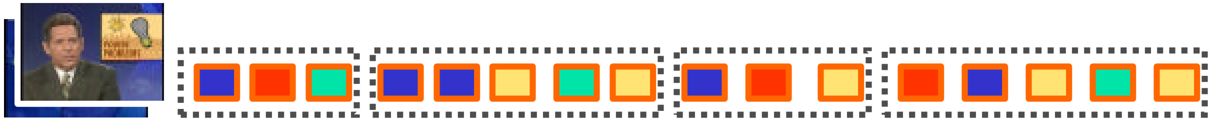


Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation

- *@TRECVID 2003 Workshop*

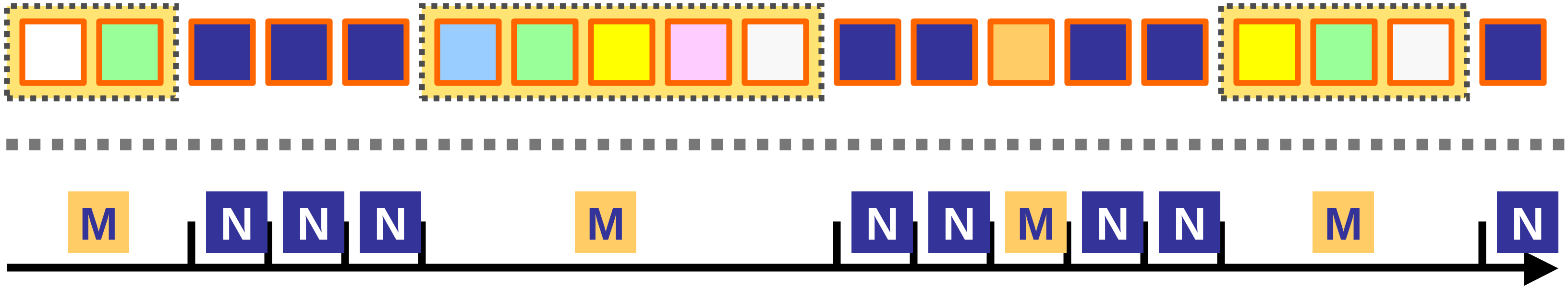
Winston Hsu[1], Shih-Fu Chang[1], Lyndon Kennedy[1], Chih-wei Huang[1], Ching-Yung Lin[2], and Giridharan Iyengar[3]

[1]Dept. of Electrical Engineering, Columbia University, New York, NY
[2]IBM T. J. Watson Research Center, Hawthorne, NY
[3]IBM T. J. Watson Research Center, Yorktown Heights, NY

