COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

IBM

Adaptive Feature Discovery for TRECVID Broadcast News Video Story Segmentation

@TRECVID Workshop 2004, Nov. 15-16

Winston Hsu[1], Lyndon Kennedy[1], Shih-Fu Chang[1], Martin Franz[3], John Smith[2], 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

http://www.ee.columbia.edu/~winston

Outlines

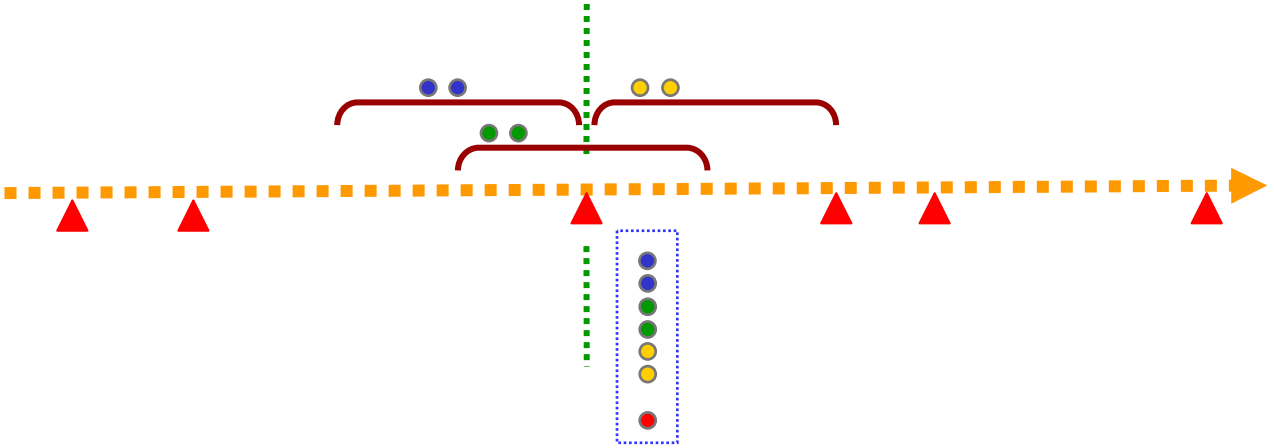
- Features and Fusion Strategies
  - Multi-modal features at different observation windows (e.g., prosody, visual cues, text)
  - Fusion with Support Vector Machines
- New focus in 2004:
  - Automatic Visual Cue Cluster Construction (VC^3 framework)
  - Ability to handle diverse production events
- Thorough error analysis for different genres
- Brief comparison with last year results

Story Segmentation Model

$$T_{pas} \oplus_{\epsilon} T_{sht}$$

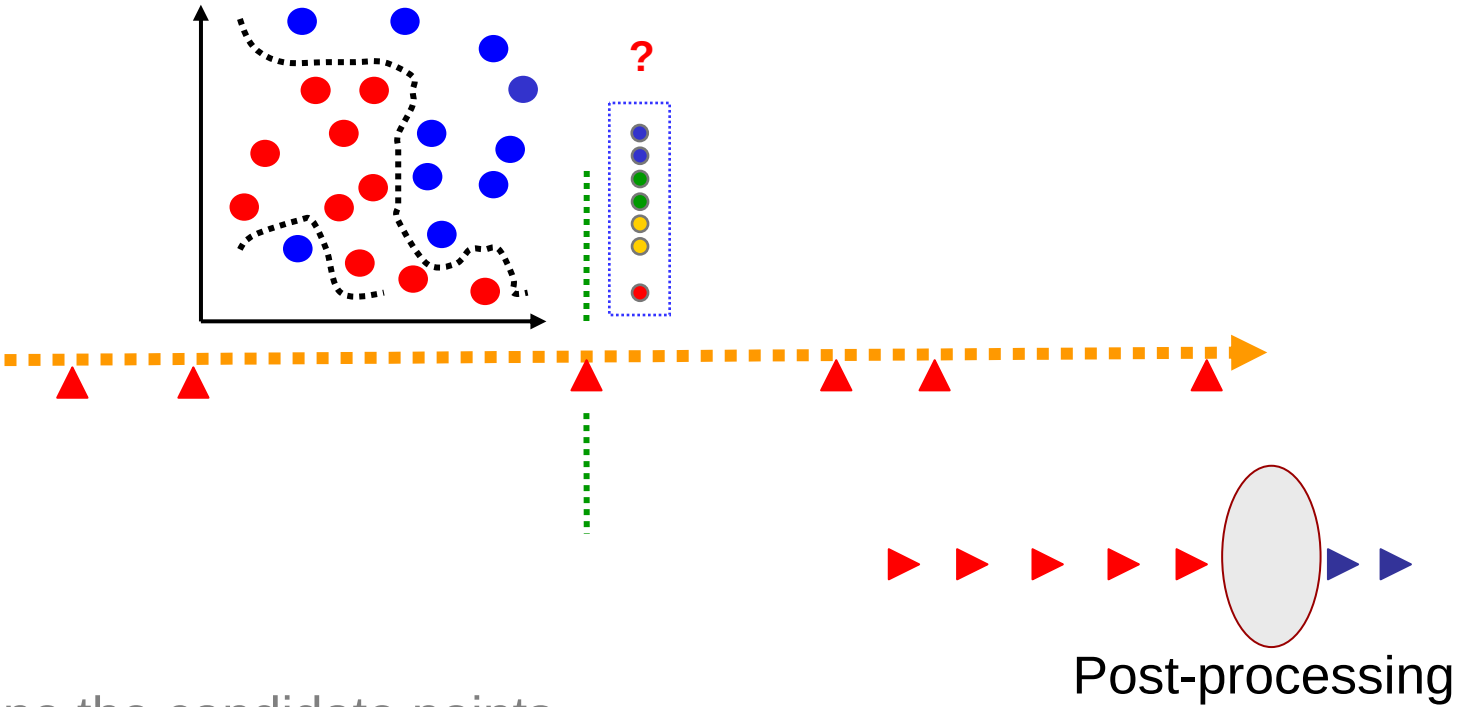
- Determine the candidate points
  - union of pauses and shot boundaries with fuzzy window 2.5 sec

