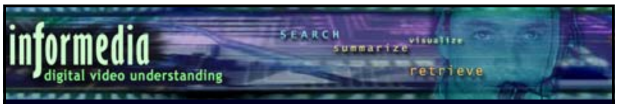


Machine Learning, Pattern Recognition, Cross-modal Analysis and Fusion

Alex Hauptmann
School of Computer Science
Carnegie Mellon University,
Pittsburgh, PA, USA
alex@cs.cmu.edu
http://www.cs.cmu.edu/~alex

Outline

Part One

- Introduction/Background
  - Motivation
  - Text, Speech, Images, Video
- Case Studies in Cross-Modal Analysis
  - Labeling Monologues, Faces and Locations

Part Two

- Bridging the Semantic Gap across Modalities
  - A Large Scale Ontology for Multimedia (LSCOM)
  - Retrieval Experiments with LSCOM
  - Active Learning of Semantic Concepts
- YouTube and Challenges for the Future

Extracting information from video sources

- Understanding how language refers to video imagery
  - Learn the kinds of visual objects that may be strongly predicted from particular expressions
  - Utilize recorded “audio descriptions” of a media’s visual content
- Identify imagery and audio components
  - Audio classifiers: transcribed speech, gender, gunfire, cheering/jeering, …
  - Image classifiers: in/outdoor, people, crowds, interviews, sports, …
  - Event classifiers: combat, rally, meeting, …
- Applying broadcast TV news ontology
  - Event ontology: functional (e.g.,role of actions) and structural (e.g. organization, sequence)
  - *Primitive* and *composite* events: descriptive name, time interval, objects participating in it

Application of Diverse, Imperfect Technologies

- Speech understanding for automatically derived transcripts
- Image understanding for video “paragraphing”; face, text and object recognition
- Natural language for segmentation, query understanding and content summarization
- Machine learning for classification and modeling
- Human computer interaction for video display, navigation and reuse

⋮

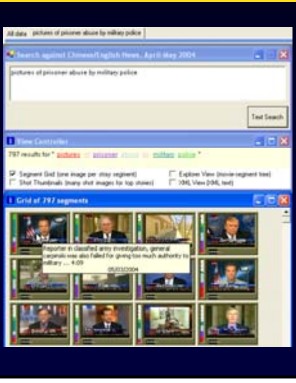
- ***Integration overcomes limitation of each***

Understanding multimedia questions

“Find scenes with George Bush exiting a car like this in New York”

- Assemble context information of query into a single structured representation, independent of modality
  - Locations may be mentioned, scenically pictured, noted on a map
  - Account for what the user already knows, observes and annotates
- Extend broadcast news video ontology for representing component relationships in multimedia queries
  - Augment with an understanding of simple relations (e.g., “like this”)

English Text Query on Video Corpus


All data | pictures of prisoner abuse by military police

Search against Chinese/English News, April-May 2004

pictures of prisoner abuse by military police

Text Search

View Controller

797 results for " pictures of prisoner abuse by military police "

☑ Segment Grid (one image per story segment)
□ Shot Thumbnails (many shot images for top stories)
□ Explorer View (movie-segment tree)
□ XML View (XML text)

Grid of 797 segments

Reporter in classified army investigation, general corporals was also failed for giving too much authority to military ... 4:09
05/03/2004

Image Search Across Multilingual Sources

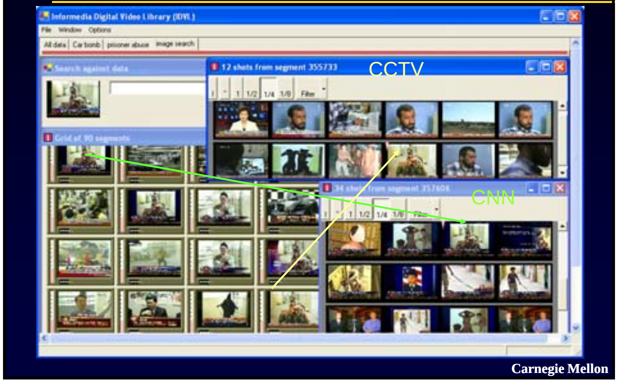
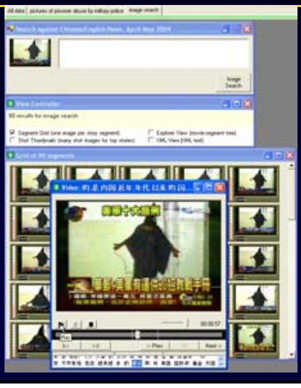

Informedia Digital Video Library (IDVL)
File Window Options
All data | Car bomb | prisoner abuse | image search
Search against data
Grid of 90 segments
12 shots from segment 355733
CCTV
34 shots from segment 357606
CNN

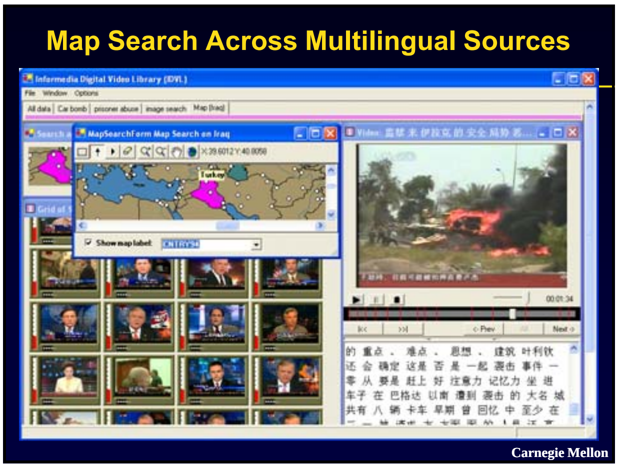
Image Query from Key-frame Image



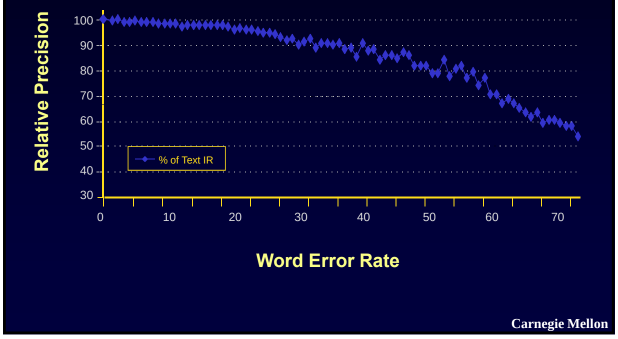
- Image similarity based
- Uses shot to launch query
- Results cut across different languages
- Find top-ranked (Chinese) news source segment

ARDA

Map Search Across Multilingual Sources


Informedia Digital Video Library (IDVL)
File Window Options
All data | Car bomb | prisoner abuse | image search | Map (Iraq)
MapSearchForm Map Search on Iraq
Turkey
Show map label
Grid of
&lt;&lt; Prev
Next &gt;&gt;

Information Retrieval Precision vs. Speech Accuracy



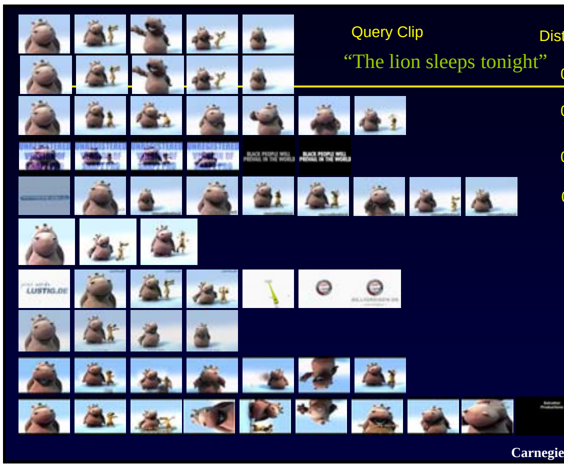
Beyond Spoken Text

Broadcast speech transcripts do not describe the video

- Rule for reporters: Let the images tell their own story

- Image Retrieval
  - Need sample images

  - Find duplicates
  - Similar colors
  - Similar layout or shape
  - Similar content



Query Clip

"The lion sleeps tonight"

