News Video Indexing and Abstraction by Specific Visual Cues: MSC and News Caption

Jiang, Fan

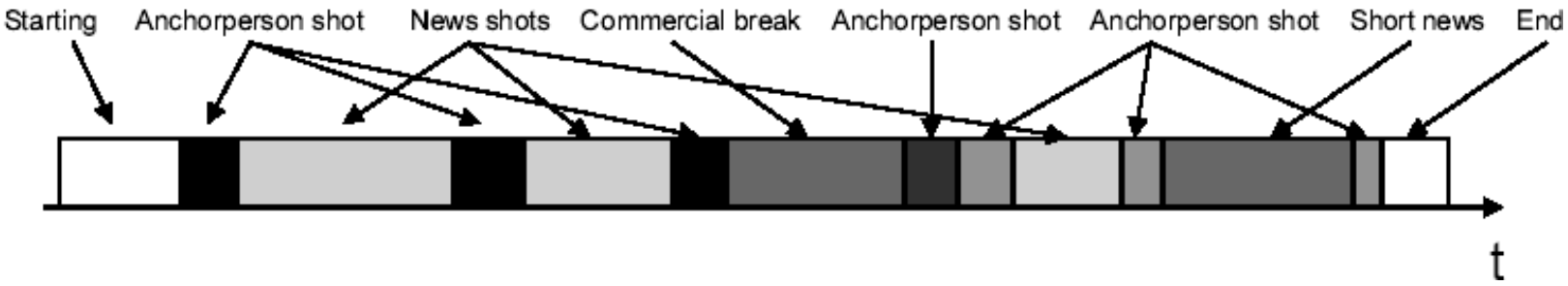
Supervisor: Zhang, Y.J.

Introduction

- "Information Explosion"
- Efficiently access video data
- Content-Based Video Retrieval
- Universal solution for high-level
- Domain-specific analysis:
  - sports, movies, news broadcasts, etc

News Video Analysis

- Common temporal structure



- Specific visual cues
  - Headline
  - Anchorperson

Technology

- Shot Boundary Detection
- Anchorperson Detection
- Video OCR
- Video Representation and Abstraction

Outline of My Work

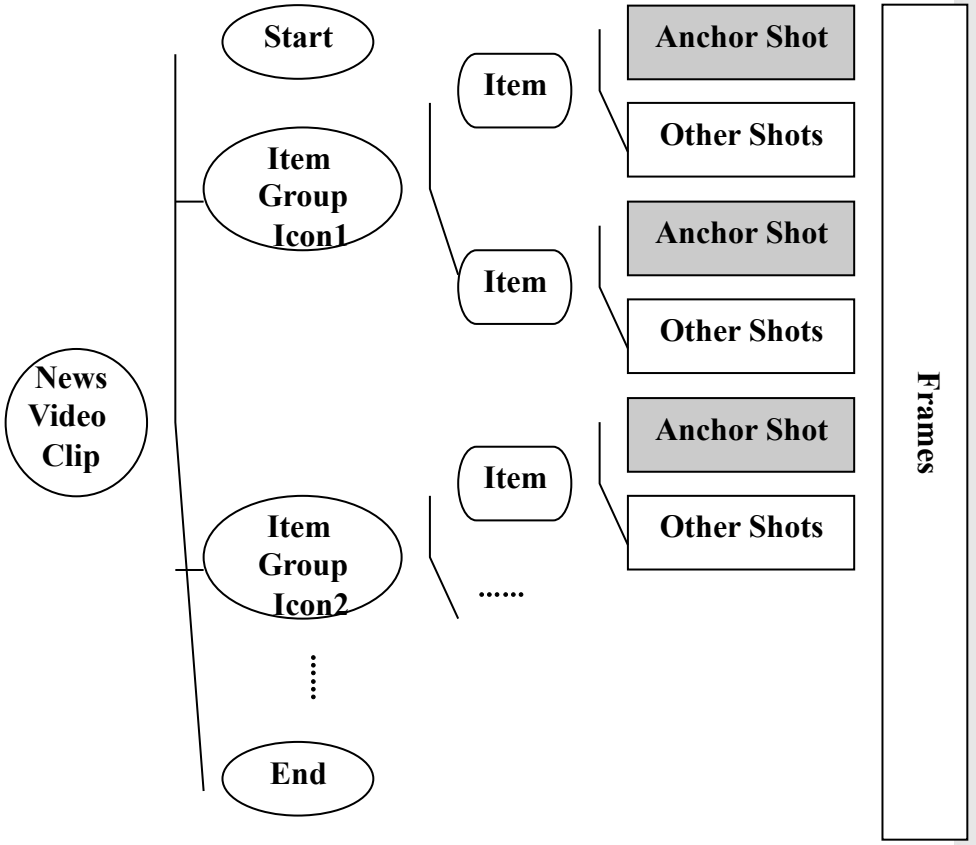
- Analysis based on two specific visual cues
  - Cue1: MSC
  - Cue2: News Caption
- Structure for news content indexing
- Generate abstraction of content