Story Segmentation Model



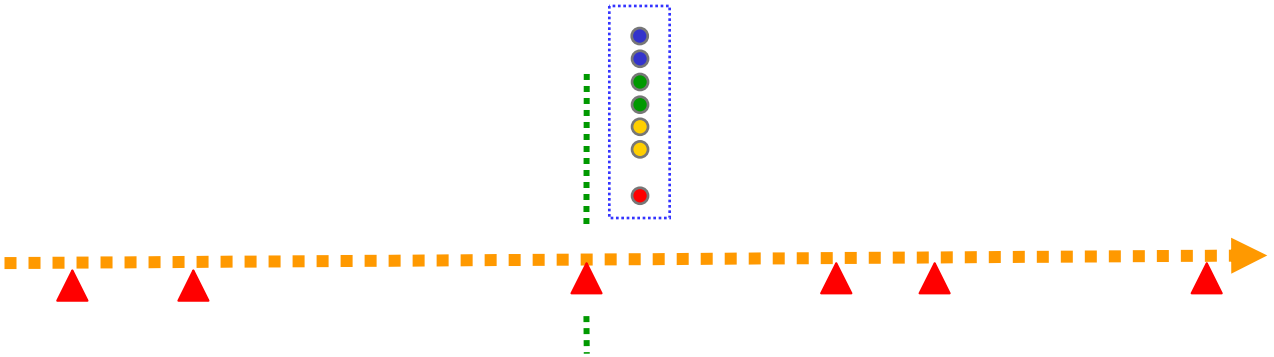
- Determine the candidate points
  - union of pauses and shot boundaries with fuzzy window 2.5 sec
- Extract and aggregate relevant features from surrounding windows
  - take into account asynchronous multi-modal futures; e.g., text, audio

Story Segmentation Model



- Determine the candidate points
  - union of pauses and shot boundaries with fuzzy window 2.5 sec
- Extract and aggregate relevant features from surrounding windows
  - take into account asynchronous multi-modal futures; e.g., text, audio
- Classify the candidate points as “boundary” or “non-boundary”
  - SVMs with RBF kernels
  - Post-processing

Raw Multi-Modal Features



Modality	Raw Features	Dim.
Visual	**Visual Cues Clusters**	15~40
	commercial	2
	motion	2
Audio	pause	1
	prosody features	30
	speaker change	1
	speech rapidity	1
Text	text story seg. scores	1

* before taking into account different observation windows

Visual Cue Cluster Construction (VC^3)

- Motivation

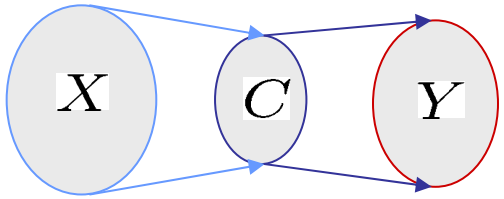
  - News channels usually have different visual production events across channels or time and are statistically relevant to story boundaries
  - Usually try different ways to manually enumerate all the production events from inspections, and then train the classifiers
    - e.g. ANCHOR, STUDIO, WEATHER, CNN_HEADLINE, …, etc.
    - Problems -> deploying on multiple channels of multiple countries …
- We hope to discover a systematic work to catch "visual cue clusters"
  - Analogously, text -> cue words or cue word clusters
  - Automatically, rather than by human inspection
  - Avoid time-consuming news production annotations

via Information Bottleneck Clustering!

VC³: the Information Bottleneck Principle

- Cluster X to C but still trying to preserve the mutual information with label space Y

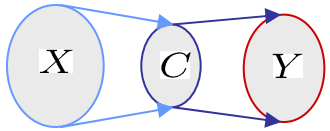
$$C^* = \underset{C}{\text{argmin}}\{I(X;C) - \beta I(Y;C)\}$$



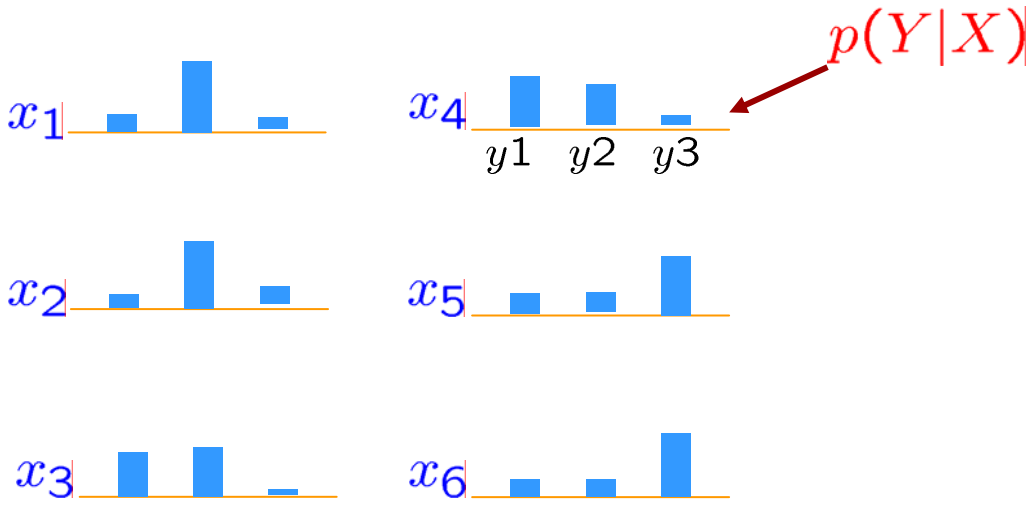
- If $\beta \to \infty$, a hard partitioning; we only care about maximizing $I(Y;C)$; that's to minimize $H(Y|C)$

$$\begin{aligned} I(Y;C) &= H(Y) - H(Y|C) \\ &= \sum_{c\in C}\sum_{y\in Y} p(c,y)\log\frac{p(c,y)}{p(c)p(y)} \\ &= \sum_{c\in C} I(Y;c) \end{aligned}$$