Distance

0

0.01

0.91

0.33

0.64

0.76

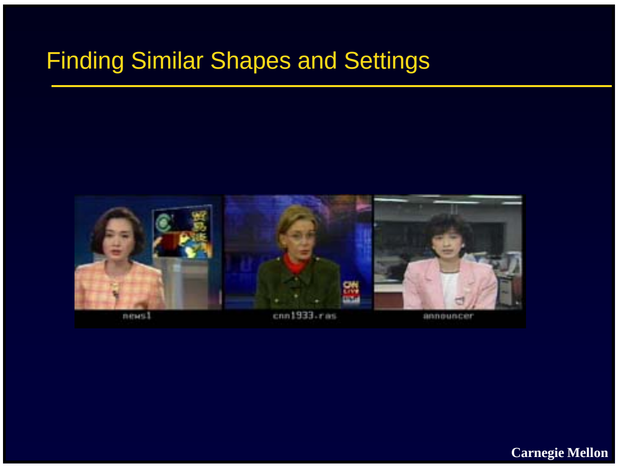
0.67

0.46

0.51

Image Similarity Challenge: Color

Carnegie Mellon

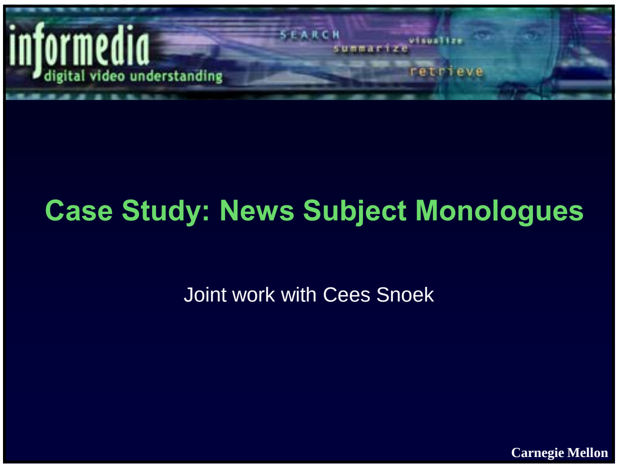


Finding Similar Shapes and Settings

Carnegie Mellon

Images containing similar content

Carnegie Mellon

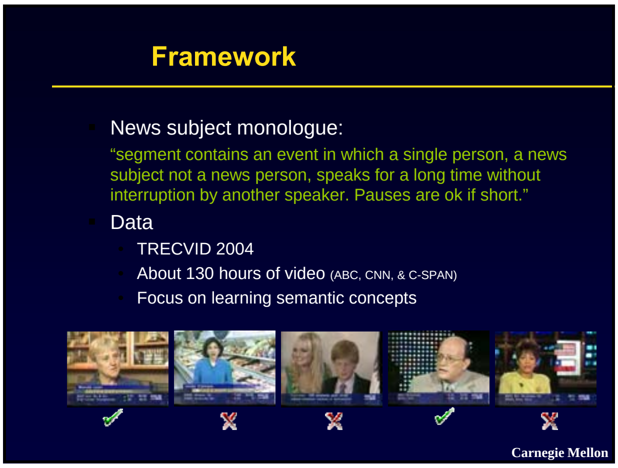


informedia
digital video understanding
SEARCH
summarize
visualize
retrieve

Case Study: News Subject Monologues

Joint work with Cees Snoek

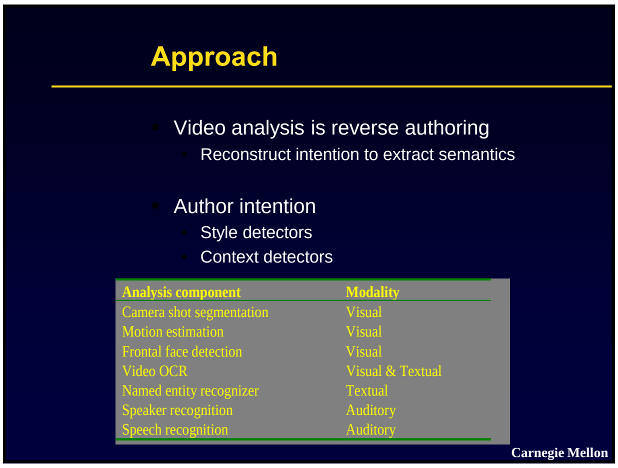
Carnegie Mellon



Framework

- News subject monologue:
  "segment contains an event in which a single person, a news subject not a news person, speaks for a long time without interruption by another speaker. Pauses are ok if short."
- Data
  - TRECVID 2004
  - About 130 hours of video (ABC, CNN, & C-SPAN)
  - Focus on learning semantic concepts

Carnegie Mellon

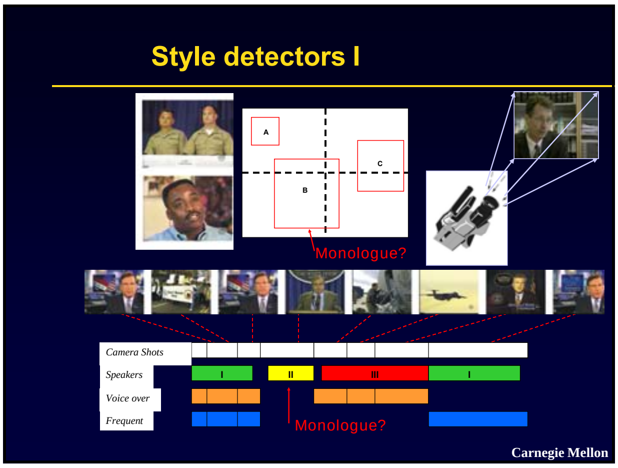


Approach

- Video analysis is reverse authoring
  - Reconstruct intention to extract semantics
- Author intention
  - Style detectors
  - Context detectors

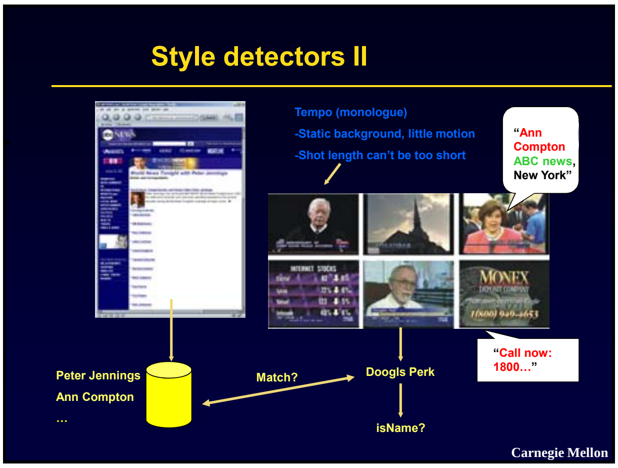
Analysis component	Modality
Camera shot segmentation	Visual
Motion estimation	Visual
Frontal face detection	Visual
Video OCR	Visual & Textual
Named entity recognizer	Textual
Speaker recognition	Auditory
Speech recognition	Auditory

Carnegie Mellon



Style detectors I

Carnegie Mellon



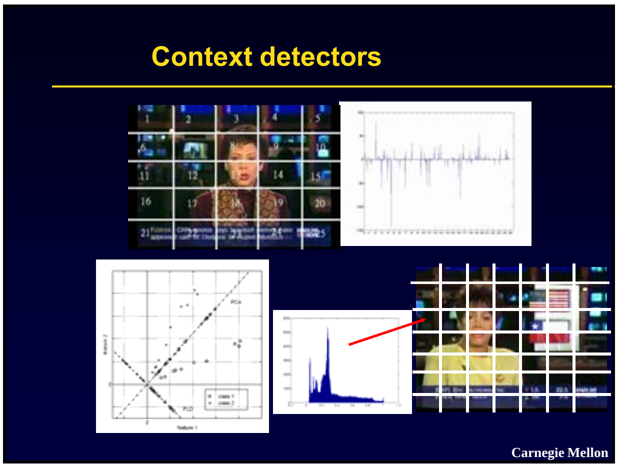
Style detectors II

Tempo (monologue)

-Static background, little motion

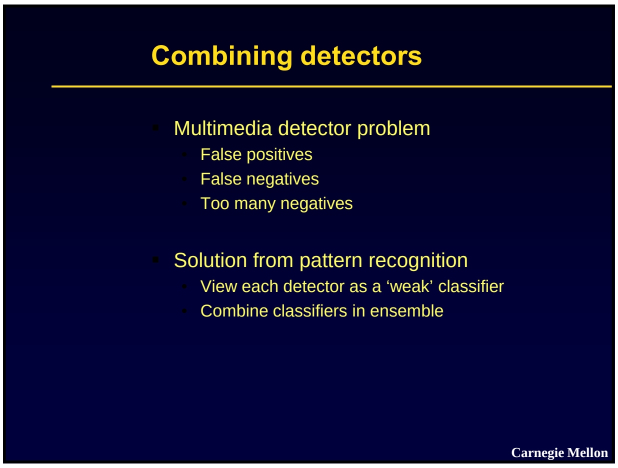
-Shot length can't be too short

Carnegie Mellon



Context detectors

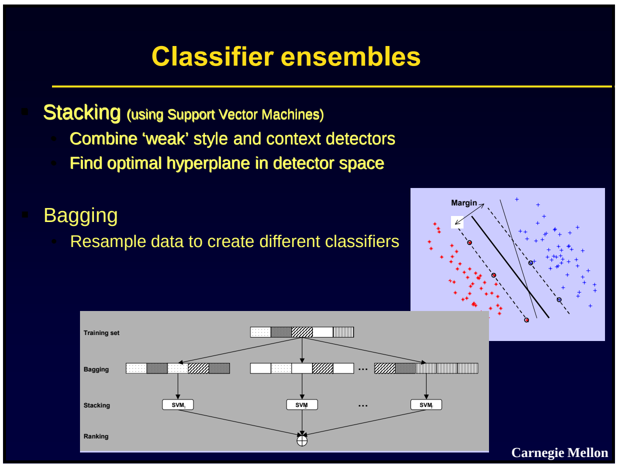
Carnegie Mellon



Combining detectors

- Multimedia detector problem
  - False positives
  - False negatives
  - Too many negatives
- Solution from pattern recognition
  - View each detector as a 'weak' classifier
  - Combine classifiers in ensemble

