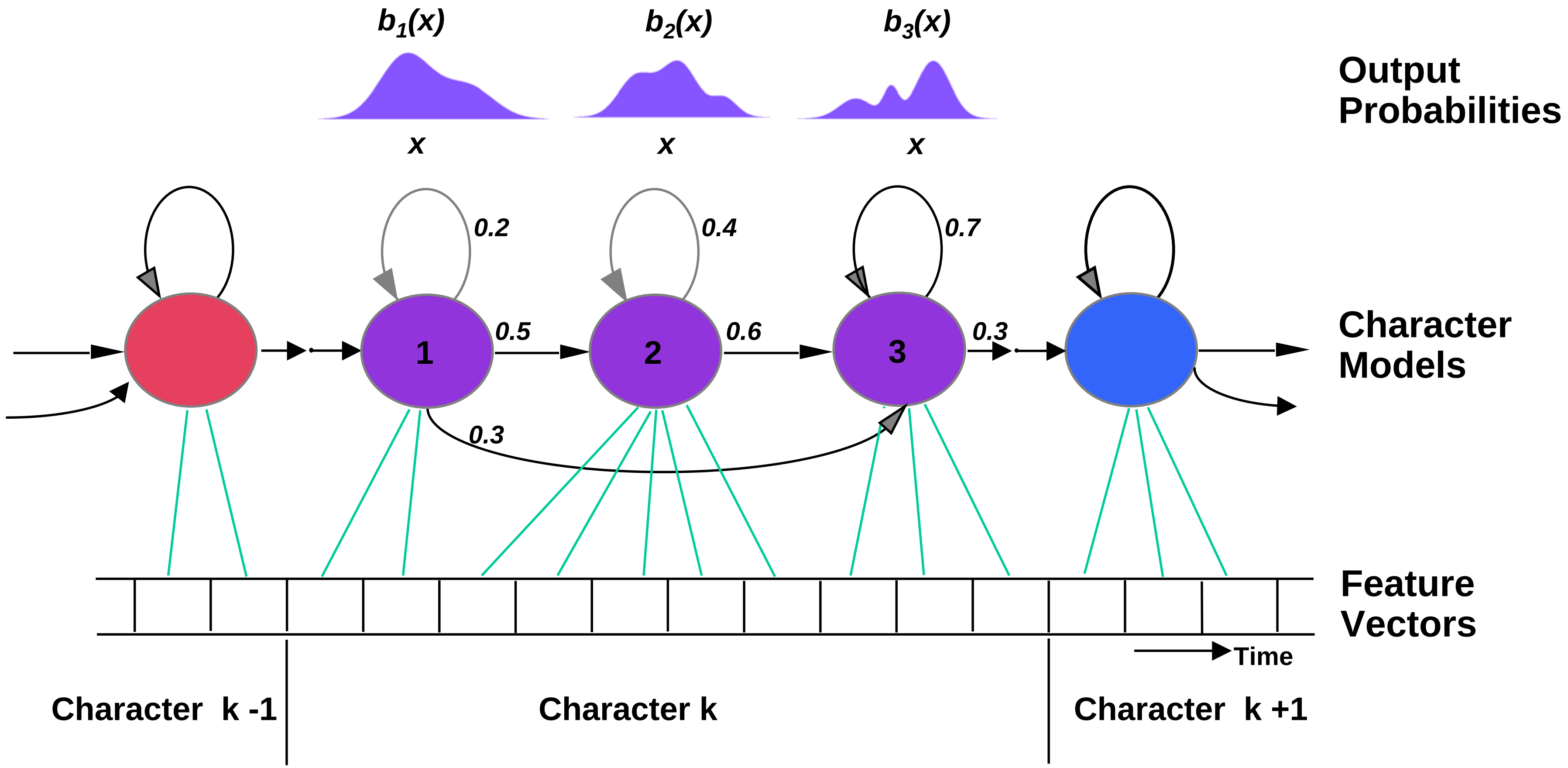
- Employs Hidden Markov Model (HMM) based BBN Byblos Optical Character Recognition (OCR) engine
  - Script-independent, trainable methodology
- Customized videotext pre-processing
  - Upsampling: 4x4 upsampling using bilinear interpolation or FFT-based filtering
  - Gray scale conversion: RGB to YIQ, with only Y (Luminance) used for converting color images to Gray scale
  - Binarization: thresholds on pixel intensity for representing the text object using binary (0 or 255) pixel intensity values