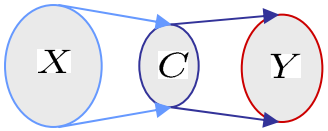
VC^3 Overview: a Simple Example



$$I(Y;X) = 11.93$$

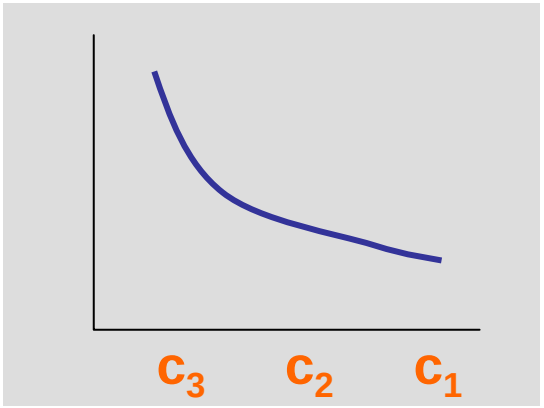


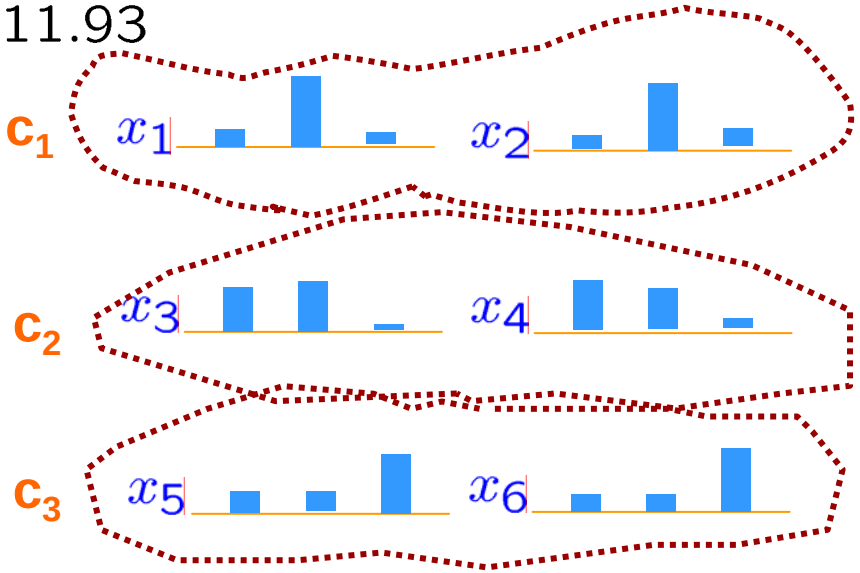
VC^3 Overview: a Simple Example



$$I(Y;X) = 11.93$$

$I(Y;c_i)$





$$I(Y;C^*) = 11.90$$

- Items (features) in the same cluster tend to be with similar probability distributions over the event labels ***Y***->semantic consistency!!
- MI contributions from different clusters -> feature selection

Columbia University in the City of New York

VC^3 Overview: Joint Probability Approximation

- For IB clustering, we essentially need $P(X,Y), P(X), P(Y)$
  - However, video features are not discrete but continuous!

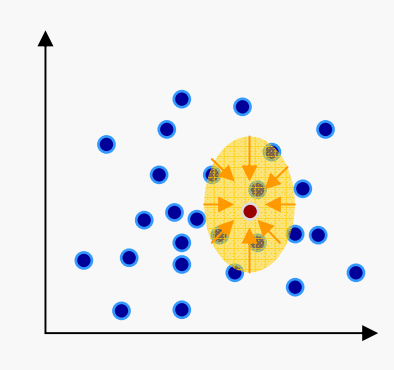
- Approximate joint probability via kernel density estimation from existent feature observations

$S = \{x_1, ..., x_i, ..., x_{|S|}\}$

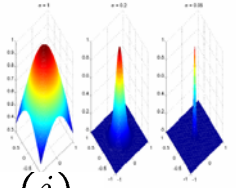
$$p(x,y) = \frac{1}{Z(x,y)} \sum_{x_i \in S} K_\sigma(x - x_i) \cdot p(y|x_i)$$

Gaussian Kernel with specific kernel bandwidth

observed event probability conditioning on the feature



- Embed prior knowledge on kernels functions and the kernel bandwidth (*D*-dimensional)



  - Gaussian Kernel (diagonal): $K_\sigma(x_r - x_i) = \prod_{j=1}^{D} \exp \frac{-||x_r^{(j)} - x_i^{(j)}||}{\sigma_j}$
  - Raw features: autocorrelogram, color moments, and Gabor texture

VC³ Overview: Cluster Examples-I

- ABC VCs for story seg.