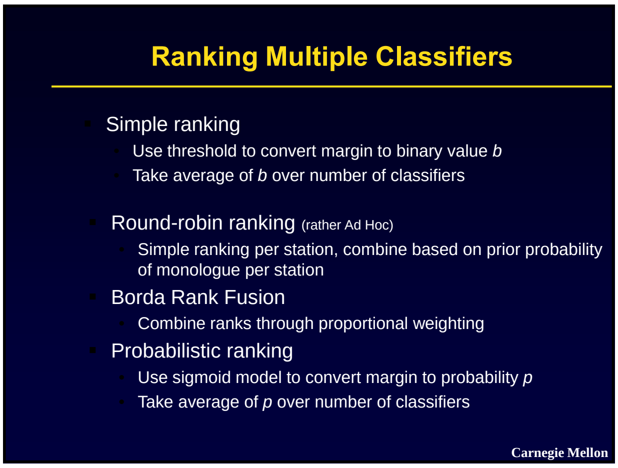
Carnegie Mellon



Classifier ensembles

- Stacking (using Support Vector Machines)
  - Combine 'weak' style and context detectors
  - Find optimal hyperplane in detector space
- Bagging
  - Resample data to create different classifiers

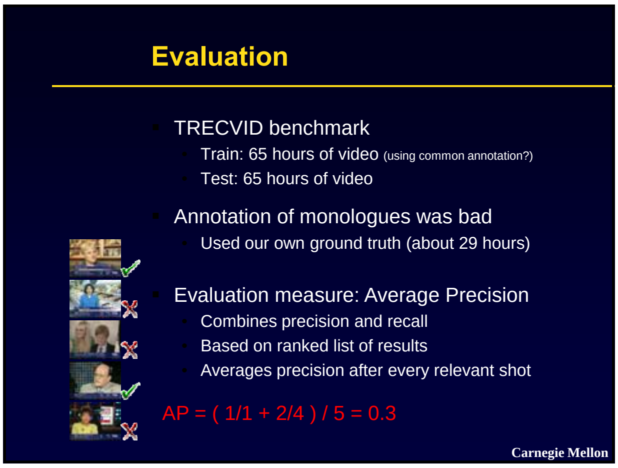
Carnegie Mellon



Ranking Multiple Classifiers

- Simple ranking
  - Use threshold to convert margin to binary value *b*
  - Take average of *b* over number of classifiers
- Round-robin ranking (rather Ad Hoc)
  - Simple ranking per station, combine based on prior probability of monologue per station
- Borda Rank Fusion
  - Combine ranks through proportional weighting
- Probabilistic ranking
  - Use sigmoid model to convert margin to probability *p*
  - Take average of *p* over number of classifiers

Carnegie Mellon

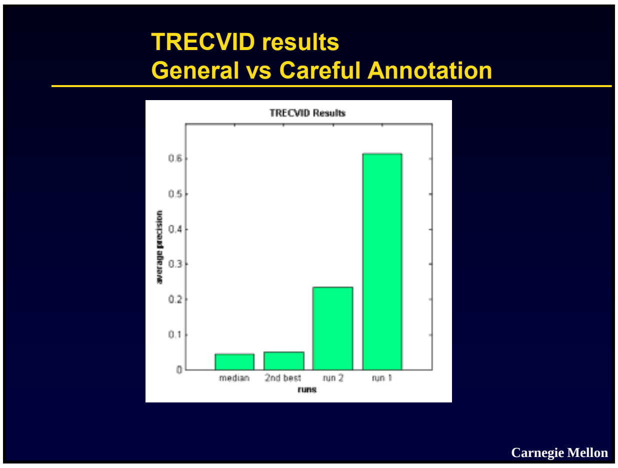


Evaluation

- TRECVID benchmark
  - Train: 65 hours of video (using common annotation?)
  - Test: 65 hours of video
- Annotation of monologues was bad
  - Used our own ground truth (about 29 hours)
- Evaluation measure: Average Precision
  - Combines precision and recall
  - Based on ranked list of results
  - Averages precision after every relevant shot

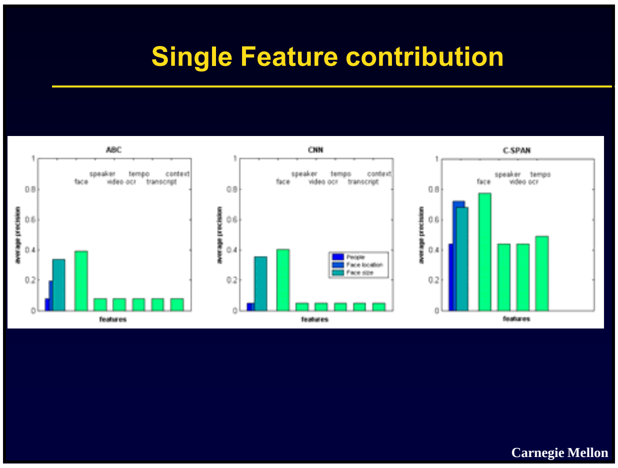
AP = (1/1 + 2/4) / 5 = 0.3

Carnegie Mellon



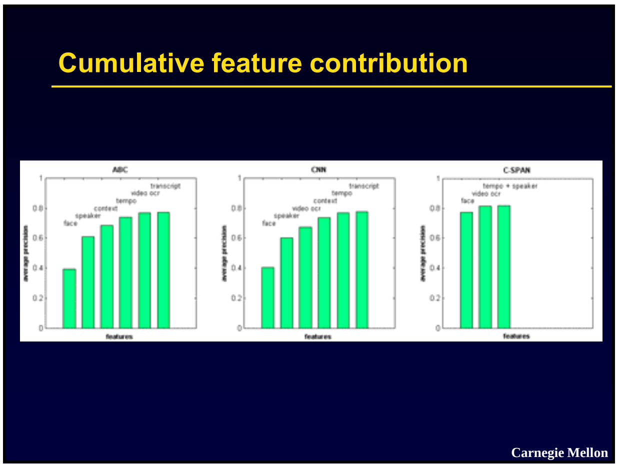
TRECVID results General vs Careful Annotation

Carnegie Mellon



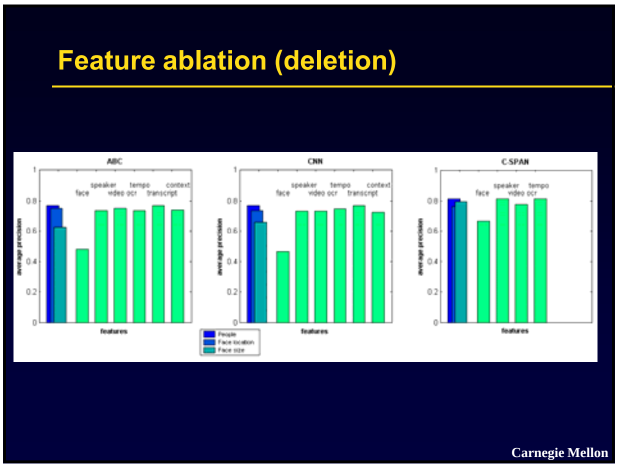
Single Feature contribution

Carnegie Mellon



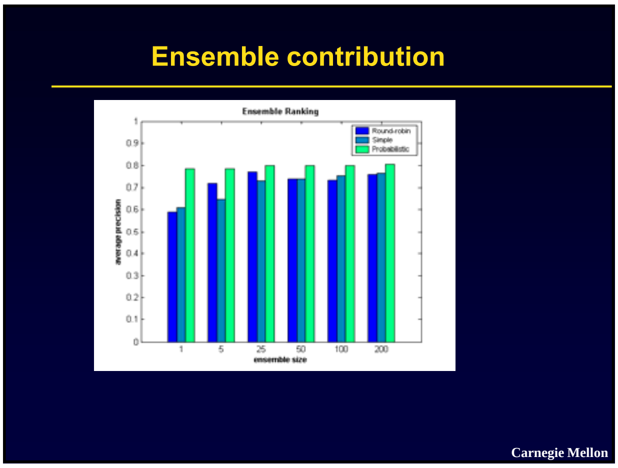
Cumulative feature contribution

Carnegie Mellon



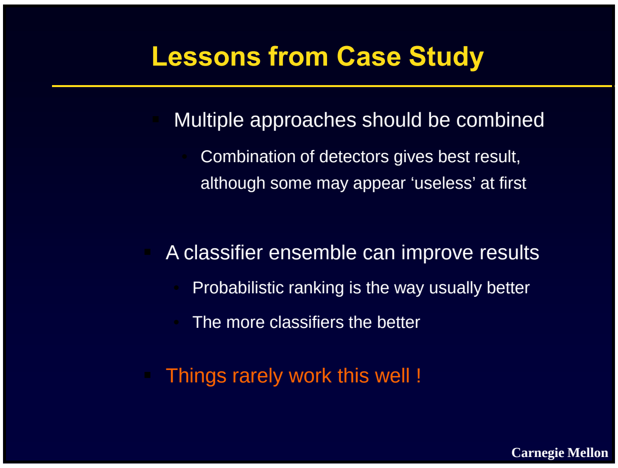
Feature ablation (deletion)

Carnegie Mellon



Ensemble contribution

Carnegie Mellon



Lessons from Case Study

- Multiple approaches should be combined
  - Combination of detectors gives best result, although some may appear ‘useless’ at first
- A classifier ensemble can improve results
  - Probabilistic ranking is the way usually better
  - The more classifiers the better
- Things rarely work this well !