Text Regions
Upsample
Gray Scale Conversion
Binarize and Pad
LineFinding and Feature Extraction
Decoding
Character Sequence
Customized Pre-processing for Videotext Recognition
Standard Steps in BBN Byblos OCR Engine

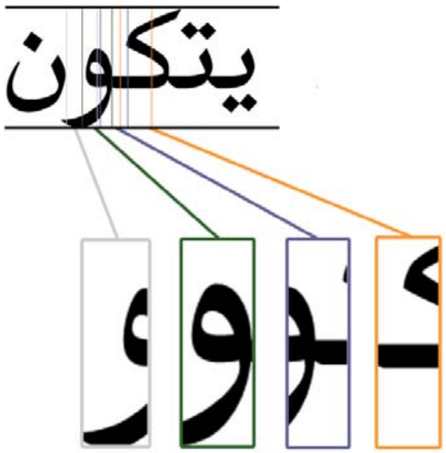
Recognition with BBN Byblos OCR System



Hidden Markov Model of a Character



Feature Extraction



- **Locate line tops and bottoms**
- **Extract narrow overlapping vertical slices of the image**
- **Compute script-independent features on each slice as input to HMM**
- **Linear Discriminant Analysis (LDA) to reduce the dimensionality of the features**

English Videotext Recognition Evaluation

- Evaluation data: Clips from 25 TDT2 videos
  - 12 CNN and 13 ABC
- Development data: 14 CNN and 14 ABC videos
  - Training: ~200K characters, 30K words
  - Test: ~18.5K characters, 3K words
  - Used hand-annotated text regions for training and test
- Submitted recognition output on automatically detected text regions
- More submission plans
  - Results on hand-annotated text regions
  - Results with fast recognition configuration

NOTE: Results in the following slides are obtained on the BBN internal test set and the Dry run test set

English Videotext Recognition – Results

- **Model configuration**
  - Single model trained on data from both channels
  - 14-state, 1 codebook per character tied-mixture (CTM) HMMs, 256 or 512 Gaussians/codebook (G/cbk)
  - Trigram character language model
- **Character Error Rate (CER) measured on 5th I-frame of the text object**
- **256 G/cbk configuration used to submit results on the evaluation data**

	%CER	
Channel	*256 G/cbk*	*512 G/cbk*
CNN	12.0	11.4
ABC	27.0	26.4
Overall	17.2	16.7

Channel Specific Modeling

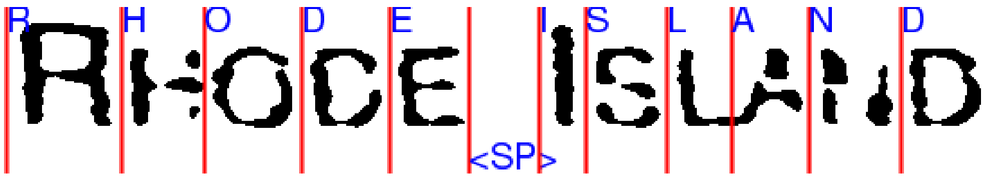
- Estimated separate set of character HMMs for ABC channel
- 14-state, 1 codebook per character HMM with 256 Gaussians/codebook
- Trigram character LM trained on both ABC and CNN

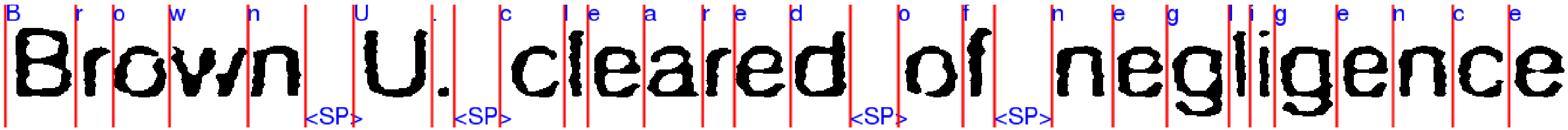
Training	*%CER (ABC only)*
ABC+CNN	27.2
ABC	24.9

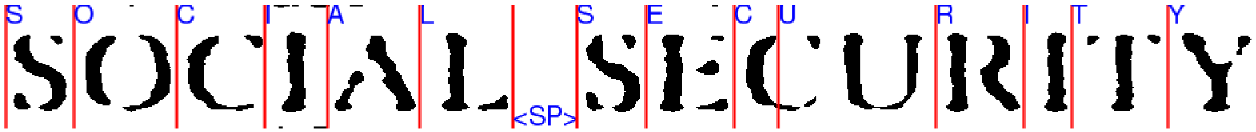
Word-level Segmentation

- Text recognition evaluation scheme uses word-level segmentation information to match detected text box to reference
  - But detection module produces boxes that contain an arbitrary number of words
- OCR decoder automatically produces frame-level (feature vector) segmentations
- Modified feature extraction and recognition software to preserve pixel boundary information
- Added new code to map frame-level segmentation to pixel location on input image

