informedia
digital video understanding

SEARCH summarize visualize retrieve

Feature Extraction Techniques
CMU at TRECVID 2004

Ming-yu Chen and Jun Yang

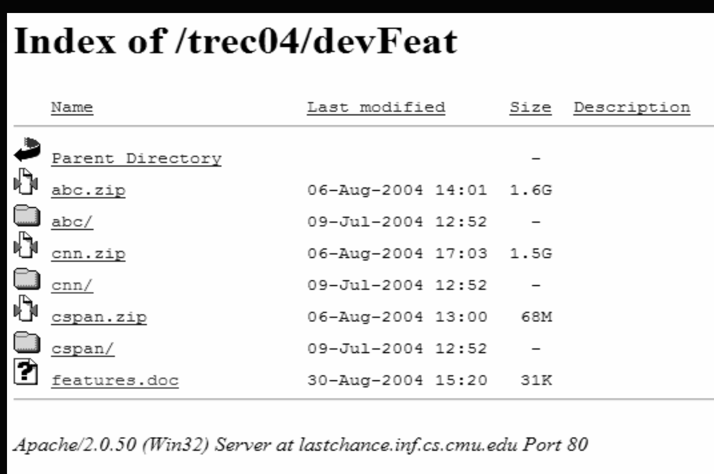
School of Computer Science
Carnegie Mellon University

Outline

- Low level features
- Generic high level feature extractions
  - Uni-modal
  - Multi-modal
  - Multi-concept
- Specialized approach for person finding
- Failure analysis

Low level features overview

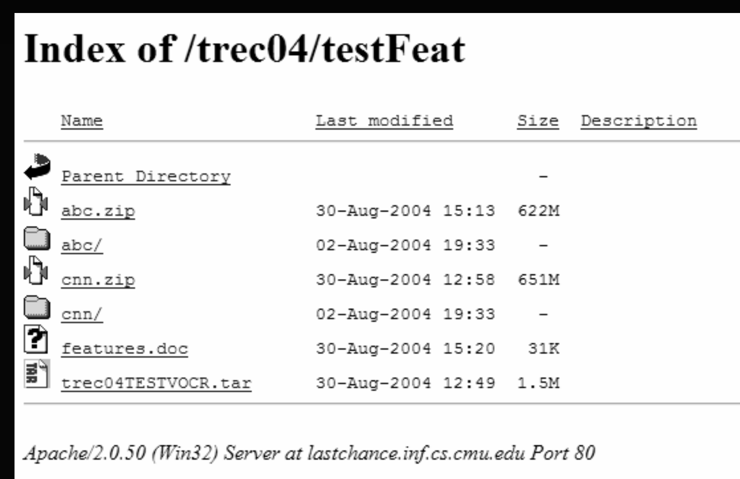
- Low level features
  - CMU distributed 16 feature sets available to all TRECVID participants
  - Development set: http://lastchance.inf.cs.cmu.edu/trec04/devFeat/
  - Test set: http://lastchance.inf.cs.cmu.edu/trec04/testFeat/
  - These features were used for all our submissions
  - We encourage people to compare against these features to eliminate confusion about better features vs better algorithms



Index of /trec04/devFeat

Name	Last modified	Size	Description
Parent Directory		-	
abc.zip	06-Aug-2004 14:01	1.6G	
abc/	09-Jul-2004 12:52	-	
cnn.zip	06-Aug-2004 17:03	1.5G	
cnn/	09-Jul-2004 12:52	-	
cspan.zip	06-Aug-2004 13:00	68M	
cspan/	09-Jul-2004 12:52	-	
features.doc	30-Aug-2004 15:20	31K	

Apache/2.0.50 (Win32) Server at lastchance.inf.cs.cmu.edu Port 80



Index of /trec04/testFeat

Name	Last modified	Size	Description
Parent Directory		-	
abc.zip	30-Aug-2004 15:13	622M	
abc/	02-Aug-2004 19:33	-	
cnn.zip	30-Aug-2004 12:58	651M	
cnn/	02-Aug-2004 19:33	-	
features.doc	30-Aug-2004 15:20	31K	
trec04TESTVOCR.tar	30-Aug-2004 12:49	1.5M	