Carnegie Mellon

informedia
digital video understanding
SEARCH
summarize
visualize
retrieve

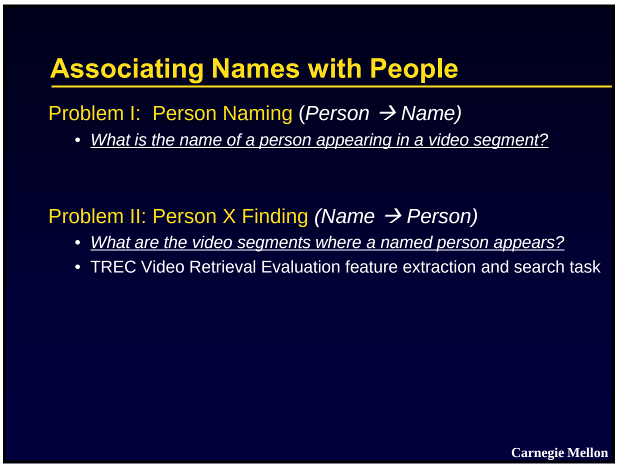
Case Study: Labeling Persons and Locations in Broadcast News

Carnegie Mellon

Naming People in the News

- News video is mainly about the activities of people
- There are a large number of people in the video
  - Approx. 50 appearances of individual person in 30-min broadcasting
  - Approx. 15-20 distinct named persons

Carnegie Mellon



Associating Names with People

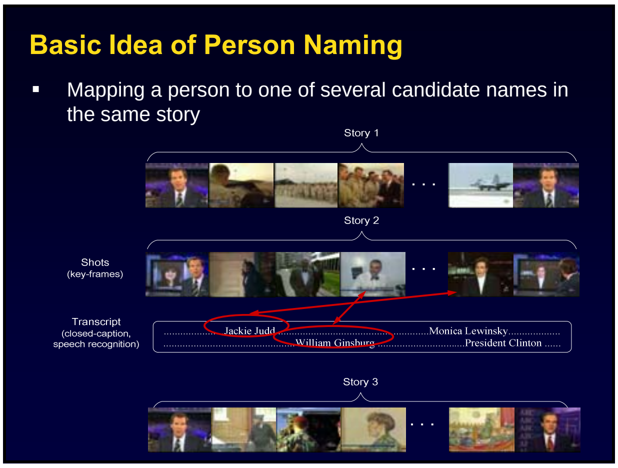
Problem I: Person Naming (*Person → Name)*

- *What is the name of a person appearing in a video segment?*

Problem II: Person X Finding *(Name → Person)*

- *What are the video segments where a named person appears?*
- TREC Video Retrieval Evaluation feature extraction and search task

Carnegie Mellon



Basic Idea of Person Naming

- Mapping a person to one of several candidate names in the same story

Simplifications

1. Monologue shots only
   - Monologue – individual speaker (anchor, reporter, news subject)
   - No shots with multiple talking faces
2. Persons with names in transcript only
   - A person’s name may appear in the transcript or in the text overlaid on screen, or never appears (anonymous)
   - No anonymous persons
   - No persons with names only in overlaid text

Carnegie Mellon

Problem Formulation

- Person naming is a classification problem
  - Choosing a person’s name from a set of candidate names
  - Learning is necessary to combing a variety of features
- Formulation A

  mapping *F: {Shot} → {Name}*

  - *{Name}* is basically infinite
    - There are always strangers emerging in news video
    - *F* is not learnable with infinite labels
  - Not every name in *{Name}* are valid candidates for a shot

Problem Formulation (cont.)

- Formulation B

  G: { <Shot, Name> } → R [-1, 1]

  where R is the degree of association between a name-shot pair

  - For a shot, the name with the largest *R* is predicted
  - A regression problem
  - Overcome all the modeling problems
    - Names can be infinite
    - Only valid candidate pairs are used
    - Features of names can be represented
- Trained using Support Vector Machine (SVM)

Probabilistic formulation

- Estimate the probability that a face is associated with a name, where the association is described by a feature set:

$$Estimate \quad P(Y=1 \mid F, N) = P(Y=1 \mid X)$$

- A typical binary classification problem on distinguishing correct and incorrect face-name associations
- Label a face with the name with the highest probability:

$$n(f_i) = \arg\max_j P(Y=1 \mid f_i, n_{ij})$$

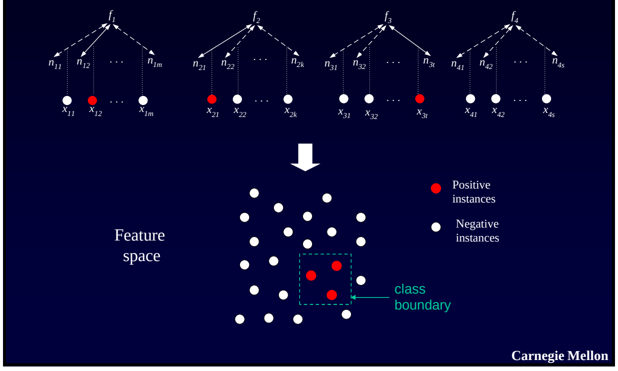
$\{n_{ij}\}$ are candidate names extracted from the same story as f_i

F: face, *N*: name, *Y*: label on face-name association, *X*: feature set

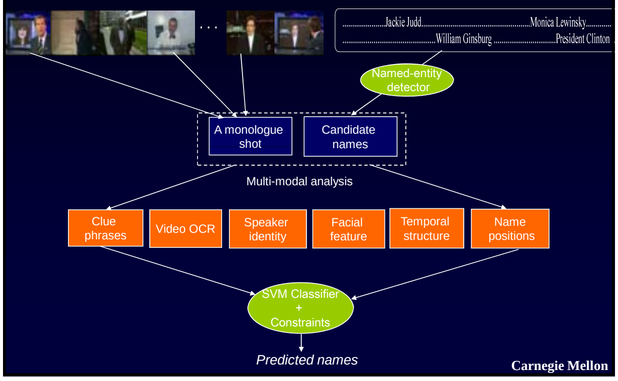
Supervised approaches

- Use supervised learning methods to build a classifier from labeled training data
- Labeled data: a set of example faces labeled with correct and incorrect names
- Learning methods: SVM, logistic regression, etc
- Manual labeling needed for good performance
  - Large volume of video data
  - Heterogeneous news programs

Supervised face labeling


f_1
n_{11} n_{12} . . . n_{1m}
x_{11} x_{12} . . . x_{1m}
f_2
n_{21} n_{22} . . . n_{2k}
x_{21} x_{22} . . . x_{2k}
f_3
n_{31} n_{32} . . . n_{3t}
x_{31} x_{32} . . . x_{3t}
f_4
n_{41} n_{42} . . . n_{4s}
x_{41} x_{42} . . . x_{4s}
Feature space
Positive instances
Negative instances
class boundary

Framework



Named-entity detector

A monologue shot

Candidate names

Multi-modal analysis

Clue phrases

Video OCR

Speaker identity

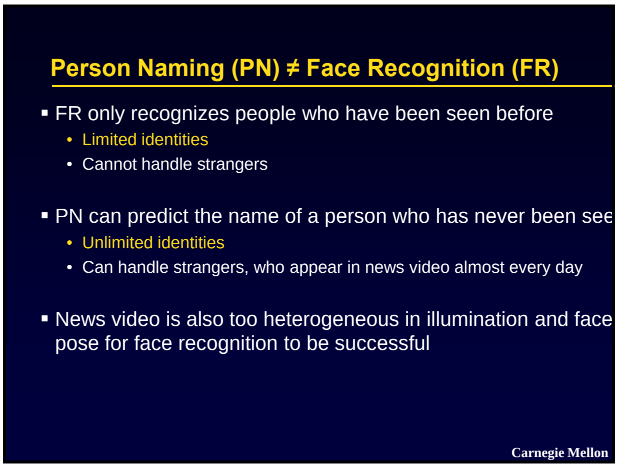
Facial feature

Temporal structure

Name positions

SVM Classifier + Constraints

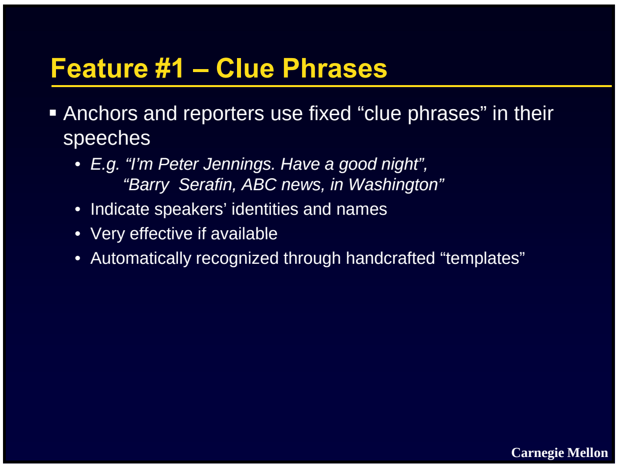
Predicted names



Person Naming (PN) ≠ Face Recognition (FR)

- FR only recognizes people who have been seen before
  - Limited identities
  - Cannot handle strangers
- PN can predict the name of a person who has never been see
  - Unlimited identities
  - Can handle strangers, who appear in news video almost every day
- News video is also too heterogeneous in illumination and face pose for face recognition to be successful