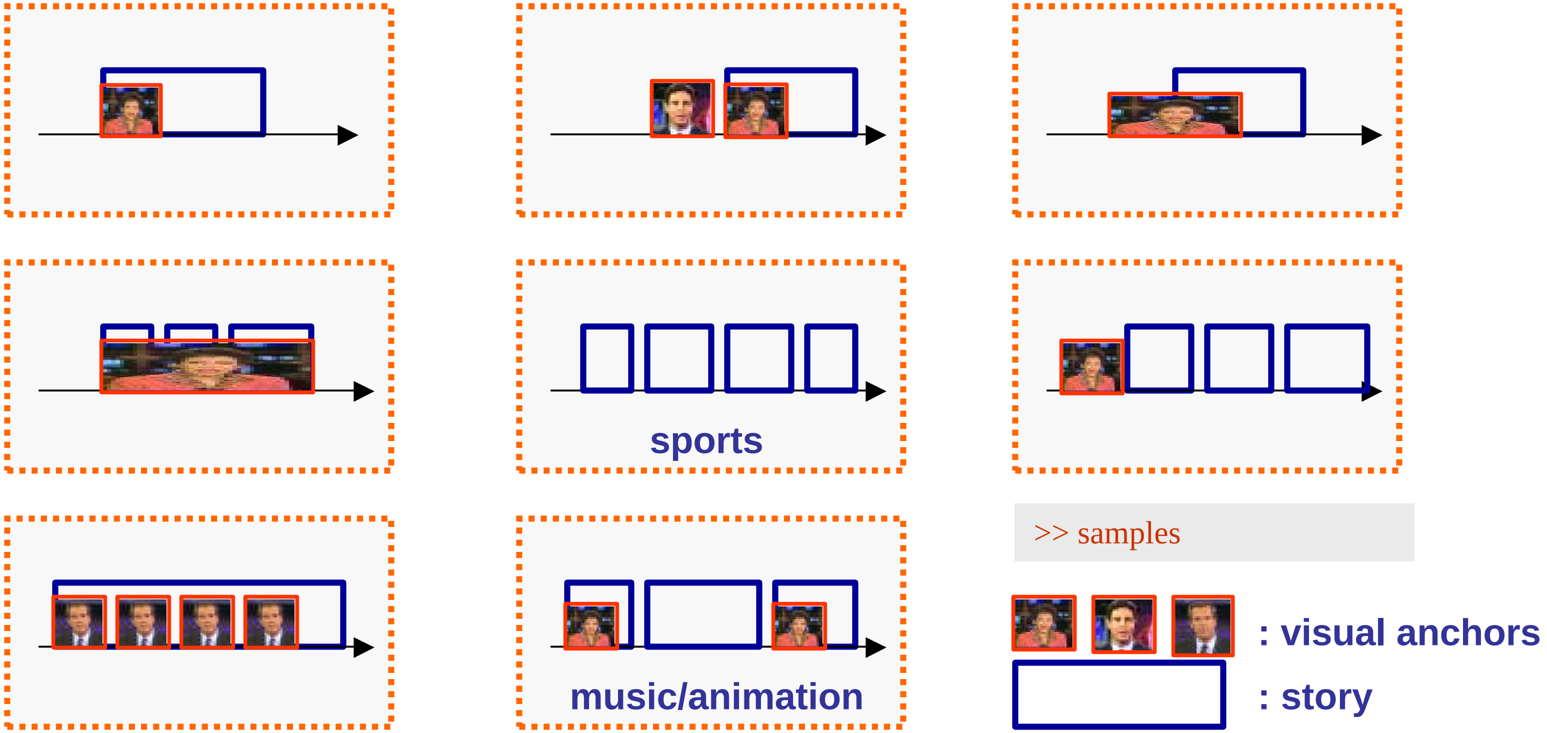
11/17/2003

Story Segmentation

- Story definition (from LDC)
  - A News story is defined as a segment of a news broadcast with a coherent news focus which contains at least two independent, declarative clauses.
  - Misc. segments like commercials, reporter chitchat, station identifications, public service, long musical (>9 sec), interludes, etc…



Challenging problems due to diverse syntax



* Visual anchors alone account for 51% and 67% of stories only on ABC/CNN

Modalities	Set	P	R	F1
Anchor Face	ABC	0.67	0.67	0.67
	CNN	0.80	0.38	0.51

Our Goal

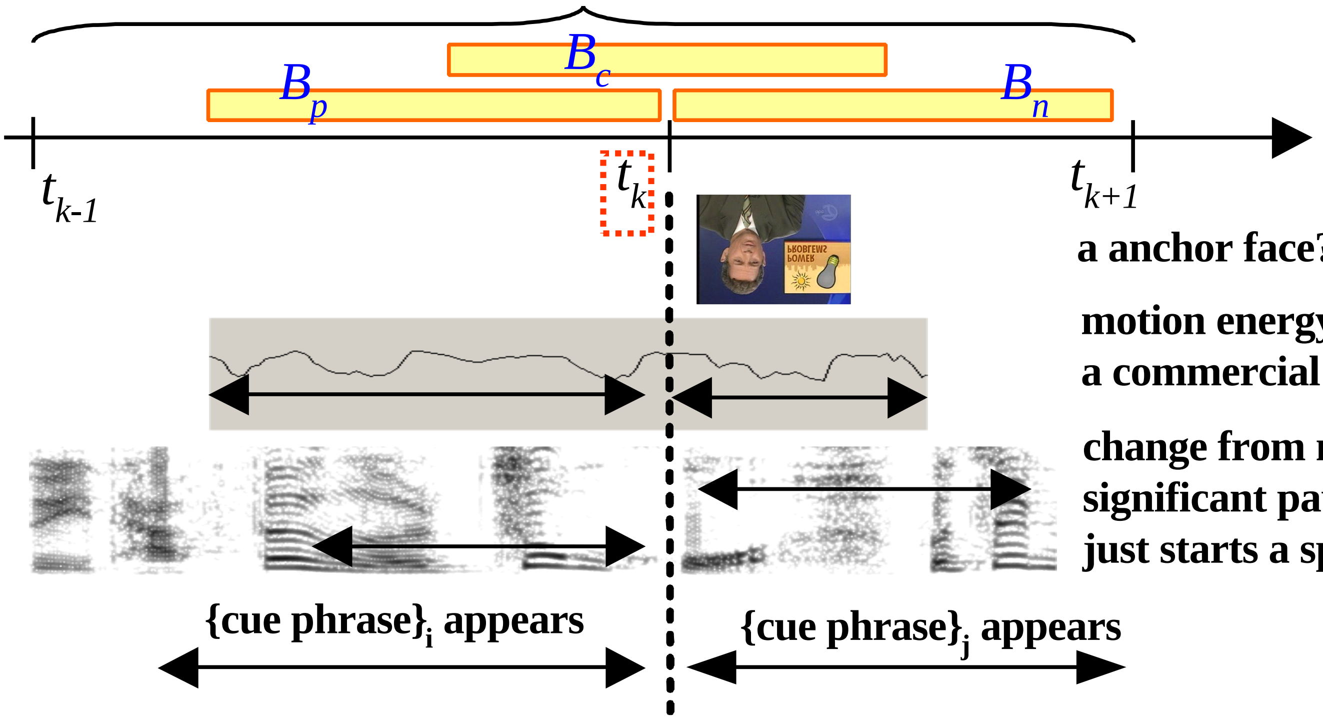
- A robust statistical framework to fuse diverse features from different modalities
- An unified framework that can be adapted to different new video sources
  - Automatically generate customized models (parameters) for CNN and ABC channels with the same framework
- An efficient mechanism for inducing dominant features for any specific domain
  - Allow us to handle large pools of features smoothly
  - Allow us to incorporate computational noisy feature detectors

