TRECVID 2010 INSTANCE RETRIEVAL PILOT

AN INTRODUCTION ....

Wessel Kraaij
TNO, Radboud University Nijmegen

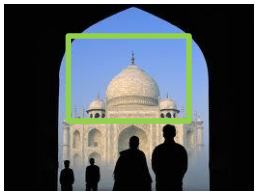
Paul Over
NIST

Background

- The many dimensions of searching and indexing video collections
  - hard tasks: search task, semantic indexing task
  - easier tasks: shot boundary detection, copy detection

- Instance search:
  - searching with a visual example (iamge or video) of a target person/location/object
  - hypothesis: systems will focus more on the target, less on the visual/semantic context

- Existing commercial applications using visual similarity
  - logo detection (sports video)
  - product / landmark recognition (images)




Differences between INS and SIN

INS	SIN
Very few training images (probably from the same clip)	Many training images from several clips
Many use cases require real time response	Concept detection can be performed off-line
Targets include unique entities (persons/locations/objects) or industrially made products	Concepts include events, people, objects, locations, scenes. Usually there is some abstraction (car)
Use cases: forensic search in surveillance/ seized video, video linking	Automatic indexing to support search.

Task

Example use case: browsing a video archive, you find a video of a person, place, or thing of interest to you, known or unknown, and want to find more video containing the same target, but not necessarily in the same context.

For example:

- All the video taken over the years in the backyard of your house on Main Street.
- All the clips of your favorite Aunt Edna
- All the segments showing your company logo.

System task:

- Given a topic with:
  - example segmented images of the target
  - the video from which the images were taken
  - a target type (PERSON, CHARACTER, PLACE, OBJECT)
- Return a list of up to 1000 shots ranked by likelihood that they contain the topic target

Data

180 hours of Dutch educational, news magazine, and cultural programming (Netherlands Institute for Sound & Vision)

~ 60 000 shots

Containing recurring

- people as themselves (e.g., presenters, hosts, VIP's)
- people as characters (e.g., in comic skits)
- objects (including logos)
- locations

Topics

<videoInstanceTopic text="**Professor Fetze Alsvanouds from the University of Harderwijk (Aart Staartjes)**" num="**9005**" type="**CHARACTER**">

<imageExample src="**9005.1.src.JPG**" target="**9005.1.target.JPG**" mask="**9005.1.mask.png**" object="**9005.1.object.png**" outline="**9005.1.outline.png**" vertices="**9005.1.vertices.xml**" video="**BG_37796.mpg**" />

<imageExample src="**9005.2.src.JPG**" target="**9005.2.target.JPG**" mask="**9005.2.mask.png**" object="**9005.2.object.png**" outline="**9005.2.outline.png**" vertices="**9005.2.vertices.xml**" video="**BG_37796.mpg**" />

<imageExample src="**9005.3.src.JPG**" target="**9005.3.target.JPG**" mask="**9005.3.mask.png**" object="**9005.3.object.png**" outline="**9005.3.outline.png**" vertices="**9005.3.vertices.xml**" video="**BG_37796.mpg**" />

<imageExample src="**9005.4.src.JPG**" target="**9005.4.target.JPG**" mask="**9005.4.mask.png**" object="**9005.4.object.png**" outline="**9005.4.outline.png**" vertices="**9005.4.vertices.xml**" video="**BG_37796.mpg**" />

<imageExample src="**9005.5.src.JPG**" target="**9005.5.target.JPG**" mask="**9005.5.mask.png**" object="**9005.5.object.png**" outline="**9005.5.outline.png**" vertices="**9005.5.vertices.xml**" video="**BG_37796.mpg**" />