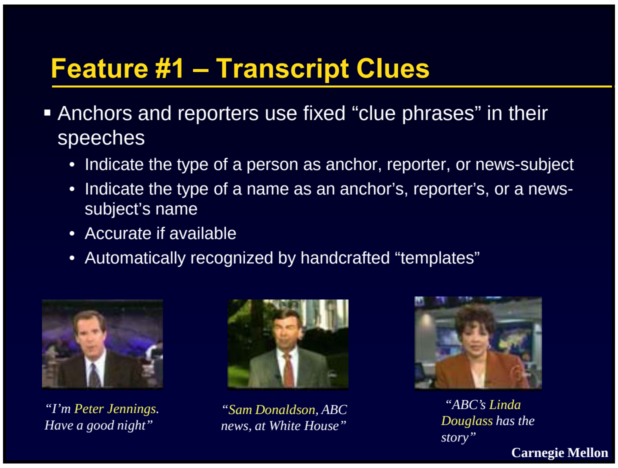
Carnegie Mellon



Feature #1 – Clue Phrases

- Anchors and reporters use fixed “clue phrases” in their speeches
  - *E.g. “I’m Peter Jennings. Have a good night”,*
    “Barry Serafin, ABC news, in Washington”
  - Indicate speakers’ identities and names
  - Very effective if available
  - Automatically recognized through handcrafted “templates”

Carnegie Mellon



Feature #1 – Transcript Clues

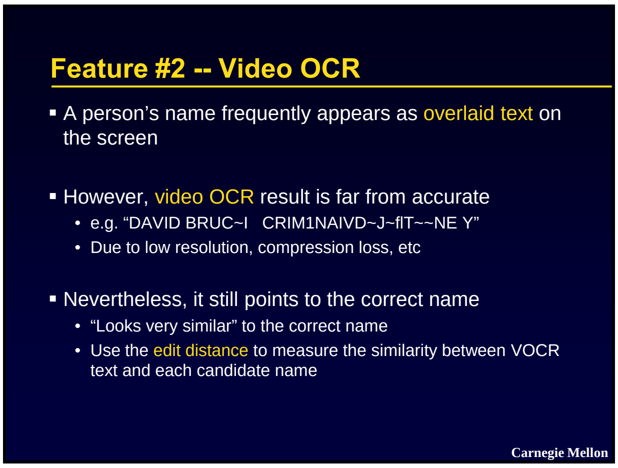
- Anchors and reporters use fixed “clue phrases” in their speeches
  - Indicate the type of a person as anchor, reporter, or news-subject
  - Indicate the type of a name as an anchor’s, reporter’s, or a news-subject’s name
  - Accurate if available
  - Automatically recognized by handcrafted “templates”

“I’m Peter Jennings. Have a good night”

“Sam Donaldson, ABC news, at White House”

“ABC’s Linda Douglass has the story”

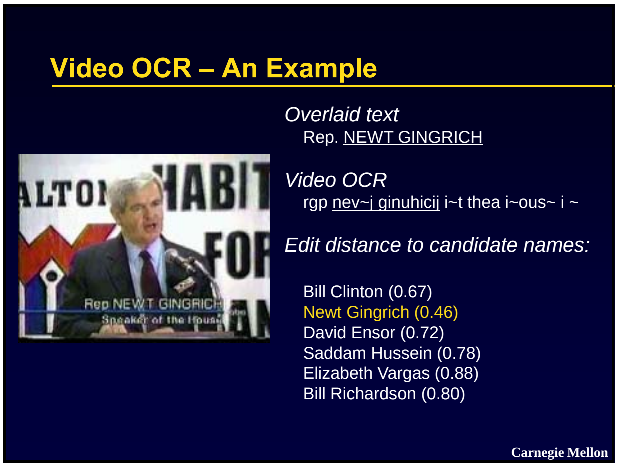
Carnegie Mellon



Feature #2 -- Video OCR

- A person’s name frequently appears as overlaid text on the screen
- However, video OCR result is far from accurate
  - e.g. “DAVID BRUC~I   CRIM1NAIVD~J~flT~~NE Y”
  - Due to low resolution, compression loss, etc
- Nevertheless, it still points to the correct name
  - “Looks very similar” to the correct name
  - Use the edit distance to measure the similarity between VOCR text and each candidate name

Carnegie Mellon



Video OCR – An Example

Overlaid text

Rep. NEWT GINGRICH

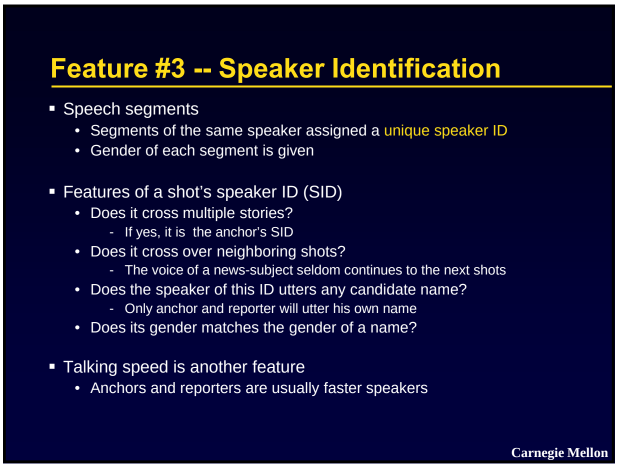
Video OCR

rgp nev~j ginuhicij i~t thea i~ous~ i ~

Edit distance to candidate names:

Bill Clinton (0.67)
Newt Gingrich (0.46)
David Ensor (0.72)
Saddam Hussein (0.78)
Elizabeth Vargas (0.88)
Bill Richardson (0.80)

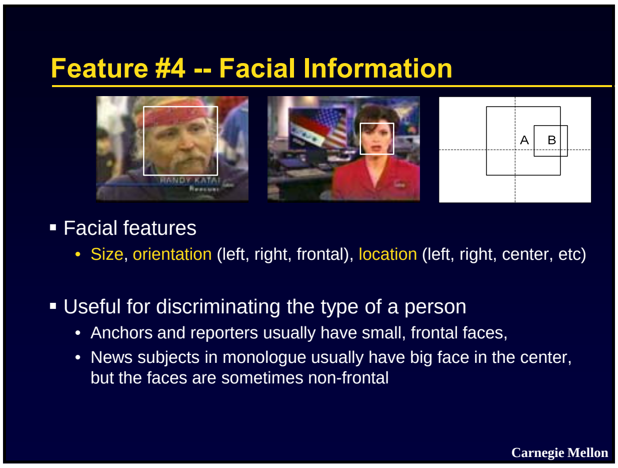
Carnegie Mellon



Feature #3 -- Speaker Identification

- Speech segments
  - Segments of the same speaker assigned a unique speaker ID
  - Gender of each segment is given
- Features of a shot’s speaker ID (SID)
  - Does it cross multiple stories?
    - If yes, it is the anchor’s SID
  - Does it cross over neighboring shots?
    - The voice of a news-subject seldom continues to the next shots
  - Does the speaker of this ID utters any candidate name?
    - Only anchor and reporter will utter his own name
  - Does its gender matches the gender of a name?
- Talking speed is another feature
  - Anchors and reporters are usually faster speakers

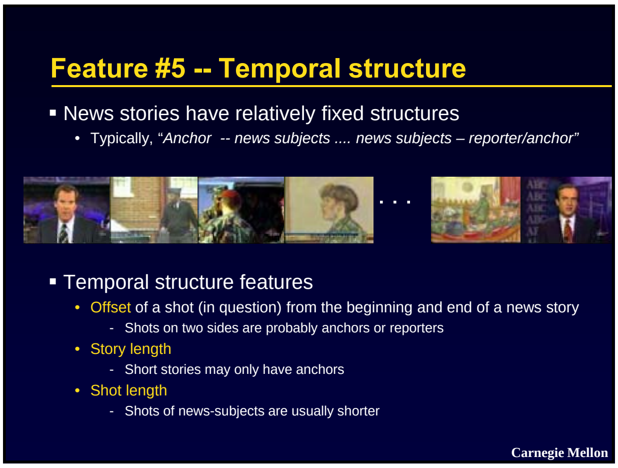
Carnegie Mellon



Feature #4 -- Facial Information

RANDY KATAI
A
B

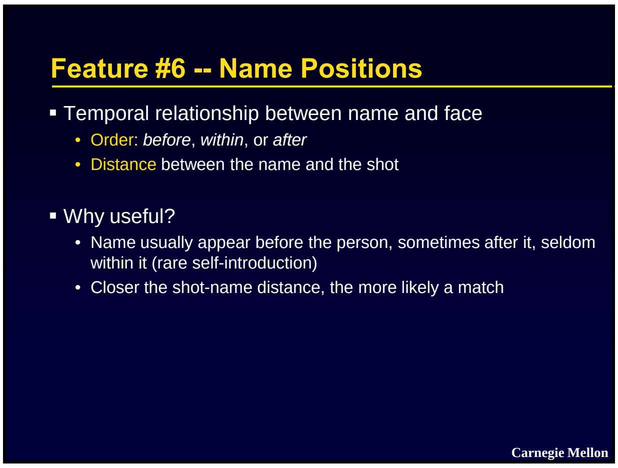
- Facial features
  - Size, orientation (left, right, frontal), location (left, right, center, etc)
- Useful for discriminating the type of a person
  - Anchors and reporters usually have small, frontal faces,
  - News subjects in monologue usually have big face in the center, but the faces are sometimes non-frontal