::More information,
"*Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation*," invited talk, Jan. 18-22, San Jose, SPIE/Electronic Imaging 2004.

Need for Multi-modal Fusion

- Issue: a story boundary at the candidate point t_k?
  - Use the perceptual multi-modal features computed from surrounding windows to infer decisions

with observation x_k to estimate posterior probability $q(b|x_k)$



a anchor face?

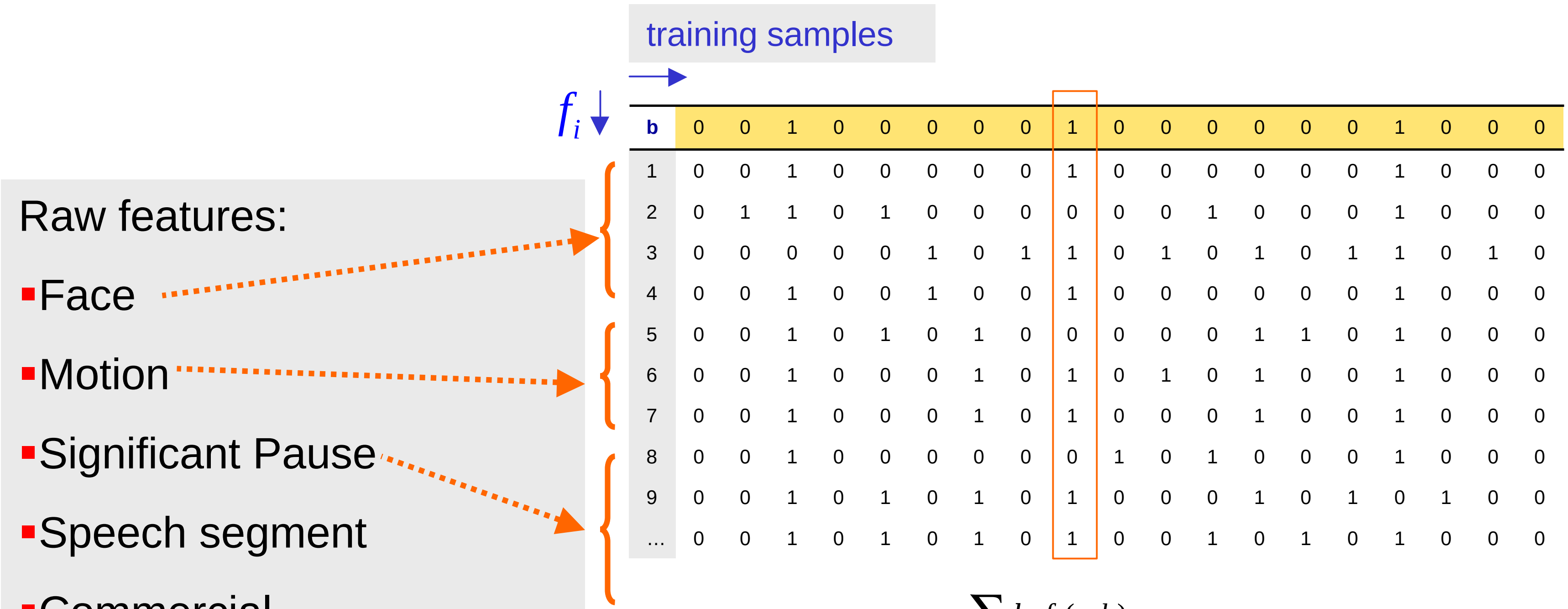
motion energy changes?
a commercial starts in 15 sec.