… </videoInstanceTopic>

Topics – segmented example images

9005.1.src.JPG

9005.1.target.JPG

9005.1.outline.png

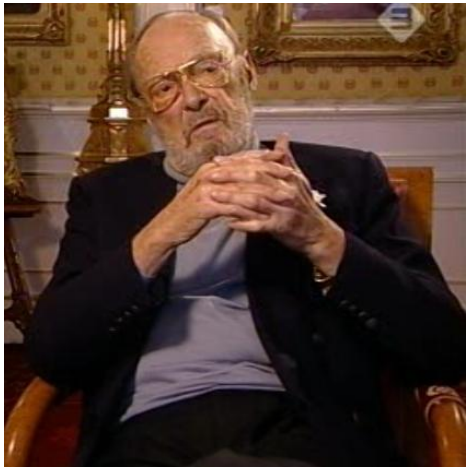
9005.1.object.png

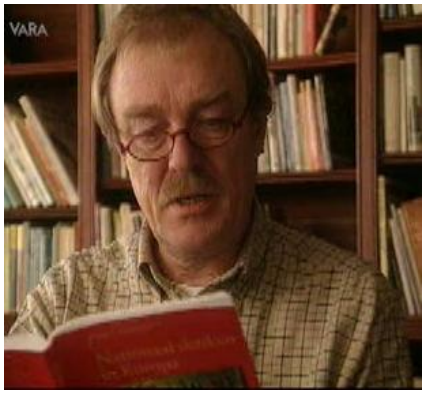
+ Outline vertex coordinates

+ Full video file name

9005.1.mask.png

Topics – 8 People (as themselves)





Topics – 5 people (as Characters)

Topics – 8 Objects

Topics – 1 Location

TV2010 Finishers (15)

CCD	INS	***	***	---	***	AT&T Labs – Research
CCD	INS	KIS	---	SED	SIN	Beijing University of Posts and Telecom.-MCPRL
---	INS	KIS	---	---	---	Dublin City University
***	INS	KIS	---	---	***	Hungarian Academy of Sciences
---	INS	KIS	MED	---	SIN	Informatics and Telematics Inst.
---	INS	---	---	***	SIN	JOANNEUM RESEARCH
---	INS	KIS	MED	***	SIN	KB Video Retrieval (Etter Solutions LLC)
---	INS	***	***	---	SIN	Laboratoire d'Informatique de Grenoble for IRIM
CCD	INS	---	---	***	---	Nanjing University
CCD	INS	***	***	***	SIN	National Inst. of Informatics
---	INS	---	---	---	---	NTT Communication Science Laboratories-NII
---	INS	---	---	---	---	TNO ICT - Multimedia Technology
***	INS	---	---	---	---	Tokushima University
---	INS	KIS	***	***	SIN	University of Amsterdam
***	INS	***	***	***	***	Xi'an Jiaotong University