Hierarchical Structure

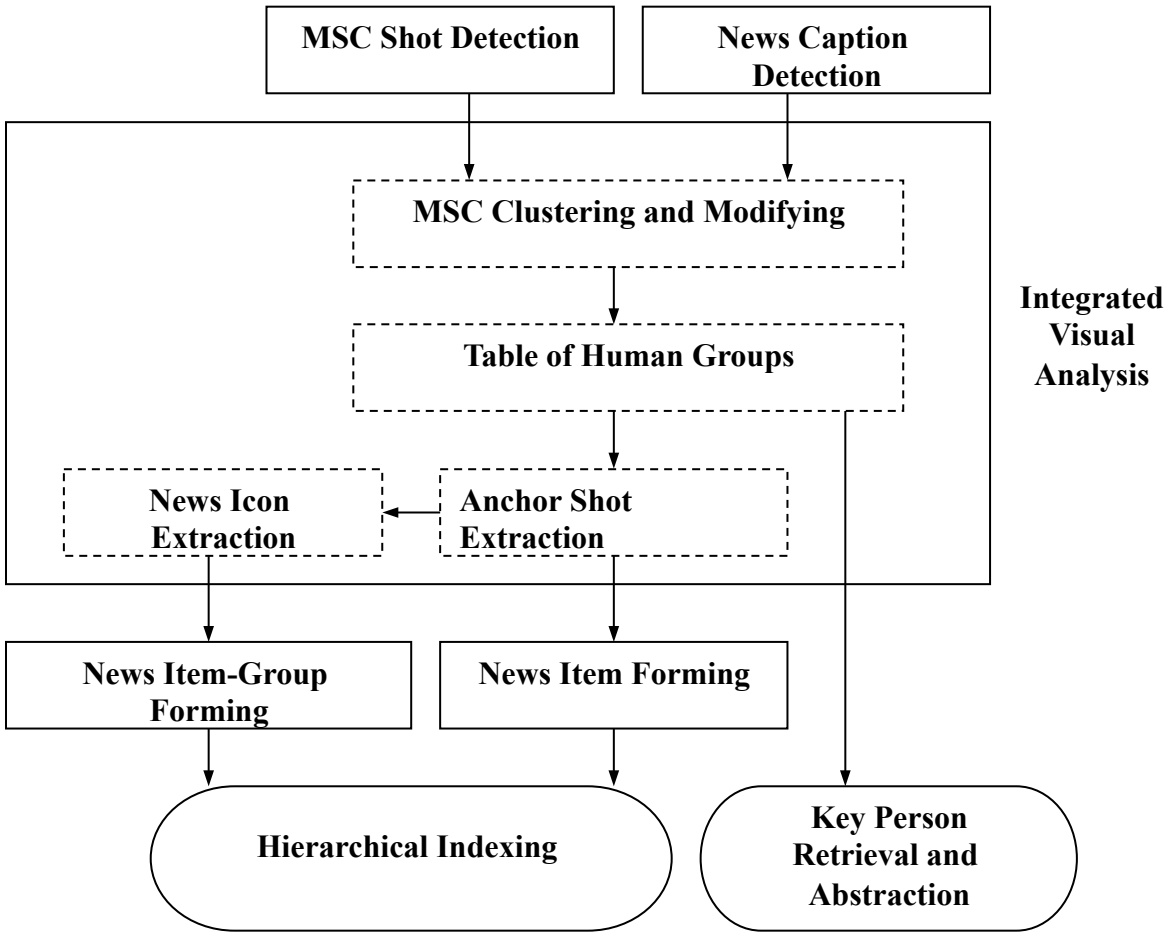
- Anchor shot

News Icon

News Caption



Overall Framework



MSC Shot Detection (1)

- Definition of Main Speaker Close-Up (MSC)

  - anchorperson detection interfered by "talking heads"
    - reporters, interviewees, and lecturers

"A single camera-focused talking head in screen" ⇒ **MSC**

MSC Shot Detection (2)

- Functions of MSC extraction
  - Main speakers serve as abstraction
    - key person indexing and retrieval
  - Preprocessing of anchor detection
  - Facilitate news icon extraction
- Detection approaches
  - Face detection in video
    - complicated, time-consuming
  - Model fitting based on statistical features
    - color / shape

MSC Shot Detection (3)

- My approach: Head-Motion-Model fitting
  - Spatial and motional features
    - **Shot activity:** lower than common shots, stronger than stationary scene shots;
    - **Activity concentrates** on a single dominant talking head in center of screen;
    - **Fixed location of talking heads**:

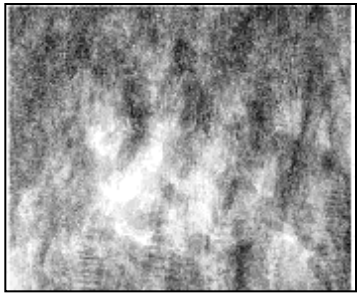
      Middle (M) / Right (R) / Left (L)
  - Need a shot activity measurement
    - measure frame dissimilarity
    - reflect motion distribution