change from music to speech?
significant pause occurs?
just starts a speech segment?

{cue phrase}$_i$ appears

{cue phrase}$_j$ appears

Our Proposed Framework – Exponential Model w/ Perceptual Binary Features



training samples

f_i

Raw features:

- Face
- Motion
- Significant Pause
- Speech segment
- Commercial
- Text segmentation score
- …

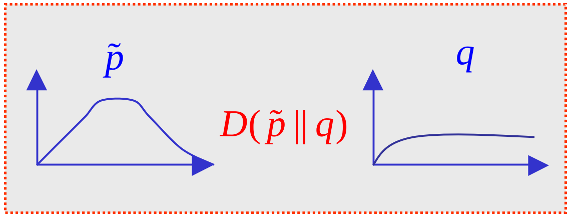
b	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
2	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
3	0	0	0	0	0	1	0	1	1	0	1	0	1	0	1	1	0	1	0
4	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
5	0	0	1	0	1	0	1	0	0	0	0	0	1	1	0	1	0	0	0
6	0	0	1	0	0	0	1	0	1	0	1	0	1	0	0	1	0	0	0
7	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	0
8	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1	0	0	0
9	0	0	1	0	1	0	1	0	1	0	0	0	1	0	1	0	1	0	0
…	0	0	1	0	1	0	1	0	1	0	0	1	0	1	0	1	0	0	0

$$q_l(b \mid x) = \frac{1}{Z_l(x)} e^{\sum_i l_i \cdot f_i(x,b)}, \; f_i(x,b), b \in \{0,1\}$$

* Use supervised learning to find optimal exponential weight l_i for binary feature i

Parameter Estimation

- Estimate $q_l(b|x)$ from training data $T=\{(x_k,b_k)\}$ by minimizing Kullback-Leibler divergence, defined as



$$D(\tilde{p}\,\|\,q_l)=\sum_x\sum_b \tilde{p}(b,x)\log\frac{\tilde{p}(b|x)}{q_l(b|x)}$$

$$=-\sum_x\sum_b \tilde{p}(x,b)\log q_l(b|x)+\text{constant}(\tilde{p})$$

- Also maximize the log-likelihood (with max extreme “0”)

$$L_{\tilde{p}}(q_l)\equiv\sum_x\sum_b \tilde{p}(x,b)\log q_l(b|x)$$

estimated model

empirical distribution

- Iteratively find l_i

$$l_i' = l_i + \Delta l_i \qquad \Delta l_i = \frac{1}{M}\log\left(\frac{\sum_{x,b}\tilde{p}(x,b)\,f_i(x,b)}{\sum_{x,b}\tilde{p}(x)\,q_l(b|x)\,f_i(x,b)}\right)$$

- Because of the convexity of objective function, the iterative process is guaranteed to the global optima.
- Our Matlab implementations show efficient convergence in ~30 mins when using 30 features and 11,705 training samples

Feature Selection

- Input: collection of candidate features, training samples, and the desired model size
- Output: selected features and their corresponding exponential weights
- Current model q augmented with feature h with weight a ,

$$q_{a,h}(b \mid x) = \frac{e^{a\,h(x,b)} q(b \mid x)}{Z_a(x)}$$

- Select the candidate which improves current model q the most, at each iteration;

$$h^{*} = \arg\max_{h \in C} \left\{ \sup_{a} \left\{ D(\tilde{p} \,\|\, q) - D(\tilde{p} \,\|\, q_{a,h}) \right\} \right\}$$