** : group applied but didn't submit

-- : group didn't apply for the task

Evaluation

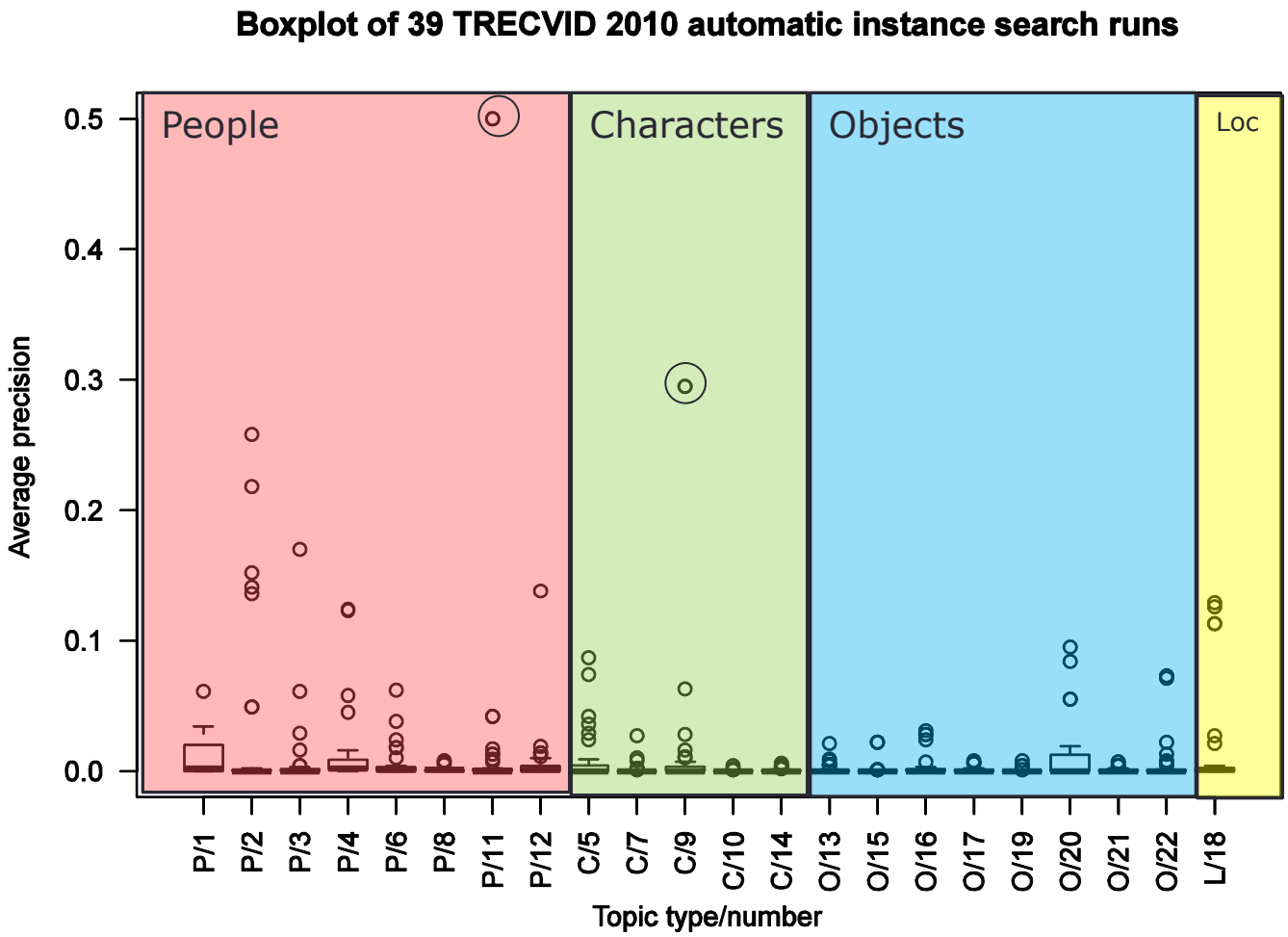
For each topic, the submissions were pooled and judged down to at least rank 100 (on average to rank 130), resulting in 68770 shots.

10 NIST assessors played the clips and determined if they contained the topic target or not.

1208 clips (=avg. 55 / topic) did contain the topic target.

trec_eval was used to calculate average precision, recall, precision, etc.

Evaluation – results by topic/type - automatic



P/1 George W. Bush (61)
P/2 George H. W. Bush (28)
P/3 J. P. Balkenende (140)
P/4 Bart Bosch (140)
P/6 Prince Bernhard (36)
P/8 Jeroen Kramer
P/11 Colin Powell (4)
P/12 Midas Dekkers (174)

C/5 Professor Fetze Alsvanouds (36)
C/7 The Cook (14)
C/9 Two old ladies, Ta en To (9)
C/10 one of two officeworkers (68)
C/14 Boy Zonderman (15)

O/13 IKEA logo on clothing (25)
O/15 black robes with white bibs (28)
O/16 zebra stripes on pedestrian crossing (27)
O/17 KLM Logo (20)
O/19 Kappa Logo (6)
O/20 Umbro Logo (38)
O/21 tank (28)
O/22 Willem Wever van (15)