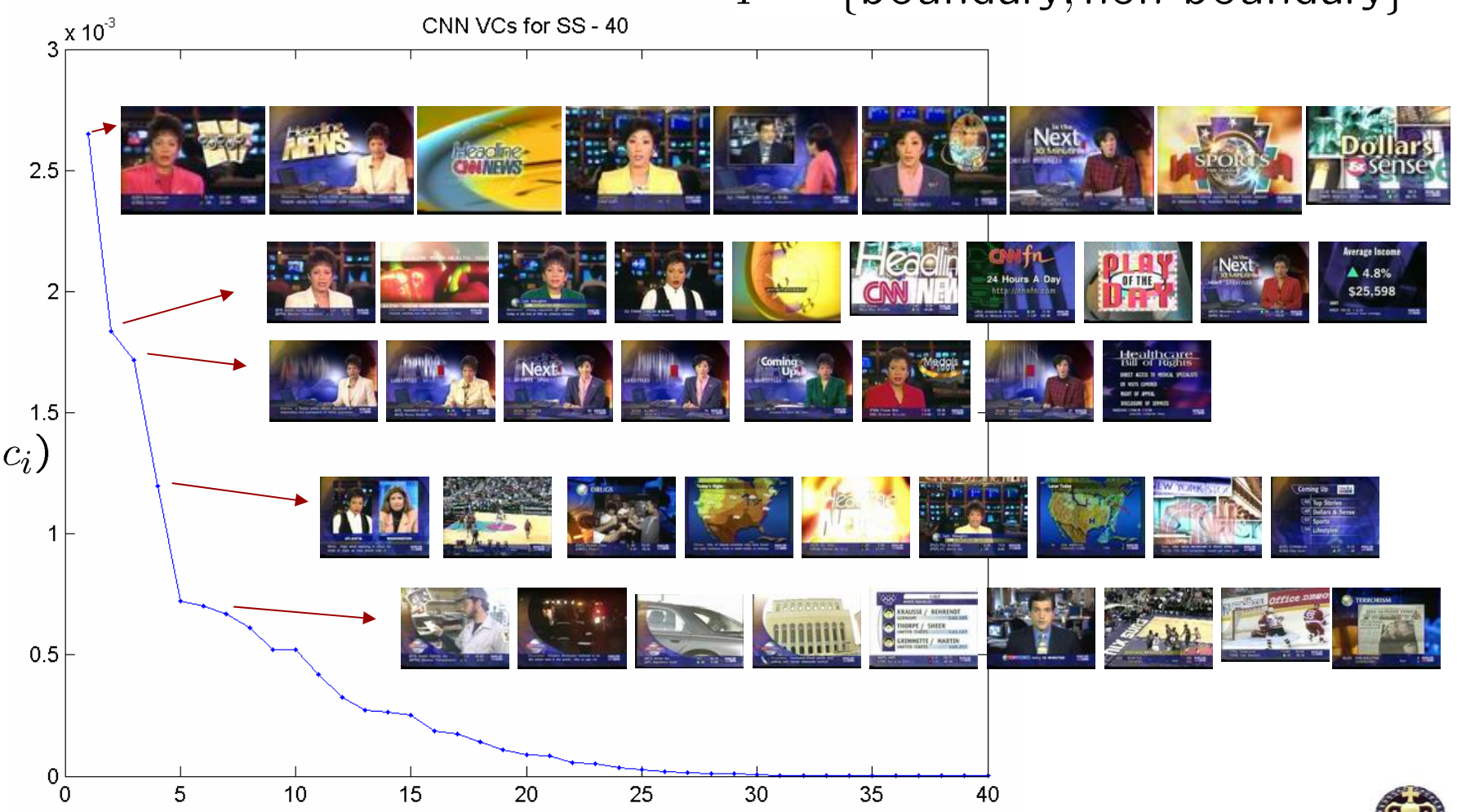
$$X = \{\text{key frame features}\}$$
$$Y = \{\text{boundary}, \text{non-boundary}\}$$

Columbia University in the City of New York

VC^3 Overview: Cluster Examples-II

- CNN VCs for story seg.

$$X = \{\text{key frame features}\}$$
$$Y = \{\text{boundary}, \text{non-boundary}\}$$