reduce divergence

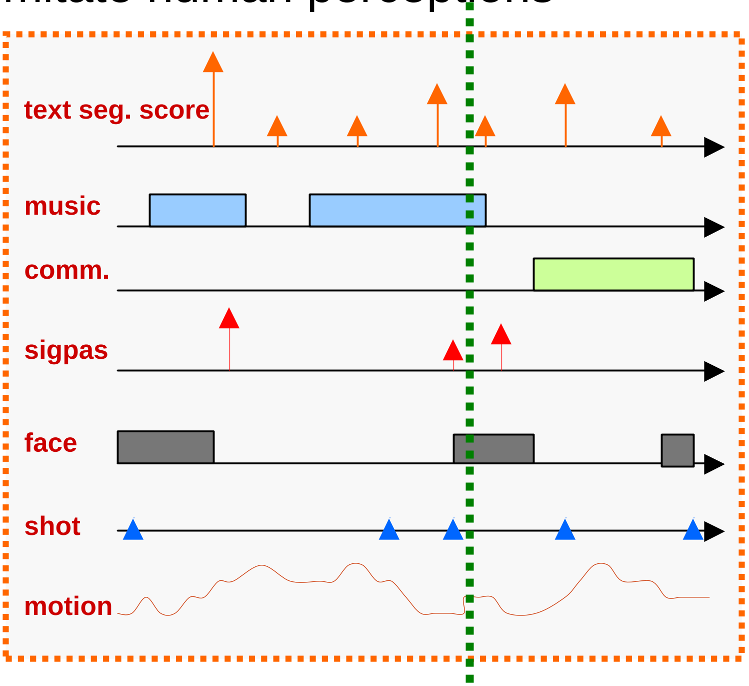
$$= \arg\max_{h \in C} \left\{ \sup_{a} \left\{ L_{\tilde{p}}(q_{a,h}) - L_{\tilde{p}}(q) \right\} \right\}$$

increase log-likelihood

Examples of Raw Features

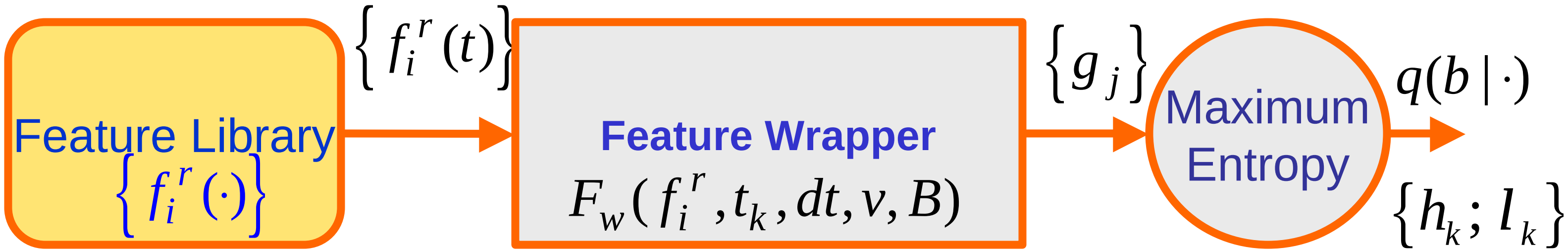
Modality	Raw Features	Time Index	Value
Video	motion	segment	real
	shot boundary	point	boolean
	face	segment	real
	commercial	segment	boolean
Audio	pause	point	real
	pitch jump	point	real
	significant pause	point	real
	musc./spch. disc.	segment	boolean
	spch seg./rapidity	segment	real
Text	ASR cue terms	point	boolean
	V-OCR cue terms	point	boolean
	text seg. score	point	real
Misc.	combinatorial	point	boolean
	sports	segment	boolean