MSC Shot Detection (4)

- Map of Average Motion in Shot (MAMS)

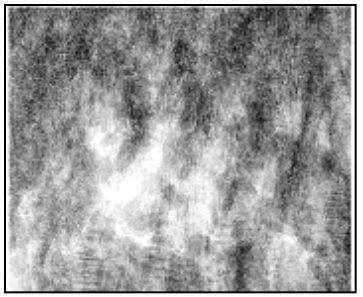
$$MAMS_h(x,y) = \frac{1}{M} \sum_{\substack{i=1 \\ i=i+v}}^{M} \left| f_i(x,y) - f_{i+v}(x,y) \right| \qquad M = \frac{N}{v}$$



MSC Shot Detection (5)

- Criteria 1



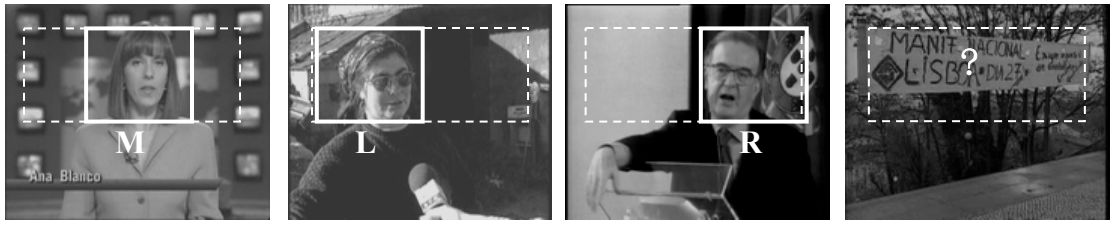
$0.0022\ (10^{-3})$ < $0.0384\ (10^{-2})$ < $0.3473\ (10^{-1})$

$(T_S = 10^{-3})$ $(T_C = 10^{-1})$

$$Ts < \frac{MAMS_n}{x \cdot y} < Tc$$

MSC Shot Detection (6)

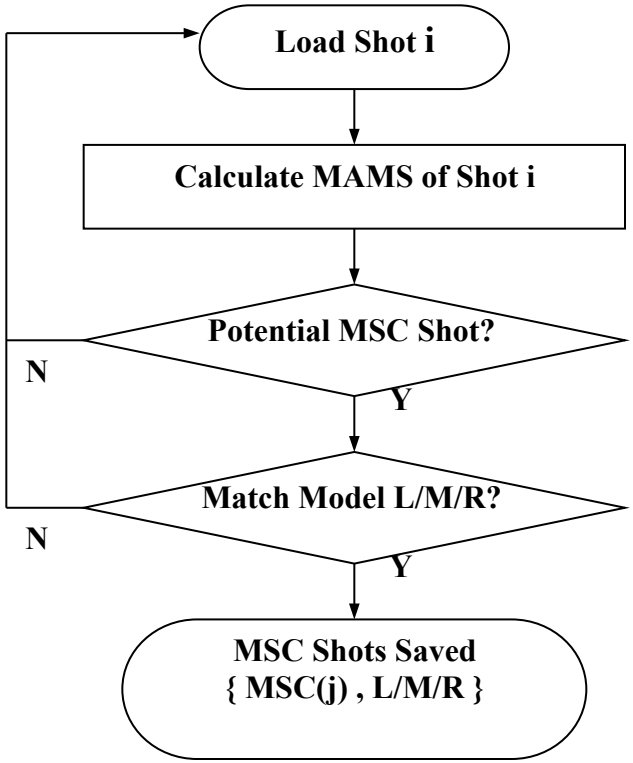
- Criteria 2



$$\frac{\max\limits_{(L/M/R)}\left\{\sum\limits_{(x,y)\in(L/M/R)Box} MAMS_n(x,y)\right\}}{\sum\limits_{(x,y)\in HeadBand} MAMS_n(x,y)} > T_H$$