Feature #5 -- Temporal structure

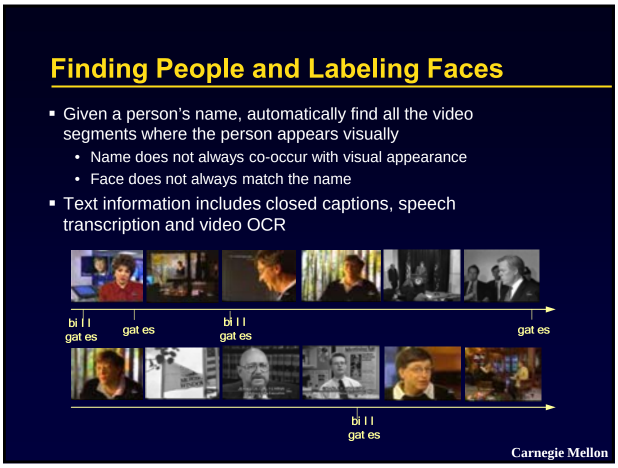
- News stories have relatively fixed structures
  - Typically, "*Anchor -- news subjects .... news subjects – reporter/anchor"*

- Temporal structure features
  - Offset of a shot (in question) from the beginning and end of a news story
    - Shots on two sides are probably anchors or reporters
  - Story length
    - Short stories may only have anchors
  - Shot length
    - Shots of news-subjects are usually shorter



Feature #6 -- Name Positions

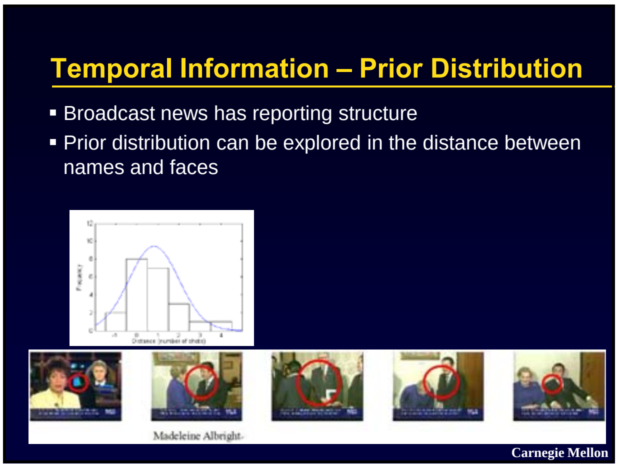
- Temporal relationship between name and face
  - Order: *before*, *within*, or *after*
  - Distance between the name and the shot
- Why useful?
  - Name usually appear before the person, sometimes after it, seldom within it (rare self-introduction)
  - Closer the shot-name distance, the more likely a match



Finding People and Labeling Faces

- Given a person's name, automatically find all the video segments where the person appears visually
  - Name does not always co-occur with visual appearance
  - Face does not always match the name
- Text information includes closed captions, speech transcription and video OCR

bi l l
gat es
gat es
bi l l
gat es
gat es
bi l l
gat es

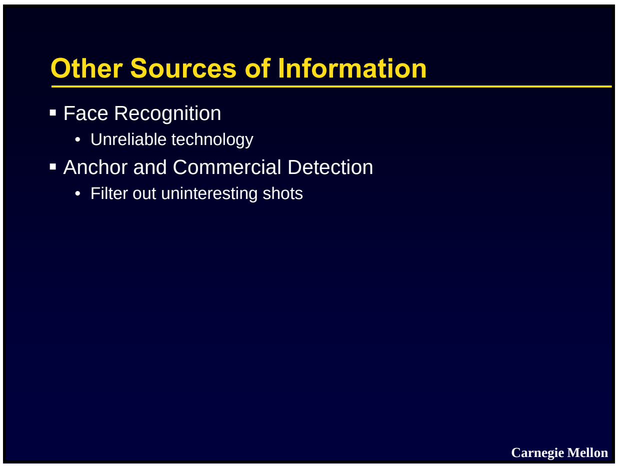


Temporal Information – Prior Distribution

- Broadcast news has reporting structure
- Prior distribution can be explored in the distance between names and faces

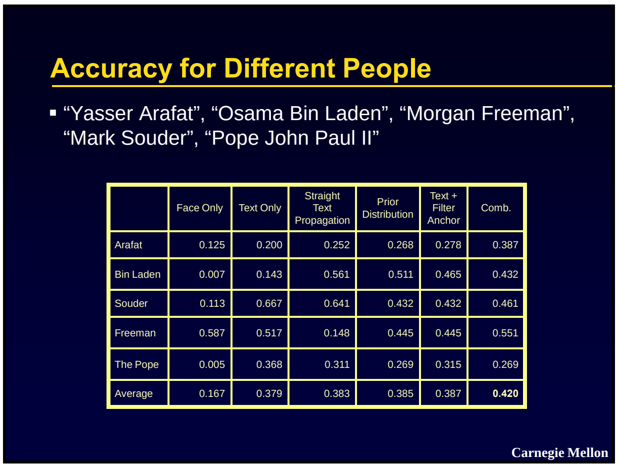
12
10
8
6
4
2
0
Frequency
-1
0
1
2
3
4
Distance (number of shots)

Madeleine Albright



Other Sources of Information

- Face Recognition
  - Unreliable technology
- Anchor and Commercial Detection
  - Filter out uninteresting shots

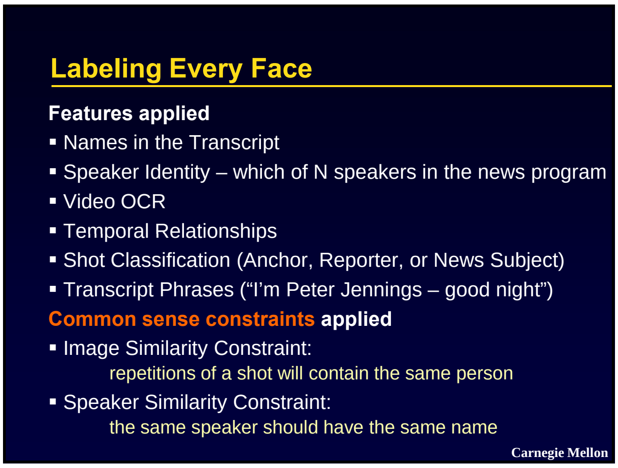


Accuracy for Different People

- "Yasser Arafat", "Osama Bin Laden", "Morgan Freeman", "Mark Souder", "Pope John Paul II"

	Face Only	Text Only	Straight Text Propagation	Prior Distribution	Text + Filter Anchor	Comb.
Arafat	0.125	0.200	0.252	0.268	0.278	0.387
Bin Laden	0.007	0.143	0.561	0.511	0.465	0.432
Souder	0.113	0.667	0.641	0.432	0.432	0.461
Freeman	0.587	0.517	0.148	0.445	0.445	0.551
The Pope	0.005	0.368	0.311	0.269	0.315	0.269
Average	0.167	0.379	0.383	0.385	0.387	**0.420**

Carnegie Mellon



Labeling Every Face

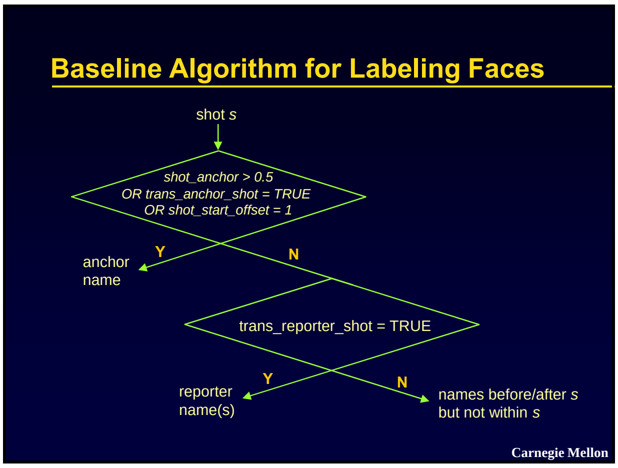
Features applied

- Names in the Transcript
- Speaker Identity – which of N speakers in the news program
- Video OCR
- Temporal Relationships
- Shot Classification (Anchor, Reporter, or News Subject)
- Transcript Phrases ("I'm Peter Jennings – good night")

Common sense constraints applied

- Image Similarity Constraint:
  repetitions of a shot will contain the same person
- Speaker Similarity Constraint:
  the same speaker should have the same name

Carnegie Mellon



Baseline Algorithm for Labeling Faces

shot *s*

shot_anchor > 0.5
OR trans_anchor_shot = TRUE
OR shot_start_offset = 1

Y

anchor name

N

trans_reporter_shot = TRUE

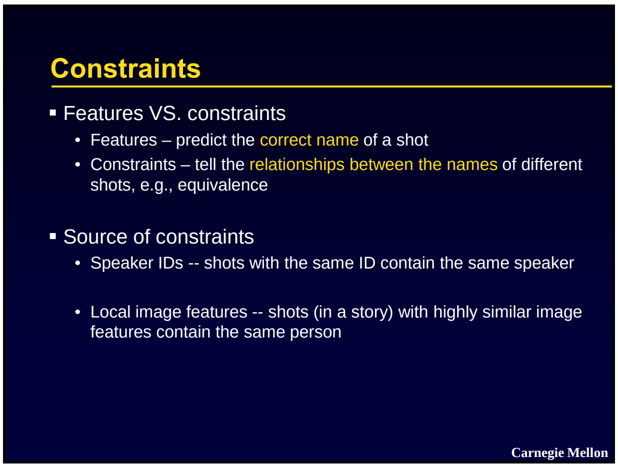
Y

reporter name(s)