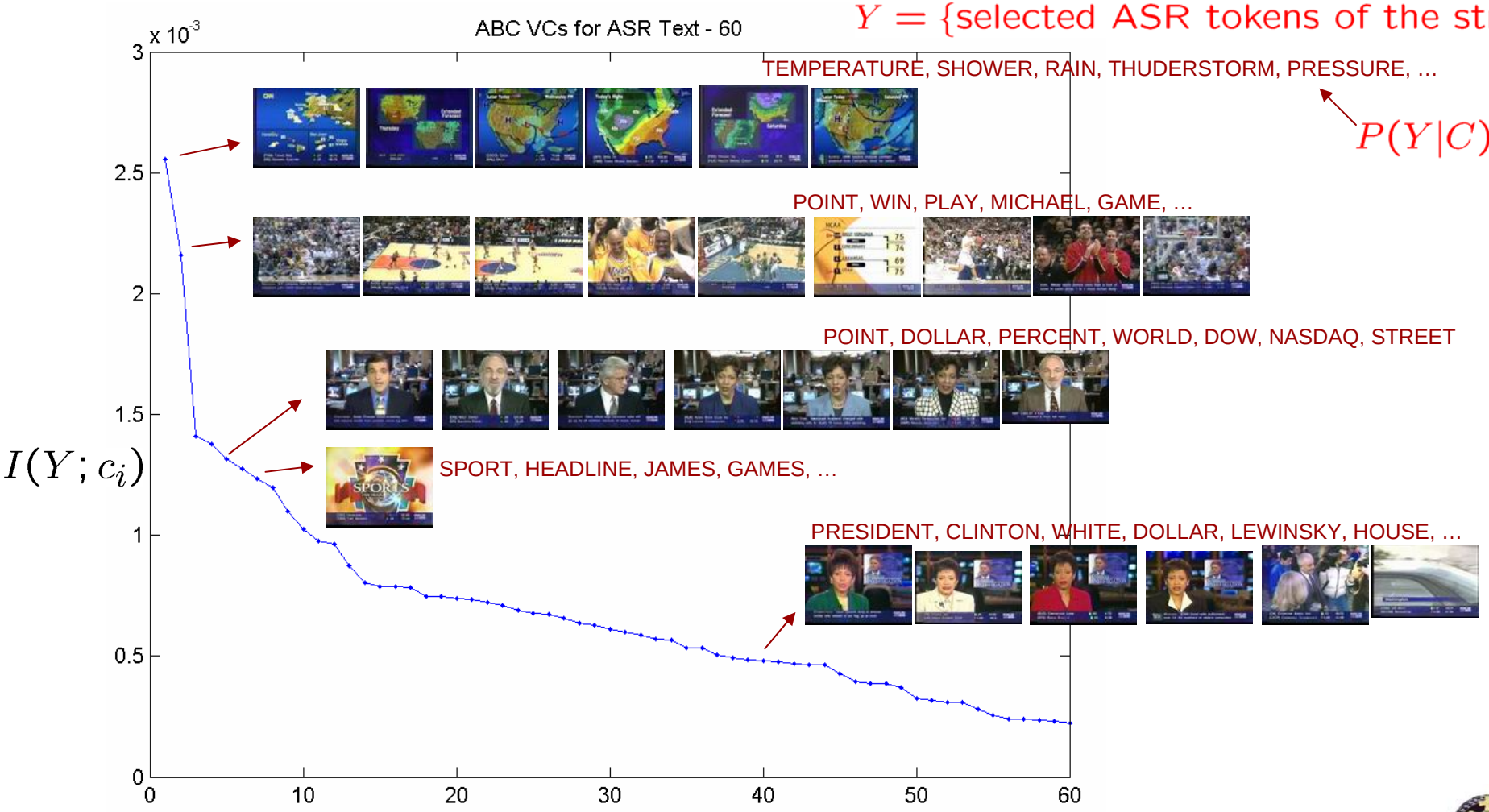


VC³ Overview: Cluster Examples-III

- CNN VCs for text association

$X = \{\text{key frame features}\}$

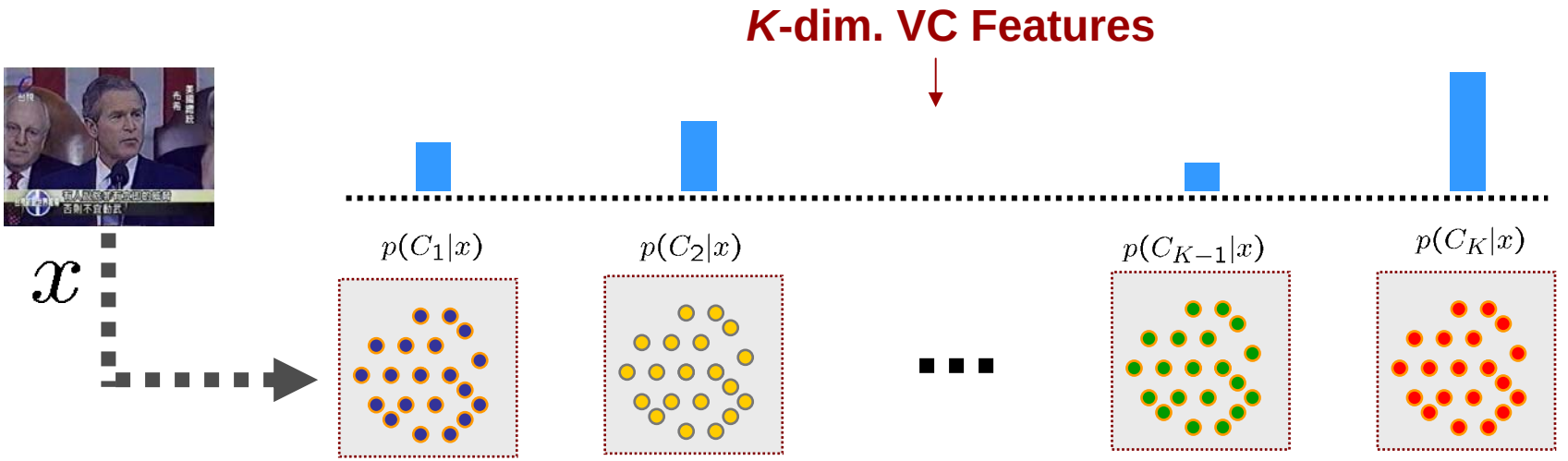
$Y = \{\text{selected ASR tokens of the stry.}\}$