Apache/2.0.50 (Win32) Server at lastchance.inf.cs.cmu.edu Port 80

Low level features

- Image features
  - Color histogram
  - Texture
  - Edge
- Audio features
  - FFT
  - MFCC
- Motion features
  - Kinetic energy
  - Optical flow
- Detector features
  - Face detection
  - VOCR detection

Image features

- 5 by 5 grids for key-frame per shot
- Color histogram (*.hsv, *.hvc, *.rgb)
  - 5 by 5, 125 bins color histogram
  - HSV, HVC, and RGB color space
  - 3125 dimensions (5*5*125)
  - row-wise grids
  - 19980202_CNN.hsv
    - 0.000000,0.036541,0.009744,0.010962,0.001218,0.000000,0.000000,0.000000,0.000000,0.000000,0.000000,0.000000,0.000000,0.000000,………………….. 0.000000,0.091776,0.055418,0.025415,0.008825,0.000000,0.007413,0.000353,0.000000,0.000000,0.000000,0.000000,0.000000,0.000000, …………………..

- Texture (*.texture_5x5_bhs)
  - Six orientated Gabor filters
- Edge (*.cannyedge)
  - Canny edge detector, 8 orientations

Audio features & Motion features

- Every 20 msecs (512 windows at 44100 HZ sampling rate)
  - FFT (*.FFT) – Short Time Fourier Transform
  - MFCC (*.MFCC) – Mel-Frequency cepstral coefficients
  - SFFT (*.SFFT) – simplified FFT

- Kinetic energy (*.kemotion)
  - Capture the pixel difference between frames
- Mpeg motion (*.mpgmotion)
  - Mpeg motion vector extracted from p-frame
- Optical flow (*.opmotion)
  - Capture optical flow in each grid

Detector features

- Face detector (*.faceinfo)
  - Detecting faces in the images

- VOCR detector (*.vocrinfo and *.mpg.txt)
  - Detecting and recognizing VOCR

Closed caption alignment and Shot mapping

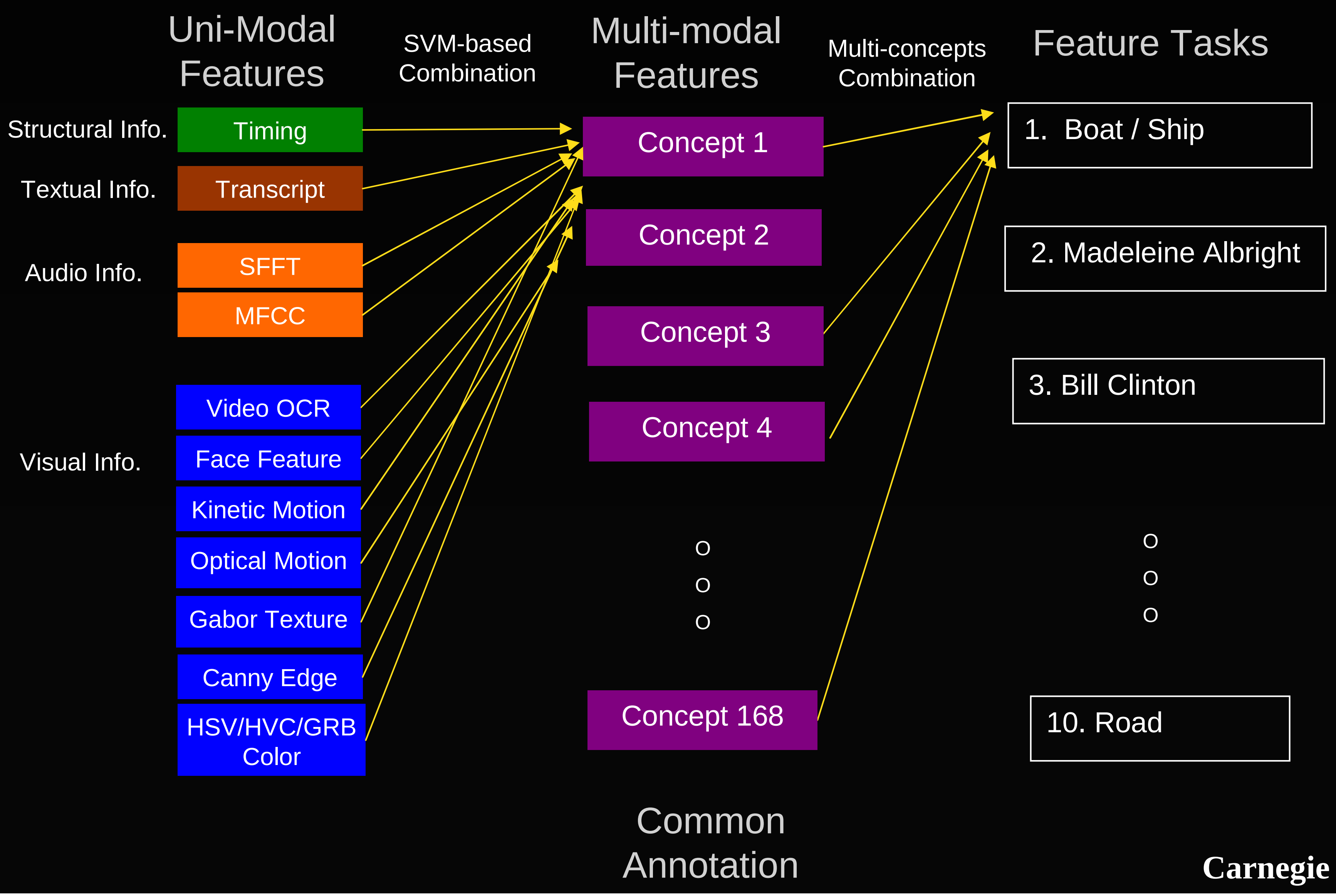
- Closed caption alignment (*.wordtime)
  - Each word in the closed caption file is assigned an approximate time in millisecond
- Shot Break (*.shots)
  - Provides the mapping table of the shot