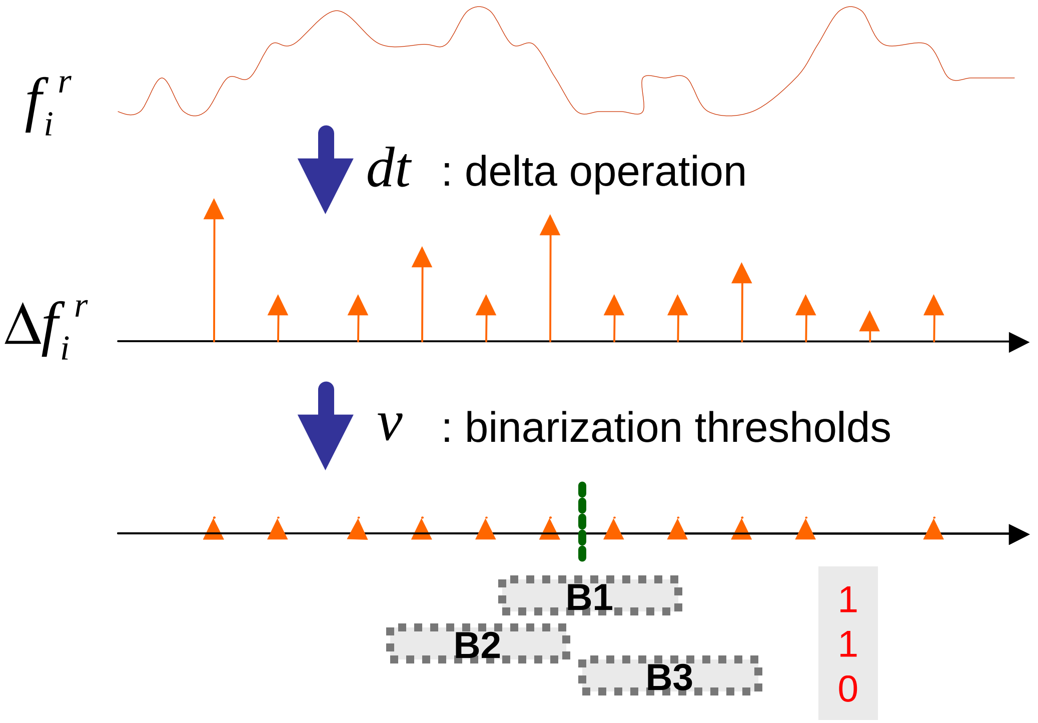
- Features exist at different time scales, asynchronous points
- Need an unified wrapper to convert them to consistent representation & imitate human perceptions



candidate point

Feature Wrapper





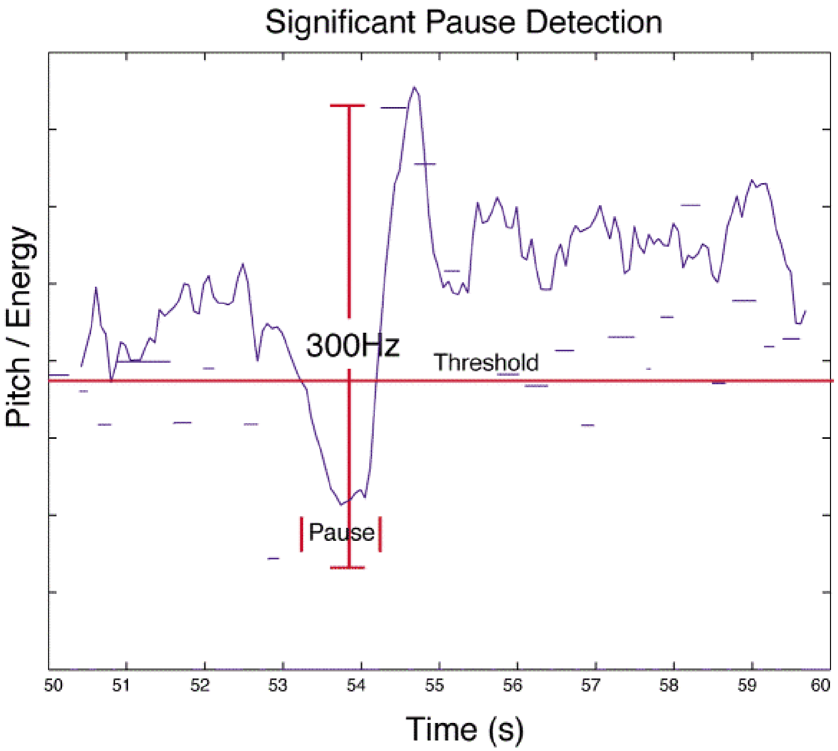
■delta interval:	dt
■observation windows:	B
■ binarization levels:	v
■ raw features:	f_i^r
■ candidate point:	t_k

Selected Features (from CNN)