Columbia University in the City of New York

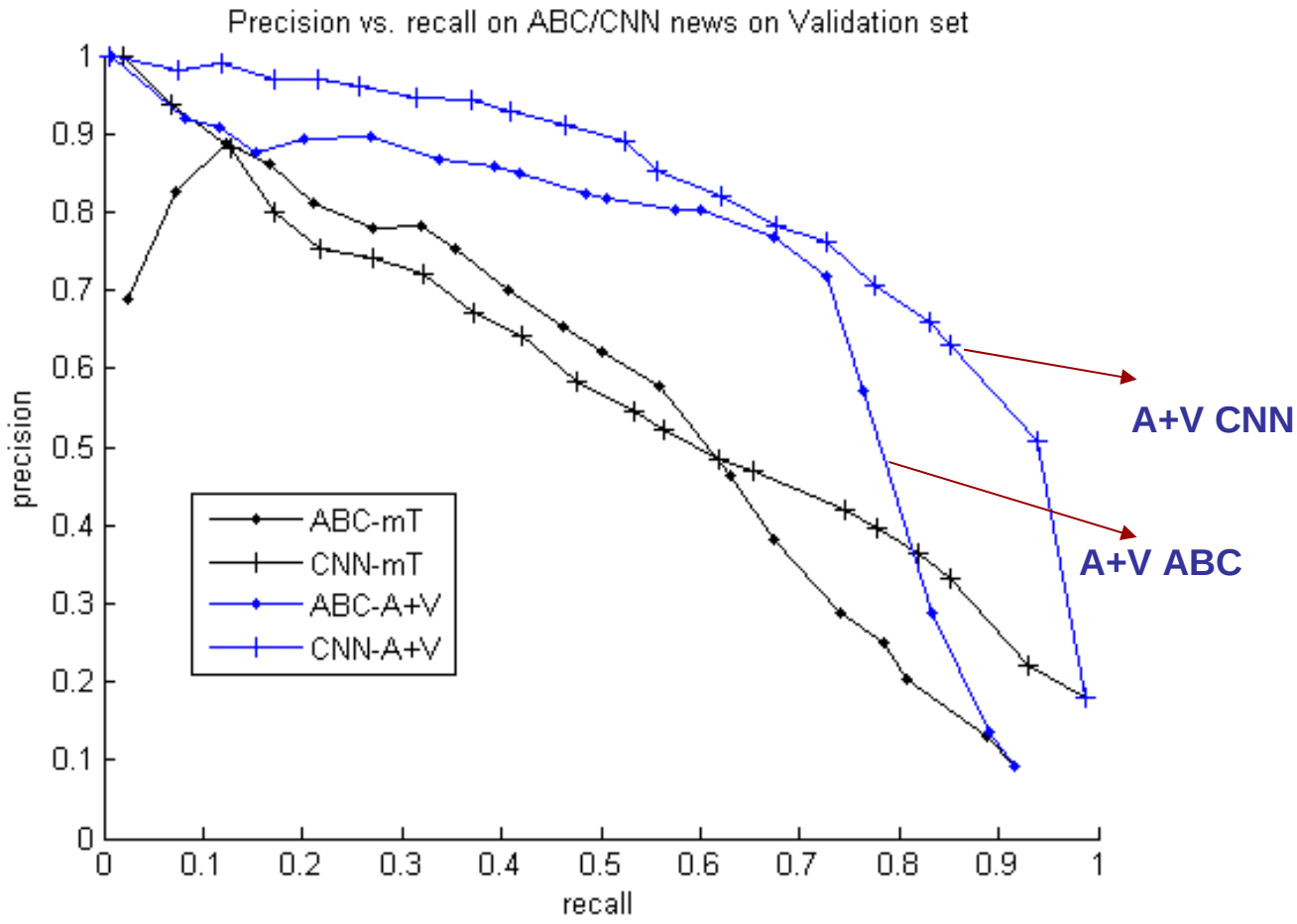
VC^3 Overview: Feature Projection

- In feature extraction, project an image to those induced cue clusters by calculating the membership probabilities