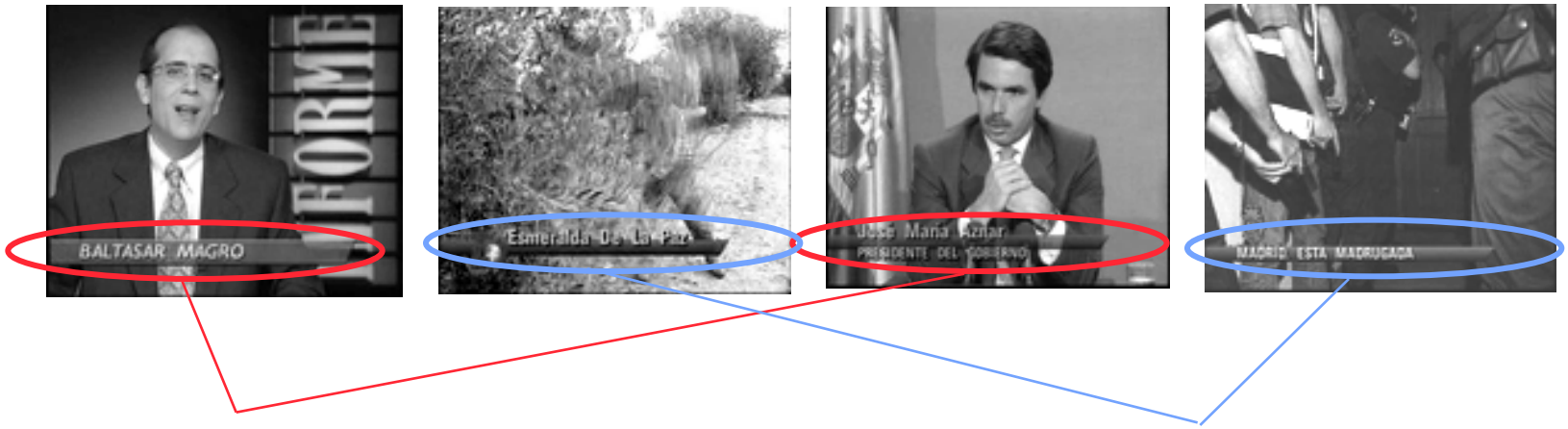
MSC Shot Detection (7)

- MSC shot detection scheme



News Caption Detection (1)

- Two scenarios:



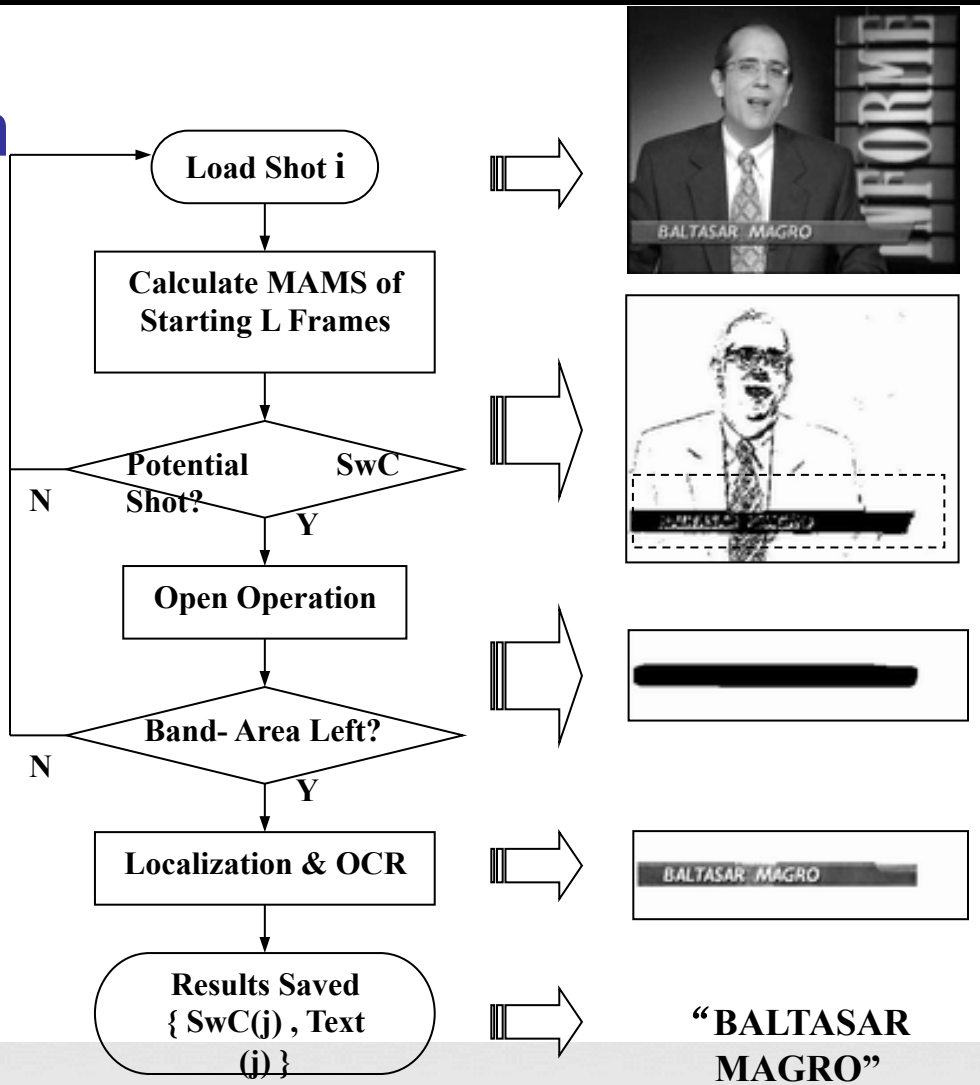
name of camera-focused person

annotation of news story

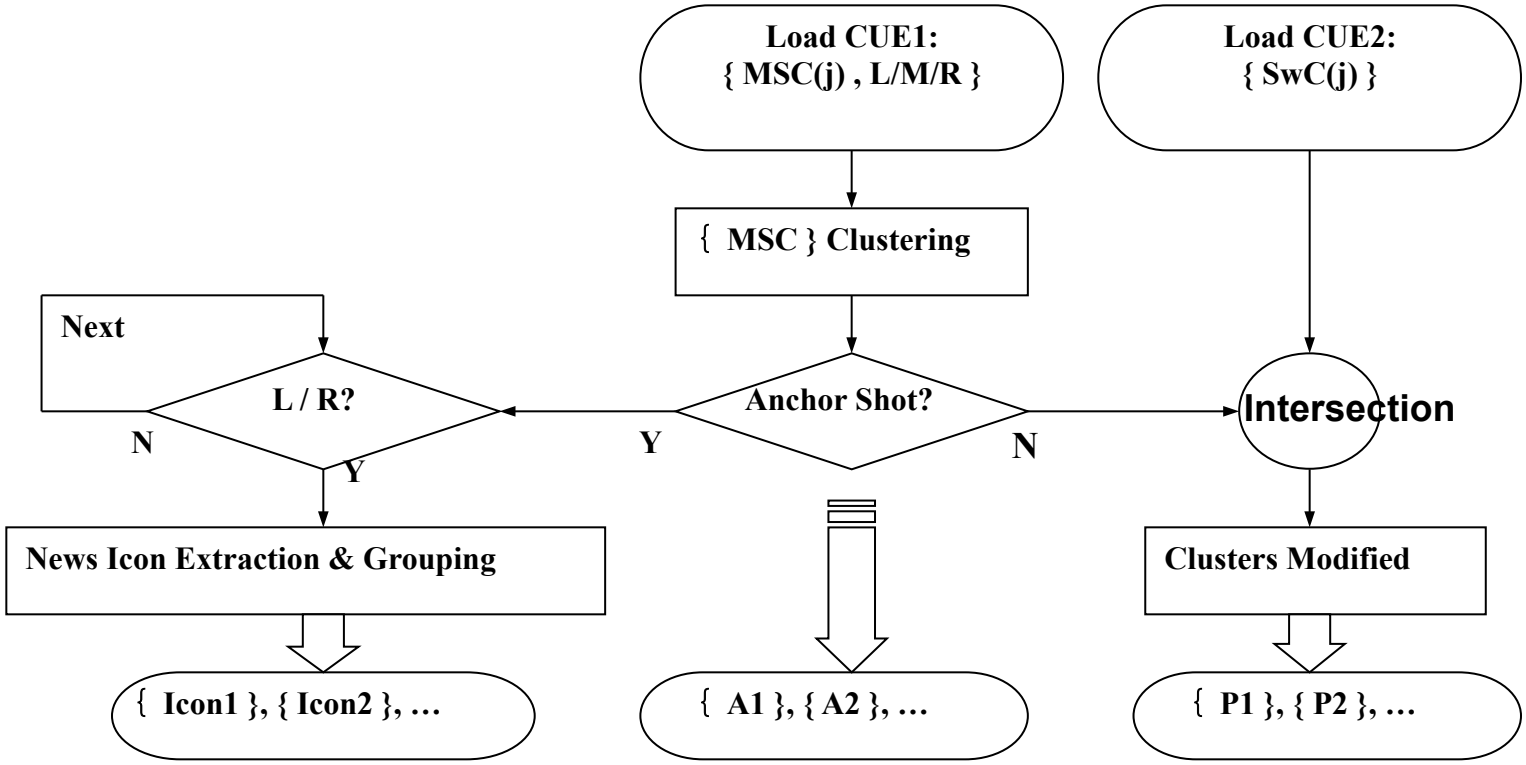
Indexing MSC clustering results

News Caption Detection (2)

- Specific-region text-detecting algorithm

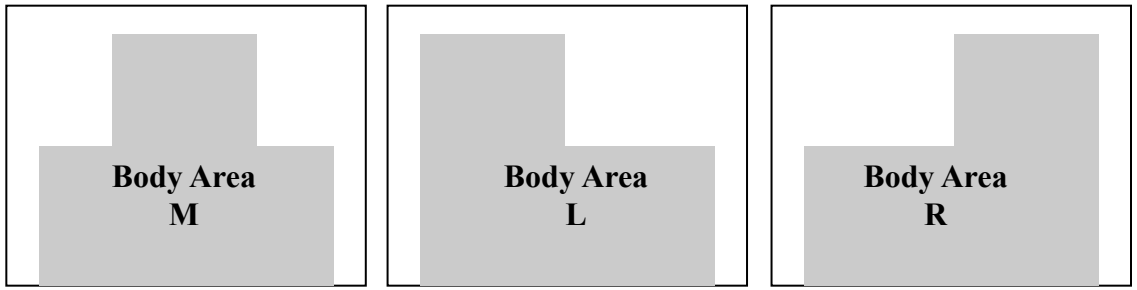


Integrated Analysis by Two Cues



MSC Clustering (1)