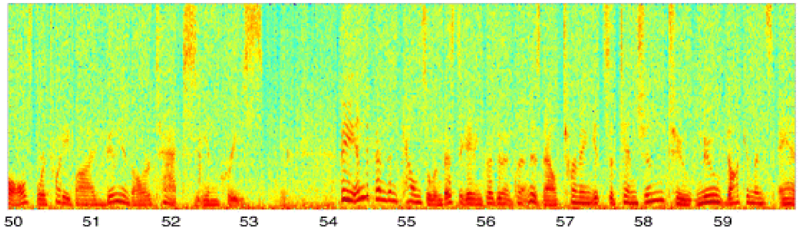
* The first 10 “A+V” features automatically discovered for the CNN channel

	no	raw feature set	gain	*l*	interpretation
>>	1	Anchor Face	0.3879	0.4771	An anchor face segment just starts after the boundary point
>>	2	Significant pause & non-commercial	0.0160	0.7471	A significant pause within the non-commercial section appears in surrounding observation window.
	3	Pause	0.0058	0.2434	An audio pause with the duration larger than 2.0 second appears after the boundary point.
	4	Significant pause	0.0024	0.7947	The surrounding observation window has a significant pause with the pitch jump intensity larger than the normalized pitch threshold 1.0 and the pause duration larger than 0.5 second.
	5	Speech segment	0.0019	-0.3566	A speech segment before the candidate point
	6	Speech segment	0.0015	0.3734	A speech segment starts in the surrounding observation window
>>	7	Commercial	0.0015	1.0782	A commercial starts in 15 to 20 seconds after the candidate point.
	8	Speech segment	0.0022	-0.4127	A speech segment ends after the candidate point
	9	Anchor face	0.0016	0.7251	An anchor face segment occupies at least 10% of next window
	10	Pause	0.0008	0.0939	The surrounding observation window has a pause with the duration larger than 0.25 second.

Significant Pause



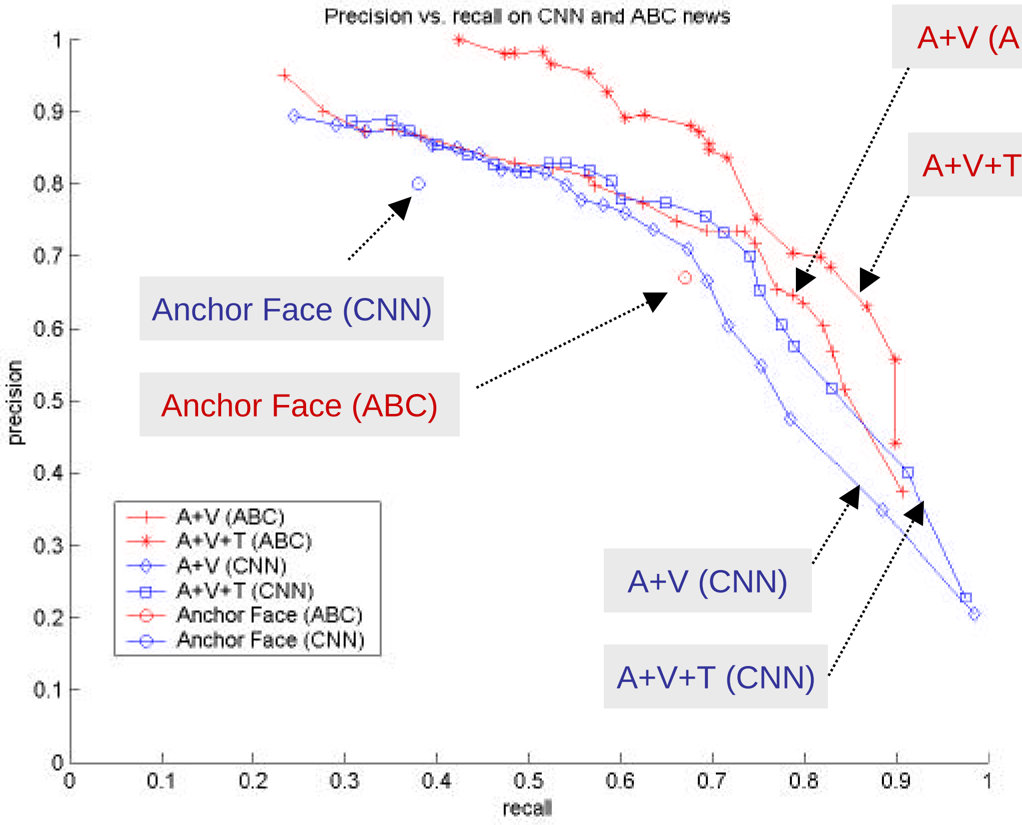
...for $23 billion tax increase. [story change] The independent counsel investigating president...