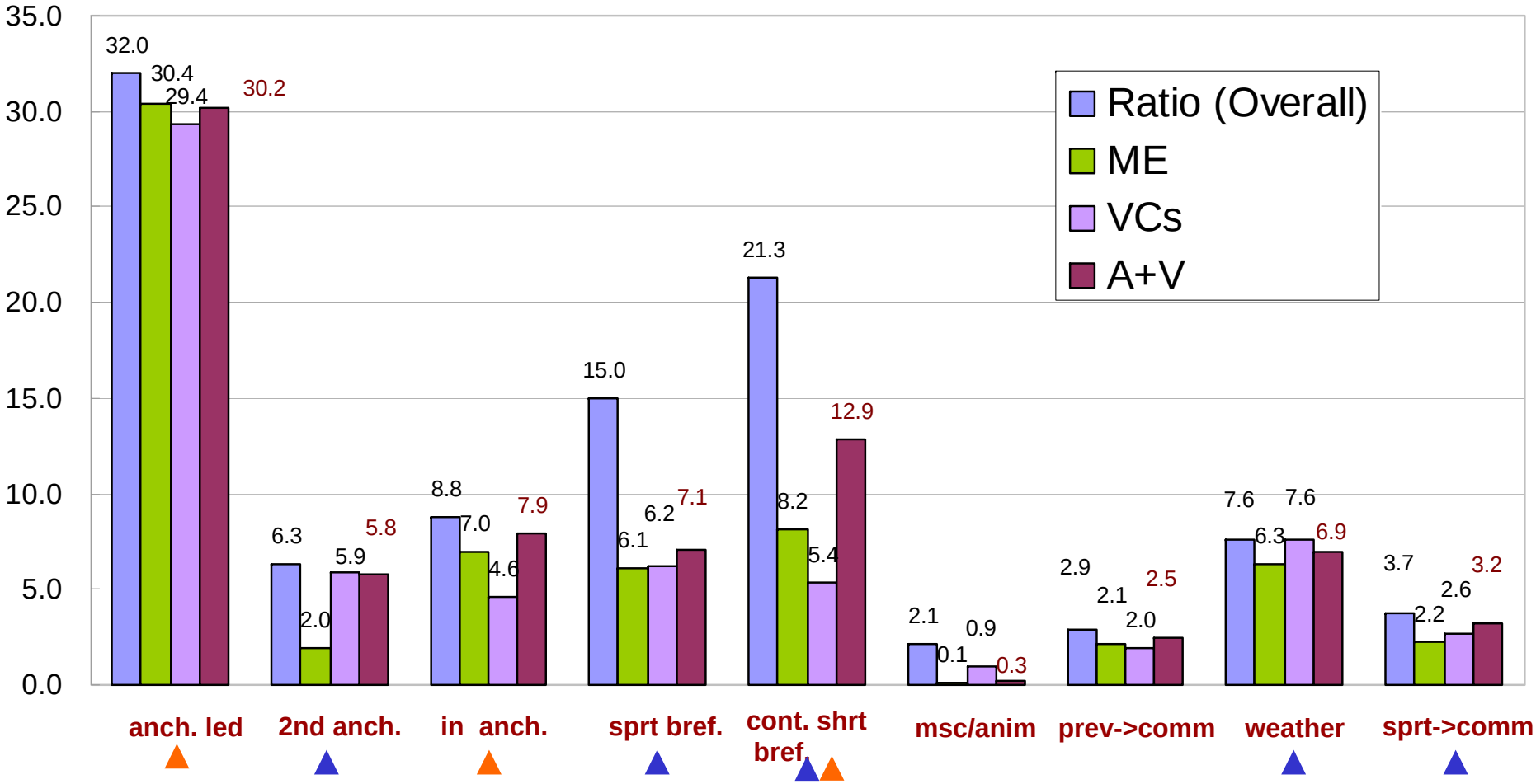


$$p(C_j|x) = \frac{p(x|C_j)p(C_j)}{p(x)} \qquad p(x|C_j) = \frac{1}{Z_j}\sum_{x_i \in C_j} K_\sigma(x - x_i)$$

Performance Overview (A+V, Validation Set)



Performance Overview (A+V, Validation Set)

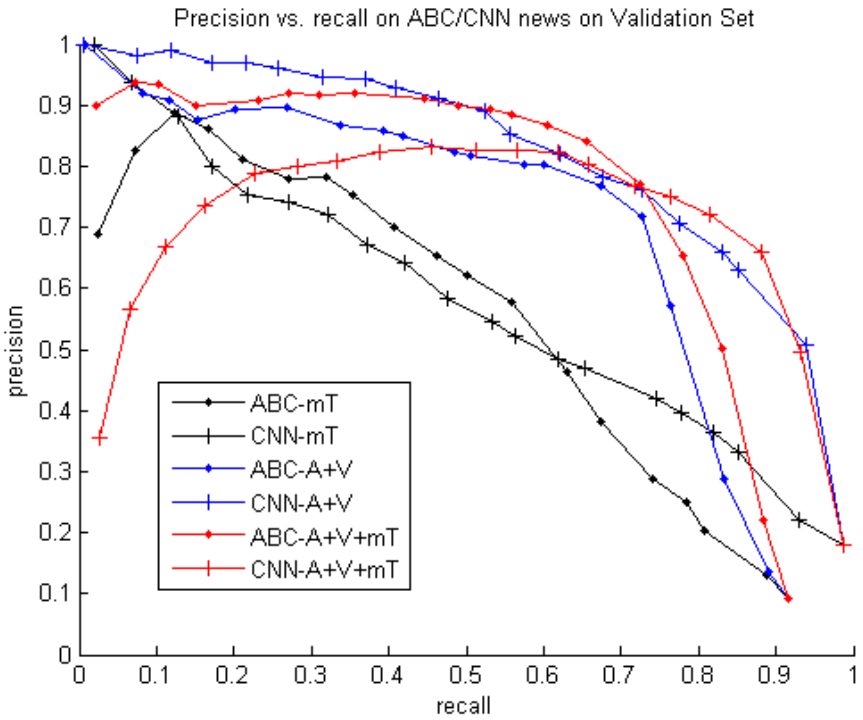


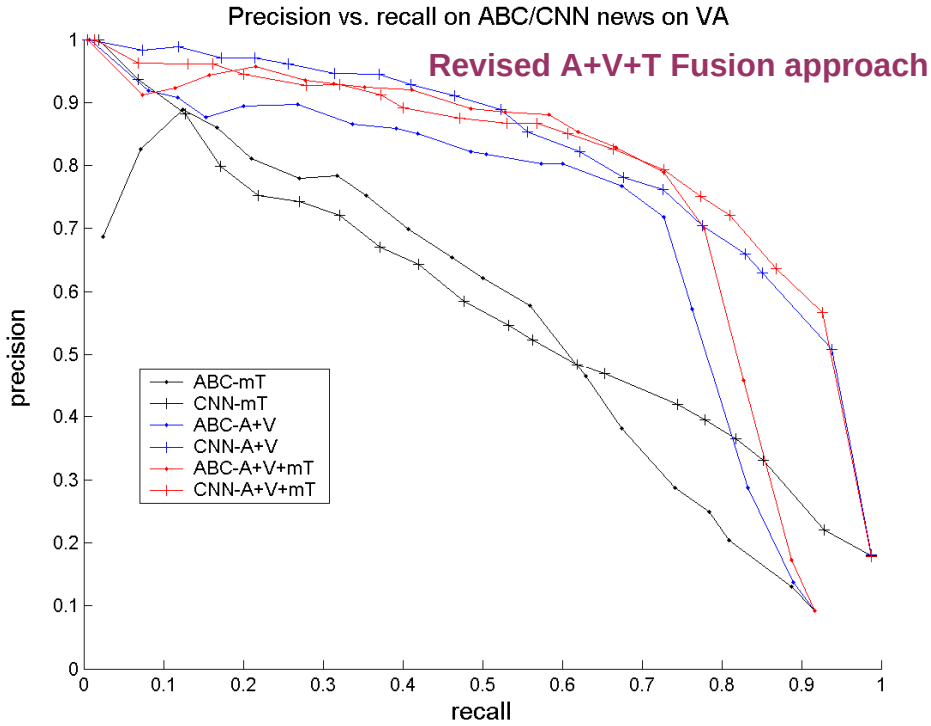
- **Annotate 749 stories into 9 types from 22 CNN videos**
- **Fixed 0.71 precision; VC(*) evaluated at shot boundaries ONLY**