- We encourage people to utilize these features
  - Eliminate confusion of better features or better algorithms
  - Encourage more participants who can emphasize their efforts on algorithms

Outline

- Low level features
- Generic high level feature extractions
  - Uni-modal
  - Multi-modal
  - Multi-concept
- Specialized approach for person finding
- Failure analysis

Generic high level feature extraction



Multi-concepts

- Learning Bayesian Networks from 168 common annotation concepts
- Pick top 4 most related concepts to combine with the target concept

Boat/Ship	Boat, Water_Body, Sky, Cloud
Train	Car_Crash, Man_Made_scene, Smoke, Road
Beach	Sky, Water_Body, Nature_Non-Vegetation, Cloud
Basket Scored	Crowd, People, Running, Non-Studio_Setting
Airplane Takeoff	Airplane, Sky, Smoke, Space_Vehicle_Launch
People Walking/running	Walking, Running, People, Person
Physical violence	Gun_Shot, Building, Gun, Explosion
Road	Car, Road_Traffic, Truck, Vehicle_Noise

Top result for TRECVID tasks

- Uni-modal gets 2 best over CMU results
- Multi-modal gets 3 best, but includes Boat/Ship which is the best overall all
- Multi-concept gets 6 best

	Boat*	Train	Beach	Basket	Airplane	Walking	Violence	Road
Uni-modal	0.097	0.001	0.013	0.503	0.021	0.015	0.005	0.036
Multi-modal	0.137	0.001	0.023	0.517	0.014	0.008	0.002	0.045
Multi-concept	0.110	0.001	0.039	0.517	0.035	0.099	0.003	0.062

Outline

- Low-level features
- Generic high-level feature extractions
- Specialized approaches for person finding
- Failure analysis