- Assumption
  - same “main speaker” has similar appearances (like clothes’ color and body size)
  - Body-Area-Models according to MSC detection results



MSC Clustering (2)

- Color Histogram Intersection (CHI)

$$Sim\left\{\underset{L/M/R}{MSC1},\underset{L/M/R}{MSC2}\right\}=\frac{\sum_{h=1}^{H}\min\left\{\overline{MSC1}_h(Y,U,V),\overline{MSC2}_h(Y,U,V)\right\}}{\sum_{h=1}^{H}\overline{MSC1}_h(Y,U,V)}$$

$$\overline{MSC}=\frac{1}{N}\sum_{i=1}^{N}\sum_{(x,y)\in L/M/R}f_i(x,y)(Y,U,V)$$

- MSC1 and MSC2 are clustered together, when

$$Sim\left\{MSC1,\ MSC2\right\}>Tg$$

MSC Clustering Modified (1)

- *Tg*: a loose value → different persons in one group?
- Modified by news captions → construct a table of key person (one person in each group)

news caption of speaker's name appears in one of MSC shots

MSC Clustering Modified (2)

- 3-step scheme
  - Step1. Intersection

$$\{MSCwC\} = \{MSC\} \cap \{SwC\}$$

  - Step2. Splitting
    - Split groups with different captions (different names) in
  - Step3. Indexing and Tabling
    - indexed textually by a person's name (caption text)
    - Indexed visually by representative close-up shot (MSCwC)

Anchor Shot Identification (1)

- Based on 3 temporal features
  - more times of repetition
  - more disperse in time
  - totally longer range of time

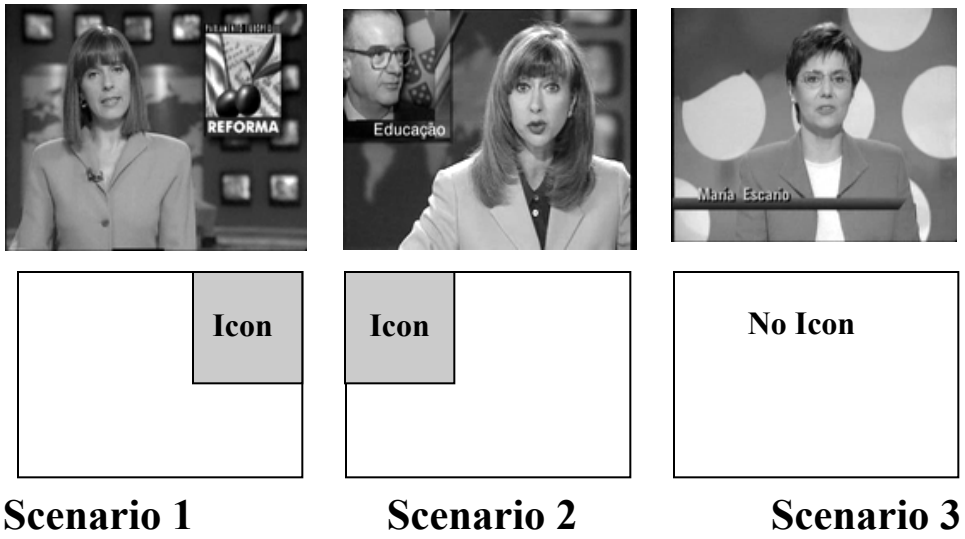
	Shot length	# of Shots (every 20 minutes)	Shot interval	Range of all shots
Anchor group	15.72sec	9 shots	12 shots	>60 shots
Any other MSC group	10.38sec	2 shots	5 shots	<10 shots

Anchor Shot Identification (2)

- Select MSC groups satisfying:
  - Total number of shots in the group is more than a predefined value (i.e. 5 shots per 20 minutes);
  - The average shot interval should be more than a predefined value (i.e. 10 shots);
  - The range of shots of the group is higher than a predefined value (i.e. 50 shots).