Word-level Segmentation Examples




B r o w n <SP> U . <SP> c l e a r e d <SP> o f <SP> n e g l i g e n c e
Brown U. cleared of negligence

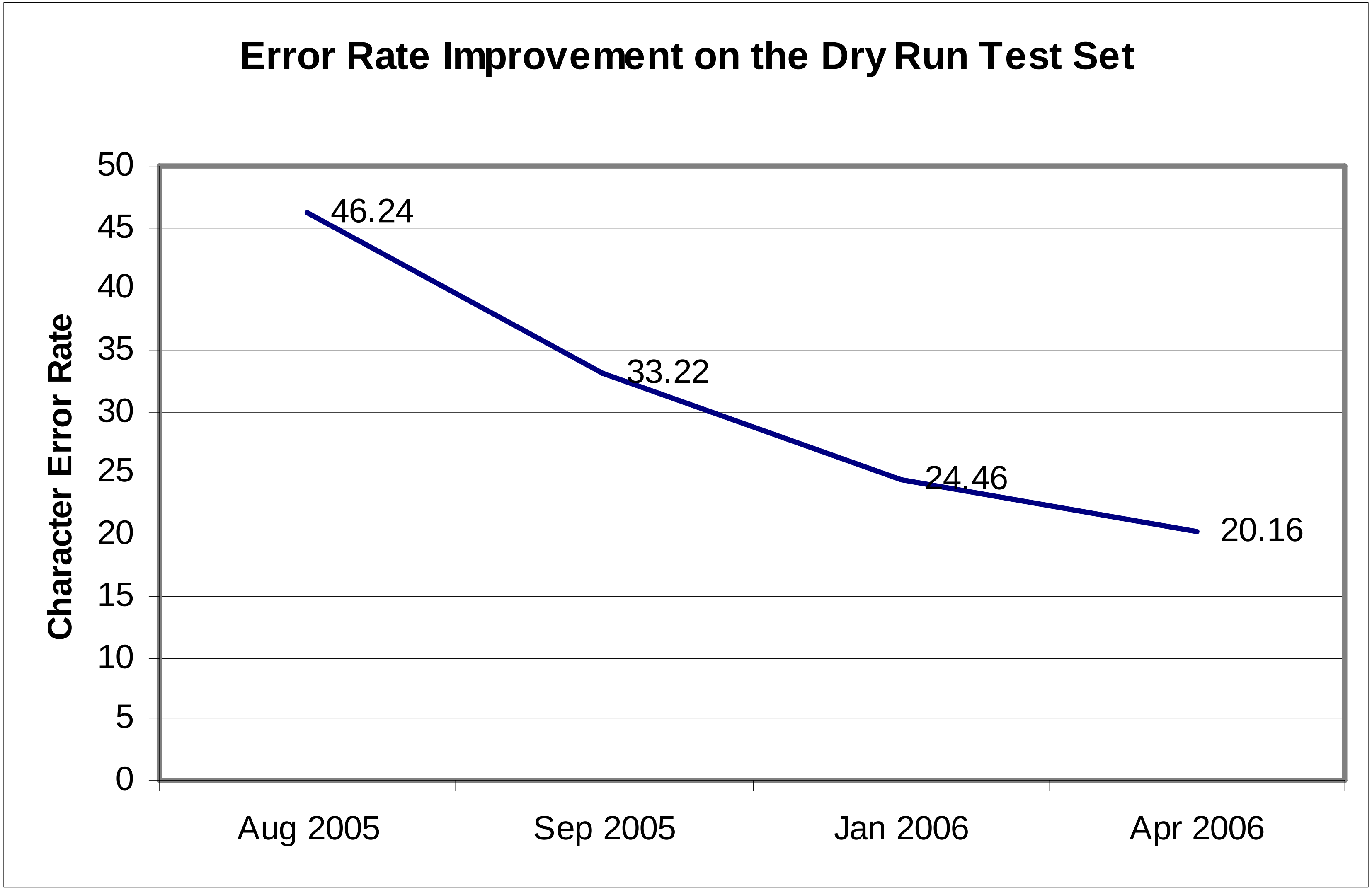


Decoding Speed Improvements on English

- **Fast Gaussian Computation (FGC) using Gaussian shortlists estimated from training data**
- **Tied-mixture (TM) model in forward pass**
  - **Forward-pass: 14-state HMMs, 1 codebook shared across all characters, 1024 Gaussians**
  - **Backward-pass: 14-state HMMs, 1 codebook per character, 512 Gaussians/codebook**