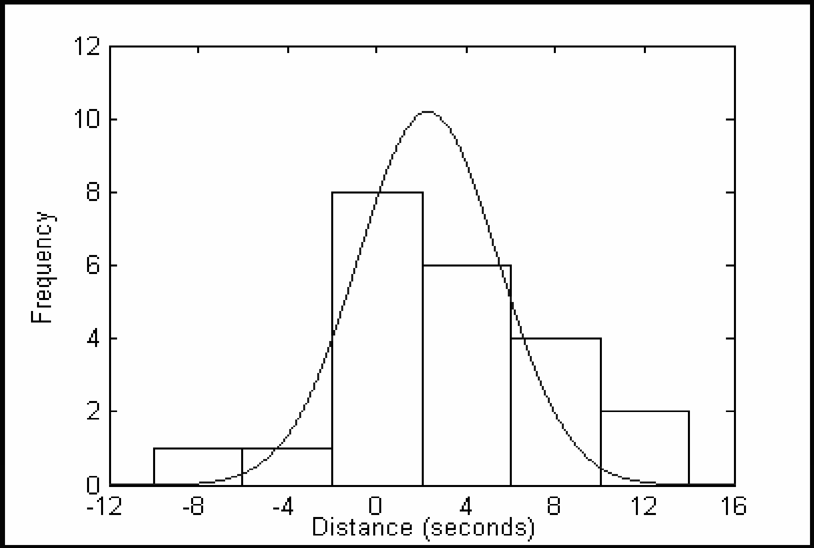
A Text Retrieval Approach

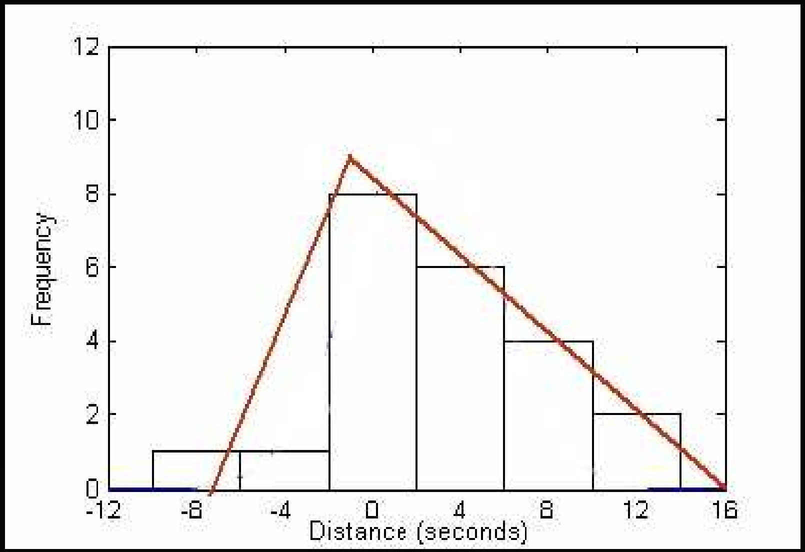
- Search for shots with names appearing in transcript
  - Vector-based IR model with TF*IDF weighting

- Temporal Mismatch Drawback
  - Faces do not always temporally co-occur with names
  - Cause false alarms and misses

Expand the Text Retrieval Results

- Propagate the text score to neighbor shots based on the distribution
- *Timing score = F (Distance (shot, name))*





before the name *name position* *after the name*

- Model 1: Gaussian model (trained using Maximum Likelihood)
- Model 2: Linear model (different gradients set on two sides)

Context Information

- Sometimes, a news story has the name but not the face
  - E.g., "…. a big pressure on Clinton administration ..."
  - Cause many false alarms
  - Related to the context "Clinton administration"
- Given the context, how likely a story has the face?
  - Collect bigrams of type "___ Clinton" or " Clinton ___"
  - Compute P_i *(Clinton appears in the story | bigram$_i$)*

Bigram	**P (face | bigram)**
Clinton says	0.627474
Clinton made	0.625652
…...	…...
Clinton administration	0.242973

Multimodal Features

- Multimodal features provide weak clues for person search
  - Face detection – shots w/o detected faces are less likely to be the results
  - Facial recognition – matching detected faces with facial model based on Eigenface representation
  - Anchor classifier – anchor shots rarely have intended faces
  - Video OCR – Fuzzy match by edit distance between video OCR and the target name