News Icon Extraction

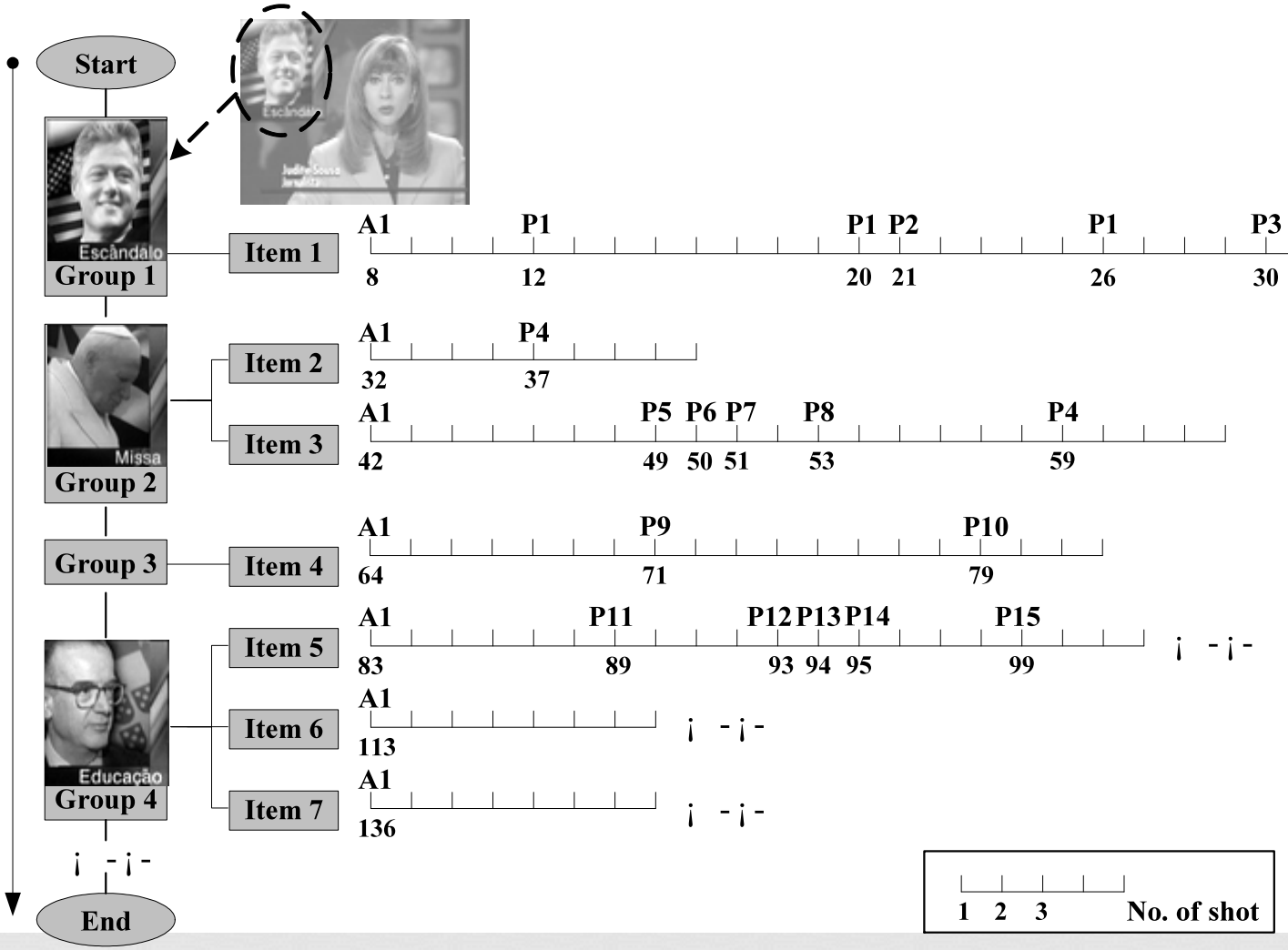
- 3 possible scenarios



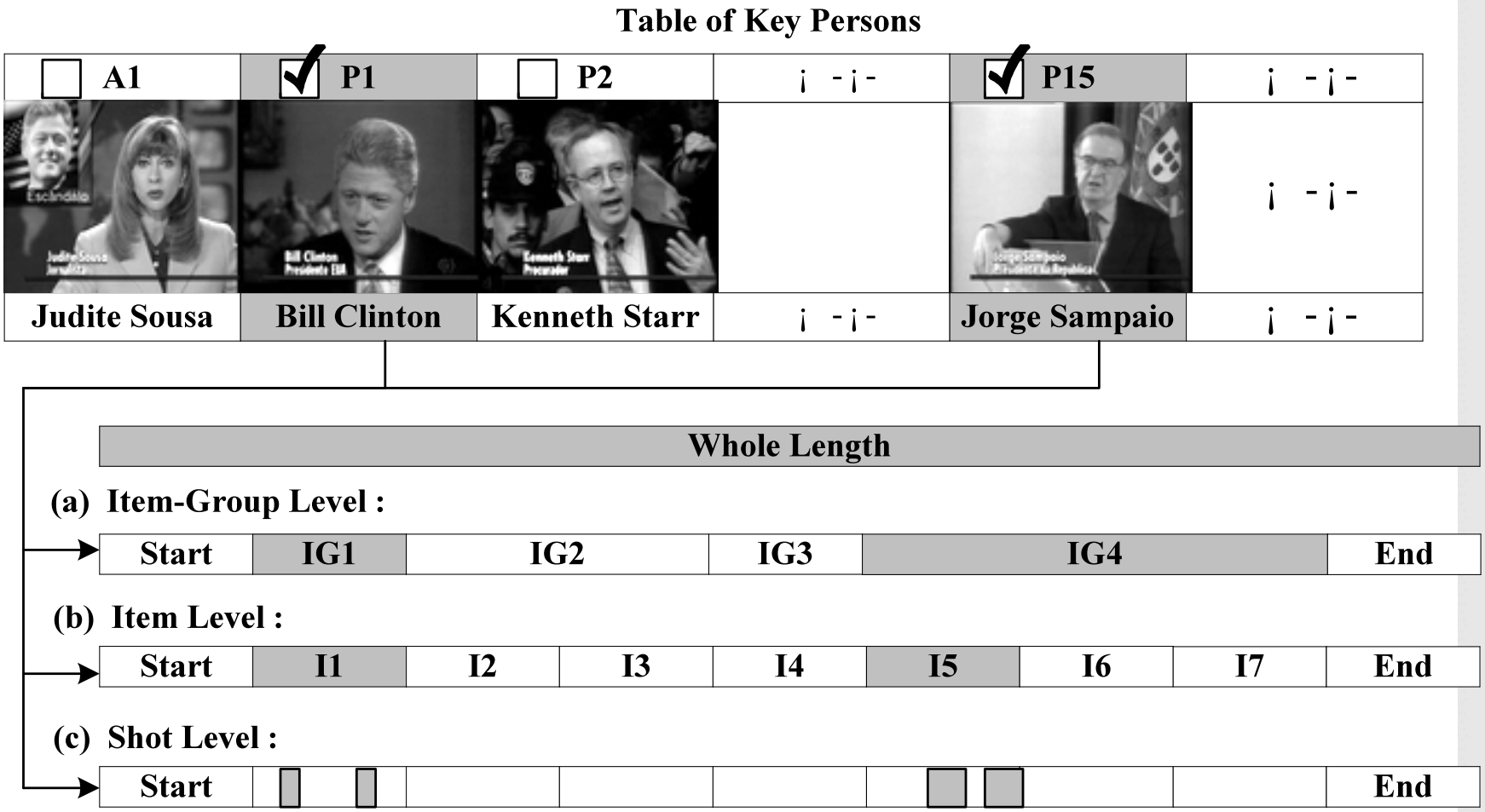
- Two adjacent news items with icons are grouped together when

$$Sim\{Icon1, Icon2\} > Ti$$

Hierarchical News Video Representation



Multi-Level Abstraction by Selective MSCs



MSC Shot Detection Experiment

News video	# of MSC shots	Candidates detected by MAMS analysis	Excluded by Head-Motion-Model fitting	Final results		
				Hits	Misses	False alarms
Clip 1	22	31	6	22	0	3
Clip 2	24	35	10	23	1	2
Clip 3	43	68	22	40	3	6

recall = 95.51% precision = 88.54%

News Caption Detection Experiment

News video	# of Shots with Caption	Candidates detected by MAMS analysis	Excluded by Band-Area detection	Final results		
				Hits	Misses	False alarms
Clip 1	21	33	10	20	1	3
Clip 2	19	28	8	19	0	1
Clip 3	32	46	11	30	2	5

recall = 95.83% precision = 88.46%

MSC Shot Clustering Results

News video	# of persons in MSC	CHI clustering results {MSC}	Corrected by {SwC}	Final results		
				Hits	Misses	False alarms
Clip 1	9 (with 1 anchor)	9	1	9	0	1
Clip 2	15 (with 2 anchors)	16	2	14	1	4
Clip 3	29 (with 1 anchor)	27	5	26	3	6

recall = 92.45% precision = 81.67%

Anchor Shot Identification Results