Configuration	*%CER*	*Char/sec.*
Baseline 1 (256 G/cbk)	17.2	23
Baseline 2 (512 G/cbk)	16.7	12
+ Fast Gaussian Computation	17.2	71
+ Tied-mixture Forward Pass	17.3	162

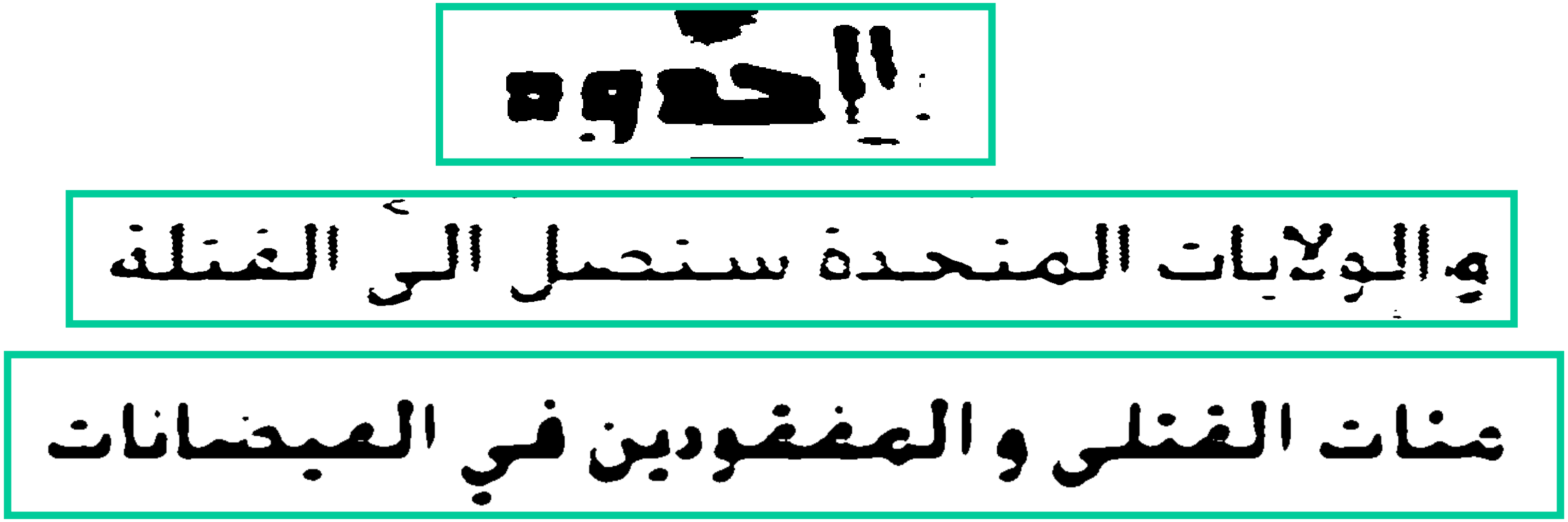
English Videotext Recognition Progress Graph



Arabic Videotext Recognition – Corpus

- **Annotated and transcribed Arabic videotext objects in recorded sequences from Al-Jazeera**
  - **Total Corpus: ~8.3K words, 48.6K characters**
  - **Training: ~7K words, 41K characters**
  - **Test: ~1.3K words, 7.6K characters**

Sample Binarized Videotext Objects



الإخوة

والولايات المتحدة ستحمل الى الغتلف

مئات القتلى و المفقودين في المستشفيات

Arabic Videotext Recognition – Results

- Modeled each presentation form of Arabic character with a separate HMM
  - Total of 167 character forms
  - Model Configuration: 14-states, 1 codebook per HMM, 256 Gaussians/codebook
- Trained Arabic-only model to evaluate performance on Arabic text
  - CER: 21.1%

Conclusions and Future Work

- Improved CER on English videotext recognition by more than a factor of 2
  - Improved upsampling, binarization, linefinding, feature set, and models
  - Increased amount of training data by a factor of 10
- Factor of ~8 speed-up in decoding rate
- Future Work
  - Improve Arabic videotext detection and recognition
  - Iteratively tune end-to-end system to improve overall performance
  - Develop videotext understanding and object classification modules