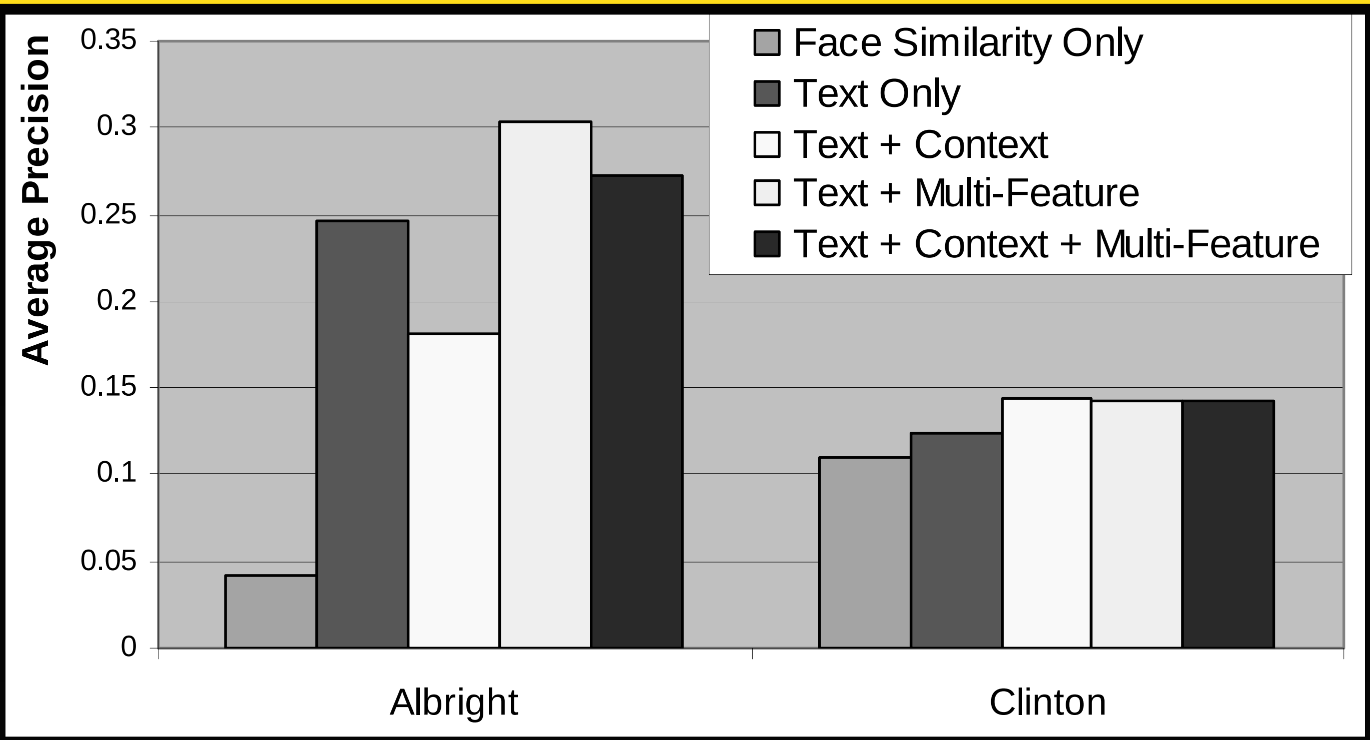
Combining Multimodal Info. with Text Search

- Updated Text Score: $R' = R * \text{Timing Score} * \text{Avg}(P_{bigram\_i})$

- Linear combination of all the features with text score
  - Features normalized into (pseudo-) probabilities [0,1]
  - Feature selection based on chi-square statistics
  - Combinational weights trained by logistic regression

Features	weight
Updated text score	6.14
Face similarity	3.94
Face detection	0.50
Anchor detection	-5.65