News video	# of anchor persons	# of anchor shots	Anchor identification results			L/M/R results
			Hits	Misses	False alarms	Errors
Clip 1	1	9	9	0	0	0
Clip 2	2	8	8	1	0	1
Clip 3	1	8	8	0	0	0

recall = 96.15% precision = 100%

MSC-Based Video Abstraction Experiment

News video	Shot Level	Item Level	Item-Group Level
Clip 1	91 shots	8 items	8 item-groups
Clip 2	90 shots	7 items	7 item-groups
Clip 3	156 shots	7 items	4 item-groups

	# of MSC groups	Shot Level	Item Level	Item-Group Level
Clip 3*	32	156 shots	7 items	4 item-groups
Abstraction with all MSCs	32	44 shots (07:21)	7 items (17:58)	4 item-groups (17:58)
Abstraction with selective MSCs	2	8 shots (00:35)	2 items (06:09)	2 item-groups (11:13)

* Whole length of Clip3 = (17:58)

Conclusion

- Analysis based on two specific visual cues
  - Cue1: MSC (detection and clustering)
  - Cue2: News Caption (intersection)
- Build a table of key persons
- Generate hierarchical structure:
  - frames, shots, news items (by anchor shot), news item-groups (by news icon) and video clip

→ Retrieval, Abstraction

Future Trends

- Broaden MSC to two-person version
- Adaptively set up head motion model
- Combine speech signals in complex situations