		Significant Pause			Uniform		
Set	e	P	R	F1	P	R	F1
ABC	5.0	**0.20**	**0.38**	**0.26**	0.10	0.22	0.14
	2.5	0.16	0.34	0.22	0.10	0.22	0.14
CNN	5.0	**0.40**	**0.45**	**0.42**	0.20	0.24	0.22
	2.5	0.37	0.43	0.39	0.20	0.24	0.22

::sigpas_seg_0202cnn

Precision vs. Recall Curves



A+V (ABC)

A+V+T (ABC)

Anchor Face (CNN)

Anchor Face (ABC)

A+V (CNN)

A+V+T (CNN)

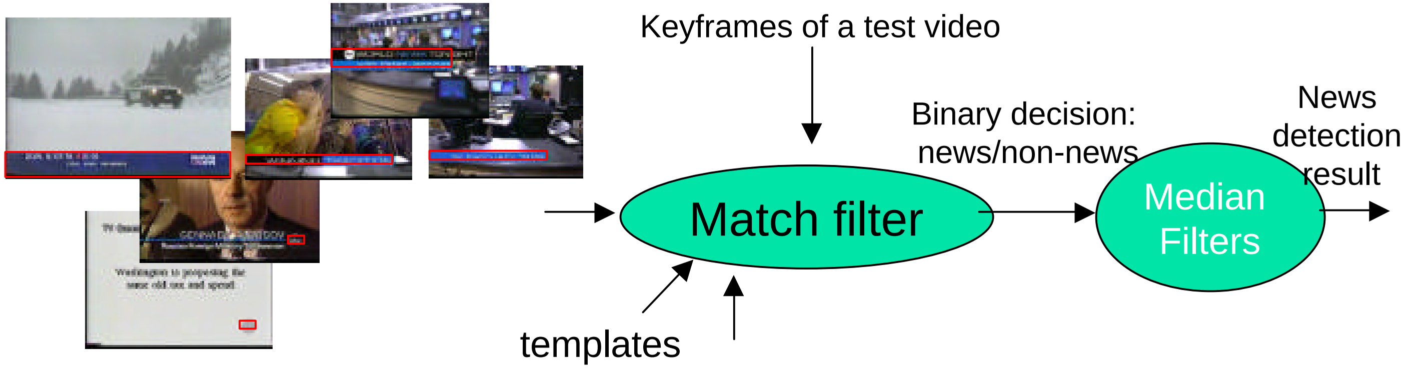
$$F1 = \frac{2 \cdot P \cdot R}{P + R}$$

	ABC	CNN
A+V	0.69	0.63
A+V+T	0.74	0.67
A+V (BM)	0.71	0.69
A+V+T (BM)	**0.76**	**0.73**

decision boundary: $q(b \mid \cdot) > b_m$

- Single point P/R is not sufficient for assessment -> need more samples of P/R curves
- Improvement by multi-modal fusion is significant
  - CNN improves more in high recall area
  - ABC improve more in high precision area

Story Typing



$$A' = (A \circ M_T) \bullet M_T$$

$$M_T = u[n] - u[n-T]$$

$$T = 450 \text{ (frames)}$$

○ : Morphological OPEN
● : Morphological CLOSE

- A story segment is assigned as “News” if overlapping with the non-commercial segments larger than a threshold

$$type_i = 1_{\left\{\frac{|S_i \cap \{S_J^N\}|}{|S_i|} > e_t\right\}}$$

Modalities	Set	P	R	F1
A+V	ABC	0.93	0.92	0.92
	CNN	0.92	0.90	0.91
A+V+T	ABC	0.89	0.94	0.91
	CNN	0.91	0.90	0.90

Summary

- We have developed a statistical framework that can be systematically applied to diverse news video sources
- The results are promising and show multi-modal improvement
  - The same framework can be used to select dominant features from any modalities flexibly
- The performance shows room for further research
  - How to go beyond 75% and reach 90%?
- Evaluation metrics should include complete P/R curves
- Future works
  - Address imbalanced data distributions
  - Explore temporal dynamics in stories
  - Expand feature pool such as speech phoneme rapidity, video OCR, and high-level concept detection, etc.

Acknowledgements

- Thanks to
  - Martin Franz of IBM Research for providing an ASR only story segmentation system
  - Dongqing Zhang of Columbia University for providing the geometric active contour face detection system
  - TRECVID 2003 organizing team for providing the evaluation platform and precious video corpora

Q & A
Thank You!!

*More information:
"*Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation*," invited talk, Jan. 18-22, San Jose, SPIE/Electronic Imaging 2004.