N

names before/after *s* but not within *s*

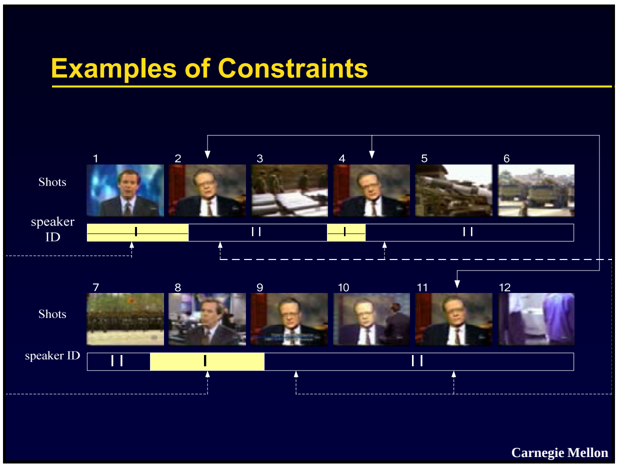
Carnegie Mellon



Constraints

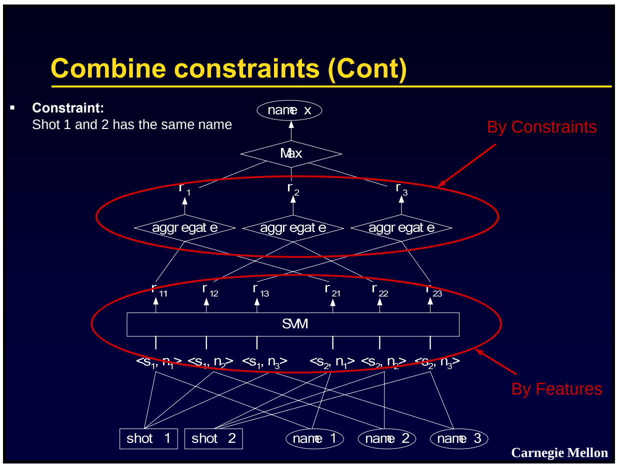
- Features VS. constraints
  - Features – predict the correct name of a shot
  - Constraints – tell the relationships between the names of different shots, e.g., equivalence
- Source of constraints
  - Speaker IDs -- shots with the same ID contain the same speaker
  - Local image features -- shots (in a story) with highly similar image features contain the same person

Carnegie Mellon



Examples of Constraints

Carnegie Mellon



Combine constraints (Cont)

- **Constraint:**
  Shot 1 and 2 has the same name

By Constraints

By Features

Carnegie Mellon

Experimental Set-up

- TREC Video Retrieval Evaluation dataset
  - 20 days of ABC Word News Tonight (10 hours)
  - 754 persons (shots) to be named
    - 237 news subjects
    - 373 anchors
    - 144 reporters
- Baseline approach
  - Name a person by the name temporally closest to it

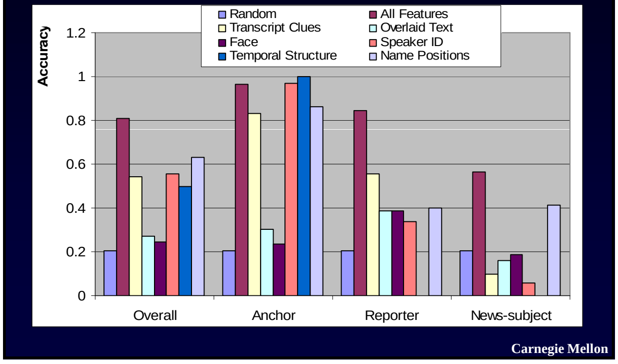
Accuracy of Person Naming

		Overall (754)	Anchor (373)	Reporter (144)	News subject (237)
Top-1 name	**Baseline**	0.561	0.834	0.359	0.256
	LM (our approach)	**0.728**	**0.958**	**0.656**	**0.424**
	LM + Constraints (Avg)	0.763	0.957	0.703	0.504
	LM + Constraints (Max)	0.771	0.957	0.734	0.512
Top-2 name	**Baseline**	0.659	0.860	0.422	0.48
	LM (our approach)	0.853	0.973	0.859	0.672
	LM + Constraints (Avg)	0.867	0.979	0.875	0.696
	LM + Constraints (Max)	0.856	0.979	0.875	0.664
Top-3 name	**Baseline**	0.710	0.877	0.515	0.56
	LM (our approach)	0.896	0.984	0.926	0.752
	LM + Constraints (Avg)	0.880	0.978	0.922	0.712
	LM + Constraints (Max)	0.875	0.978	0.922	0.696

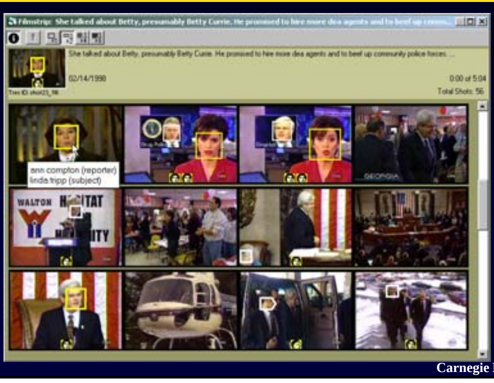
Observations

- In average, each person has 5 candidate names
  - Random baseline has 20% accuracy
  - Baseline achieves reasonably good performance
    - Poor on naming reporters and news-subjects
- Learning method (LM) is substantially better
  - Perfect on naming anchors and reporters
  - Much space for improvement on news-subjects
- Constraints are also helpful
  - Effective in boosting correct names to high ranks

Feature Contributions



Interface



Filmstrip: She talked about Betty, presumably Betty Currie. He promised to hire more dea agents and to beef up comm...

She talked about Betty, presumably Betty Currie. He promised to hire more dea agents and to beef up community police forces ...

02/14/1998

0:00 of 5:04

Total Shots: 56

ann compton (reporter)
linda tripp (subject)

Further Challenges

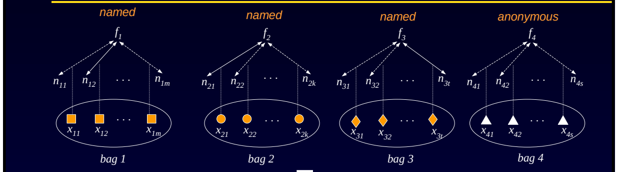
- Naming people in non-monologue shots
  - Name multiple people co-existing in a shot, speaking or not
  - A harder problem
    - many people are unnamed
    - speaker IDs can be meaningless or misleading
- Identifying unnamed people
  - Whether a person is anonymous or not
  - Whether a person looks similar to others we found
    - Similar face, similar scenes

Face labeling w/o labeled data

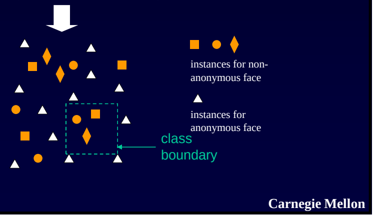
- To avoid the cost of manually labeling training data
- A missing piece in the video analysis research
- Without labels, there is still hidden information in the unlabeled data:
  1) Each face may have only one correct name
  2) It is easy to tell whether a face is named (i.e., having name in transcript) or anonymous
- How to make use of the hidden information?

Carnegie Mellon

Face labeling w/o labels


named
named
named
anonymous
f_1
f_2
f_3
f_4
n_{11} n_{12} . . . n_{1m}
n_{21} n_{22} . . . n_{2k}
n_{31} n_{32} . . . n_{3t}
n_{41} n_{42} . . . n_{4s}
x_{11} x_{12} . . . x_{1m}
x_{21} x_{22} . . . x_{2k}
x_{31} x_{32} . . . x_{3t}
x_{41} x_{42} . . . x_{4s}
bag 1
bag 2
bag 3
bag 4

Constraints: *bag 1, 2, 3* has one positive instance in it, and *bag 4* has no positive instances


instances for non-anonymous face
instances for anonymous face
class boundary

Carnegie Mellon

Multiple instance learning

- How to build a classifier from such constraints?
- Multi-Instance Learning (MIL) is the right tool -- learning with incomplete information of data labels
  - instance labels are unknown
  - instances are grouped into bags, and bag labels are known
    - if a bag is positive, at least one instance in it is positive,
    - if a bag is negative, all instances in it are negative
- MIL Methods: diverse density (DD), EM-DD, SVM variants, etc
- MIL Applications: drug activity prediction, content-based image retrieval, document classification, etc

Carnegie Mellon

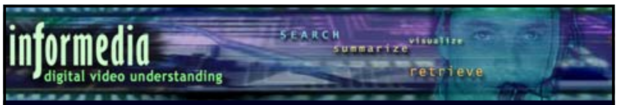
Face labeling *is* a MIL problem

- If (1) each a face-name association == an instance;

  (2) instances associated with a face == a bag of instances;

  (3) the face anonymity == bag label
- Bag labels as face anonymity required by MIL methods can be automatically obtained by heuristics
  - monologue faces w/o overlaid name are named
  - Faces appearing multiple times are mostly named
- So, MIL methods can solve face labeling w/o any human effort!