Performance Comparison



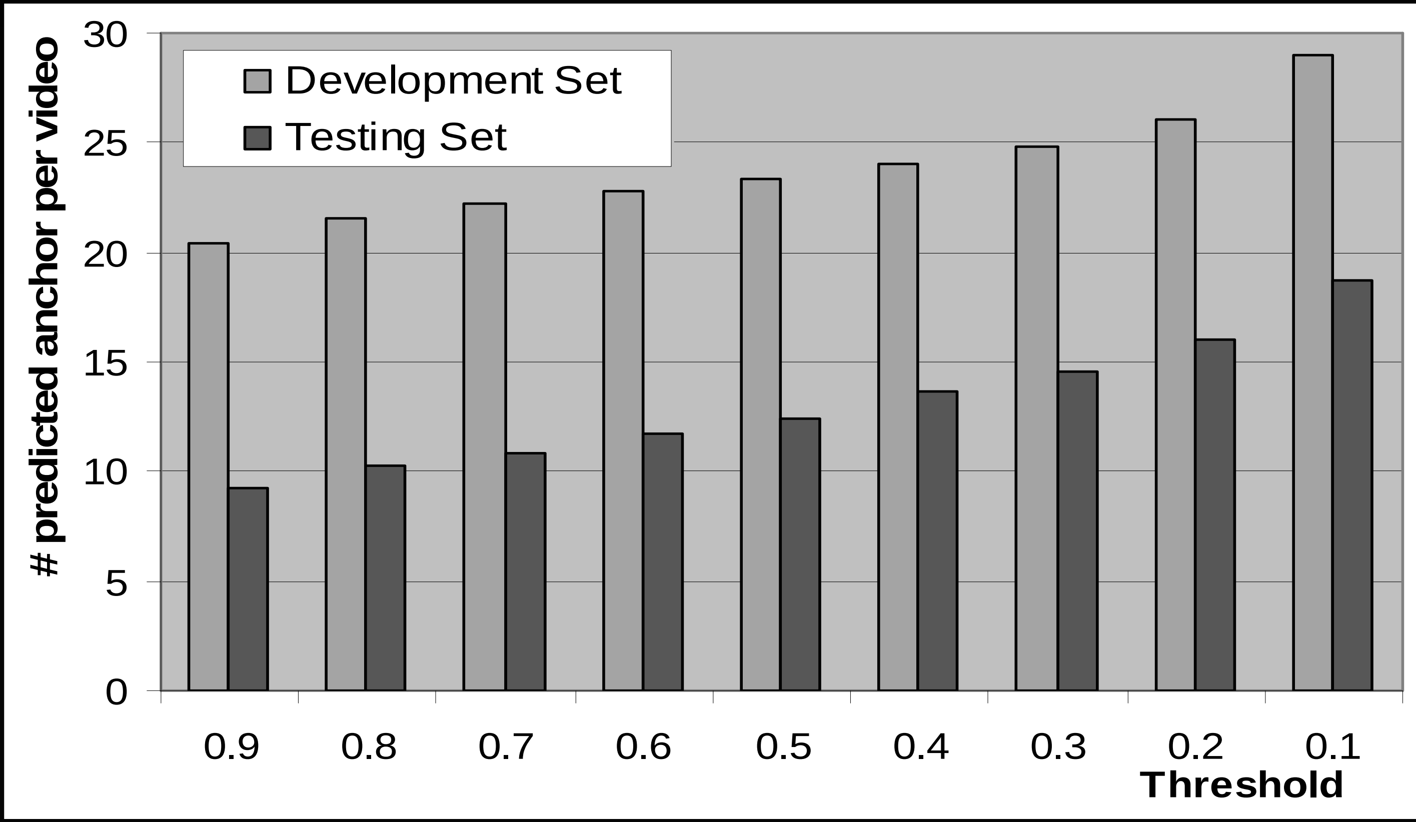
- One of our “Albright” runs is the best among all submissions
- Combining multimodal features helps both tasks
- Context helps “Clinton” but hurts “Albright”
  - Probably due to sparse training data for “Albright”

Outline

- Low-level features
- Generic high-level feature extractions
- Specialized approaches for person finding
- Failure analysis

Performance Drop on Anchor Classifier

- 95% cross-validation accuracy on development set



- 10 videos in testing set has 0 or 1 detected anchor
  - Average # of anchor shots per video is 20-30

Different Data Distribution

- Different images: change on background, anchor, clothes

Common types (development set)

Outliers (testing set)

- Similar images, but probably different MPEG encoding
  - “Peter Jennings” has similar clothes and background in both sets
  - In videos with “Peter Jennings” as the anchor
    - 19 detected per video in development set
    - 13 detected per video in testing set

Other Semantic Features

- Similar performance drop observed on Commercial, Sport News, etc
  - Compromises both high-level feature extraction and search

- Possible solutions
  - Get consistent data next year
  - Rely less on sensitive image features (color, etc)
  - Rely more on robust features -- “Video grammar”
    - Timing, e.g., sports news are in the same temporal session
    - Story structure, e.g., the first shot of a story is usually an anchor shot
    - Speech, e.g., anchor’s speaker ID is easy to identify
    - Constraints, e.g, weather forecast appear only once
  - Re-encoding MPEG video

Q & A

Thank you!