::story types

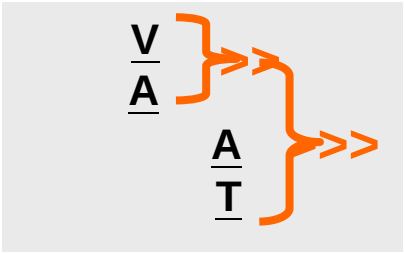
Columbia University in the City of New York

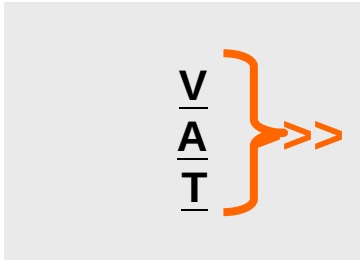
Performance Overview (A+V+T, Validation Set)





Over-fitting in the training set!!

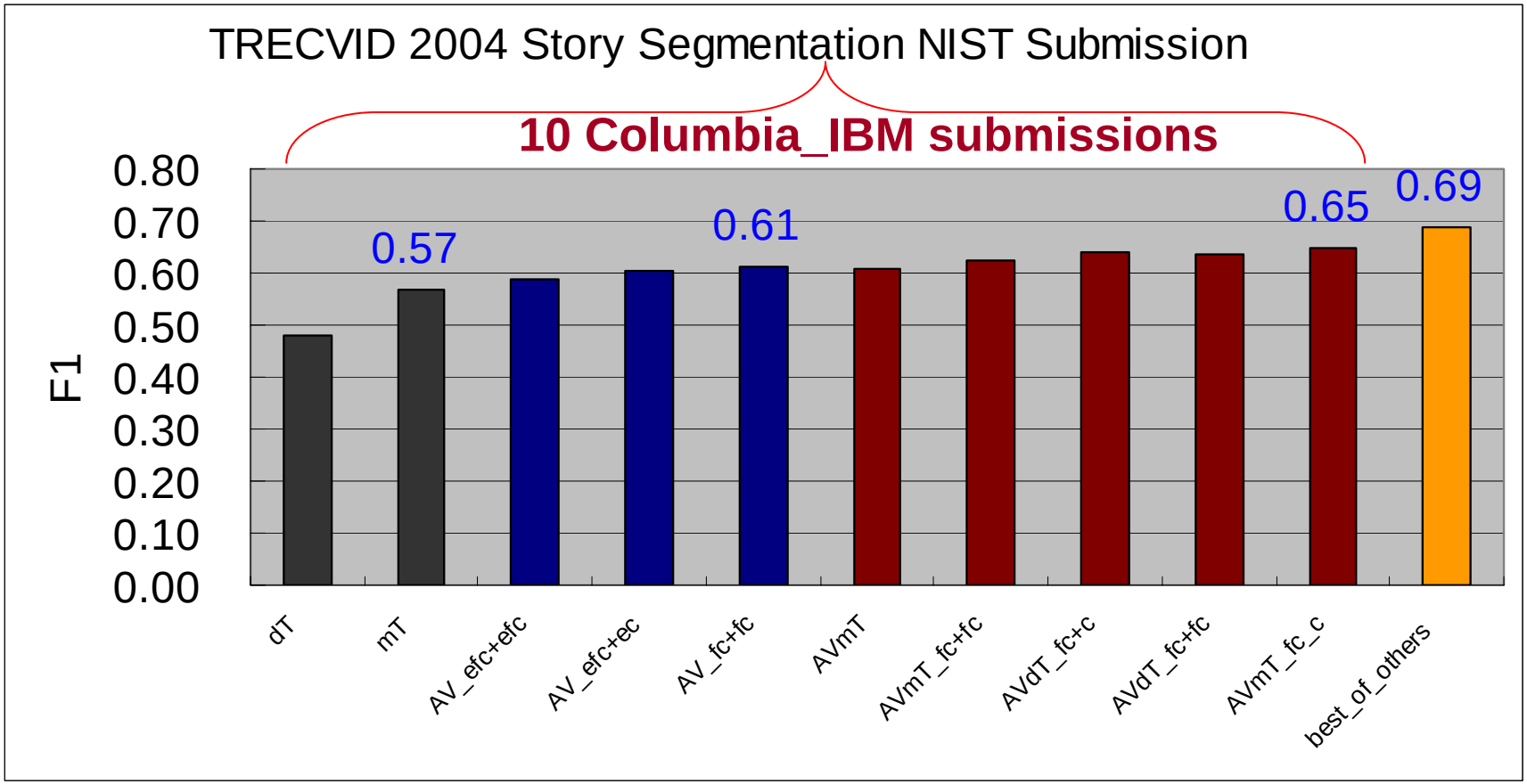




} : SVM fusion

Columbia University in the City of New York

TRECIV04 Test 04 Result



- Significant degradation (10%) comparing with our two validation sets (A+V, A+V+T: 0.72+)
- Probably due to that (1) visual patterns or raw feature had changed a lot in the test set; (2) the fusion strategy; (3) the selection of decision threshold

Summary

- Develop a novel information-theoretical framework to
  - discover visual cue clusters automatically
  - adapt to diverse production events of different channel
  - avoid manual specification/annotation of salient visual cues
- Results confirm the effectiveness of VCs in the validation set
  - But the performance degrades in the test set due to time gap
- Multi-modal fusion
  - Fusion of A and V has significant improvement
  - Fusion of AV and T improves performance in ABC only
  - Strategies for fusion are critical – simultaneous fusion is better
- Major remaining errors
  - Short sports briefings
  - Suggest merging them to a continuous story in the ground truth

< the end; thanks >