Carnegie Mellon



Location Annotation

Carnegie Mellon

Location annotation in news video

- Goal: annotate each video shot with the location of the scene
- Locations for video analysis and retrieval
  - Localized search: *"Find all the scenes showing suicide bombings in Baghdad, Iraq"*
  - Summarization: *"List all the Asian countries hit by tsunami last year"*
  - Categorizing and browsing news video by locations

Carnegie Mellon

The state-of-the-art

- Scene type classification
  - Indoor, outdoor, meetings, street, etc
  - Very restrictive vocabulary
- Matching a shot with scenes of known locations
  - Expensive, not scalable to news video
- GPS information
  - Unavailable to news video

Previous work at Informedia

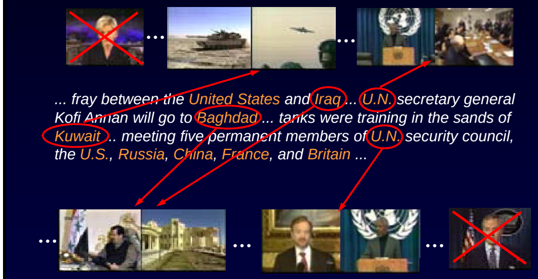
- Map-based browsing of news video by locations

- The stories are organized by the locations mentioned by the transcript

News locations ≠ mentioned locations

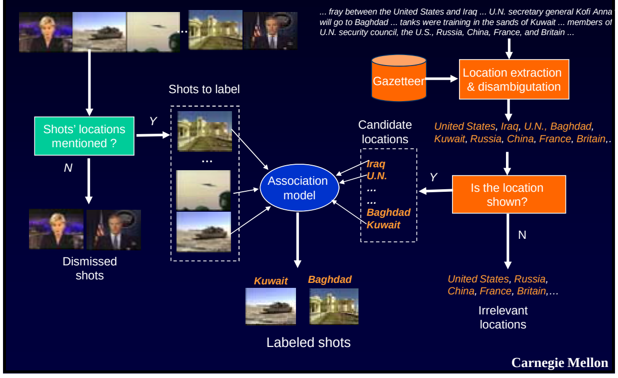
- Locations of news are typically mentioned in the transcript
  - Closed-captions, ASR text
- Mentioned locations are far from true locations
  - Multiple locations mentioned, some never “show up”
  - A story may switch between many locations
  - Shots with unmentioned locations
  - Shots without locations (e.g., artificial shots)

Location annotation: an example



... fray between the United States and Iraq ... U.N. secretary general Kofi Annan will go to Baghdad ... tanks were training in the sands of Kuwait ... meeting five permanent members of U.N. security council, the U.S., Russia, China, France, and Britain ...

System framework



... fray between the United States and Iraq ... U.N. secretary general Kofi Anna will go to Baghdad ... tanks were training in the sands of Kuwait ... members of U.N. security council, the U.S., Russia, China, France, and Britain ...

Gazetteer

Location extraction & disambigutation

United States, Iraq, U.N., Baghdad, Kuwait, Russia, China, France, Britain,..

Is the location shown?

Y

Candidate locations

Iraq
U.N.
...
...
Baghdad
Kuwait

N

United States, Russia, China, France, Britain,…

Irrelevant locations

Shots’ locations mentioned ?

Y

Shots to label

...

N

Dismissed shots

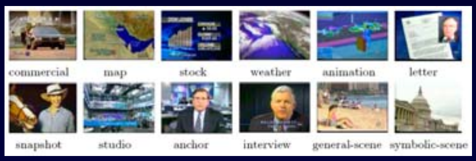
Association model

Kuwait *Baghdad*

Labeled shots

Shots w/o mentioned locations

- Case 1: Shots with no legitimate locations
  - e.g. artificial shots
- Case 2: Shot with unimportant or self-contained locations
  - e.g. studio shots, general scene



Find shots w/o mentioned locations

- A SVM classifier to find shots w/o mentioned locations using heuristics
  - Semantic concepts: *anchor, commercial, studio*
  - Motion feature: *pixel difference*
  - Story genre: *politics, technology, health, sports, business*
- Performance
  - Data set: 6219 ABC News shots (Trecvid 2004)
  - Result: 89.7% accuracy
    - Find 4072 shots w/o locations, miss only 492

Location extraction

- Extracting locations using BBN named-entity detector
  - locations, e.g., *"California"*
  - self-contained organizations, e.g., *"Capitol Hill"*
- Locations are ambiguous
  - Synonymity – multiple expressions of the same location
    - e.g. *"United Kingdom"* and *"Great Britain"*, *"Los Angeles"* and *"LA"*, *"Mosel, Iraq"* and *"Mosul, Iraq"*
  - Polysemy – multiple locations with the same name
    - e.g. *"London, UK"* vs. *"London, Ontario"*, *"Georgia"* as a state or a country
    - A serious problem in U.S. – 24 Paris, 63 Springfileds

Location Disambigutation

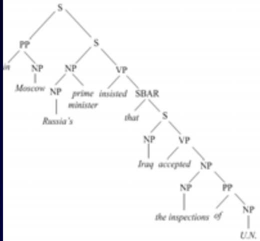
- Transform a location term into a physical location
- Resolve synonymity based on a gazetteer
  - Location → canonical location, e.g., *"U.S."* → *"United States"*
  - Manually adding rules
- Resolve polysemy by context
  - Locations mentioned in the proximity
    - e.g. *"Ontario"* immediately after or close to *"London"*
    - *e.g. "Georgia" near "North Carolina"*
  - Default reference
    - *e.g. "Paris", "Baghdad", "Damascus"*

Which location is shown?

- Example: *"In Moscow, Russia's prime minister insisted that Iraq accepted the inspections of United Nation"*
- Syntactic analysis helps
  - Prepositional phrase -- likely, depending on preposition
    - e.g. *"In Moscow", "of United Nation"*
  - Subject/object – unlikely
    - e.g. *"Iraq accepted …"*
  - Modifier – maybe
    - *e.g. "Russia's prime minister"*
- Other heuristics: location type, speaker, etc

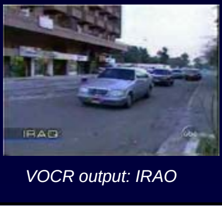
Analysis of syntactic structure

- Derive the syntactic role of a location from the parse tree
  - Link grammar parser



Mapping locations to shots

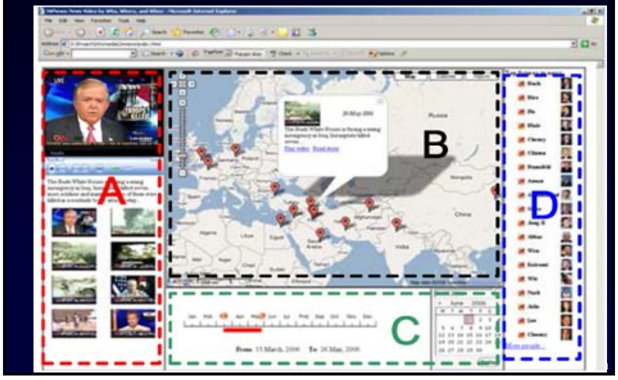
- An association model between locations and shots
- Supervised method based on multimodal features
  - Temporal distance between shot & location
  - String similarity between location & VOCR text



Edit distance:
Iraq: 0.25
U.S.: 1.0
France: 0.67
Russia: 1.0

VOCR output: IRAO

Informedia News- Google Map Interface

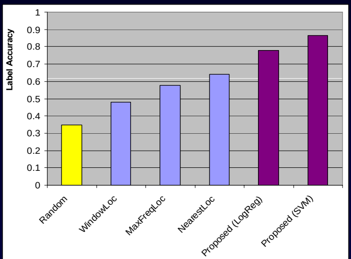


Experiments in TRECVID

- Data set: 10-hour ABC News in TRECVID 2004
  - 6219 shots, among which 1768 has location (s)
- Baseline approaches
  - WindowLoc: label a shot by all the locations in a window
  - MaxFreqLoc: label a shot by the most frequently appeared location in the story
  - NearestLoc: label a shot by the nearest location

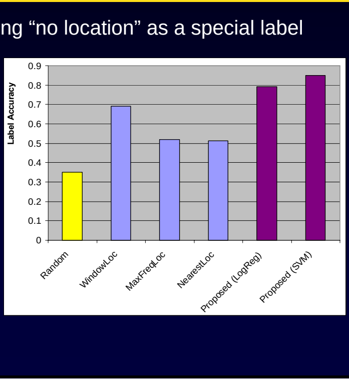
Performance on shots with locations

- Accuracy: ratio that the most likely location of a shot is correct



Performance on all shots

- Treating “no location” as a special label



Conclusion

- Annotating news video locations is feasible
  - Built on mature techniques
    - story segmentation, named-entity detection, etc
  - Good performance: around 80-90%
  - High efficiency
    - offline effort of labeling training data
    - fast predictions
  - General applicability

informedia digital video understanding SEARCH summarize visualize retrieve

Questions?