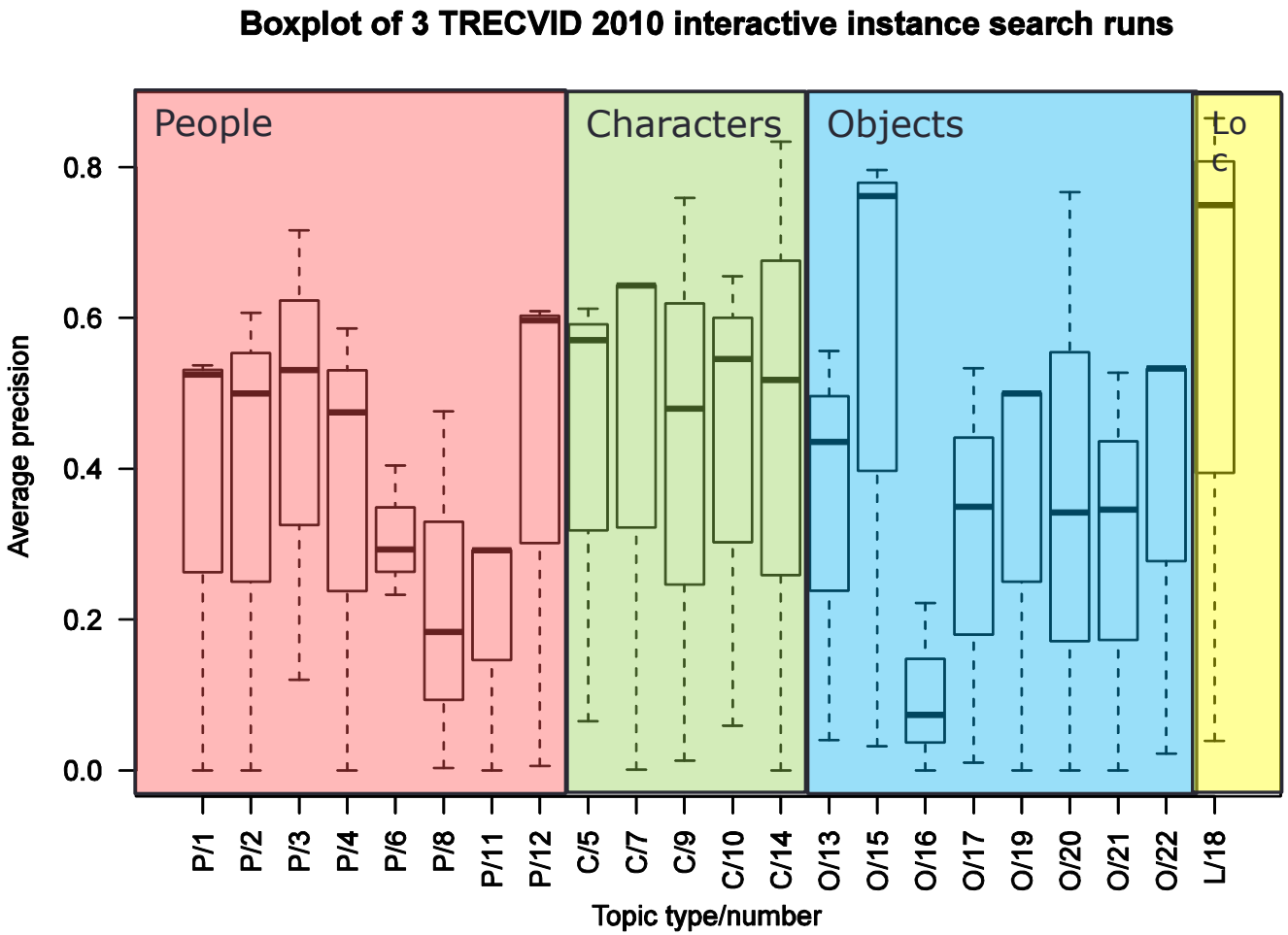
L/18=interior of Dutch parliament (52)

Evaluation – top half based on MAP

			MAP	MedianAP
I X N	ITI-CERTH	2	0.534	0.535
I X N	ITI-CERTH	1	0.524	0.532
I X N	XJTU_1	1	0.029	0.005
F X N	NII.kaori	2	0.033	0.000
F X N	NII.kaori	1	0.033	0.000
F X N	MCPRBUPT1	3	0.026	0.005
F X N	bpacad	3	0.026	0.001
F X N	MCPRBUPT1	1	0.025	0.005
F X N	bpacad	2	0.023	0.001
F X Y	KBVR_4	4	0.012	0.000
F X N	UvA_2	3	0.011	0.001
F X N	UvA_2	2	0.011	0.001
F X Y	KBVR_1	1	0.010	0.000
F X N	UvA_2	4	0.010	0.001

Mean is not very informative. It is due to a small number of non-zero scores.

Evaluation – results by topic/type - interactive



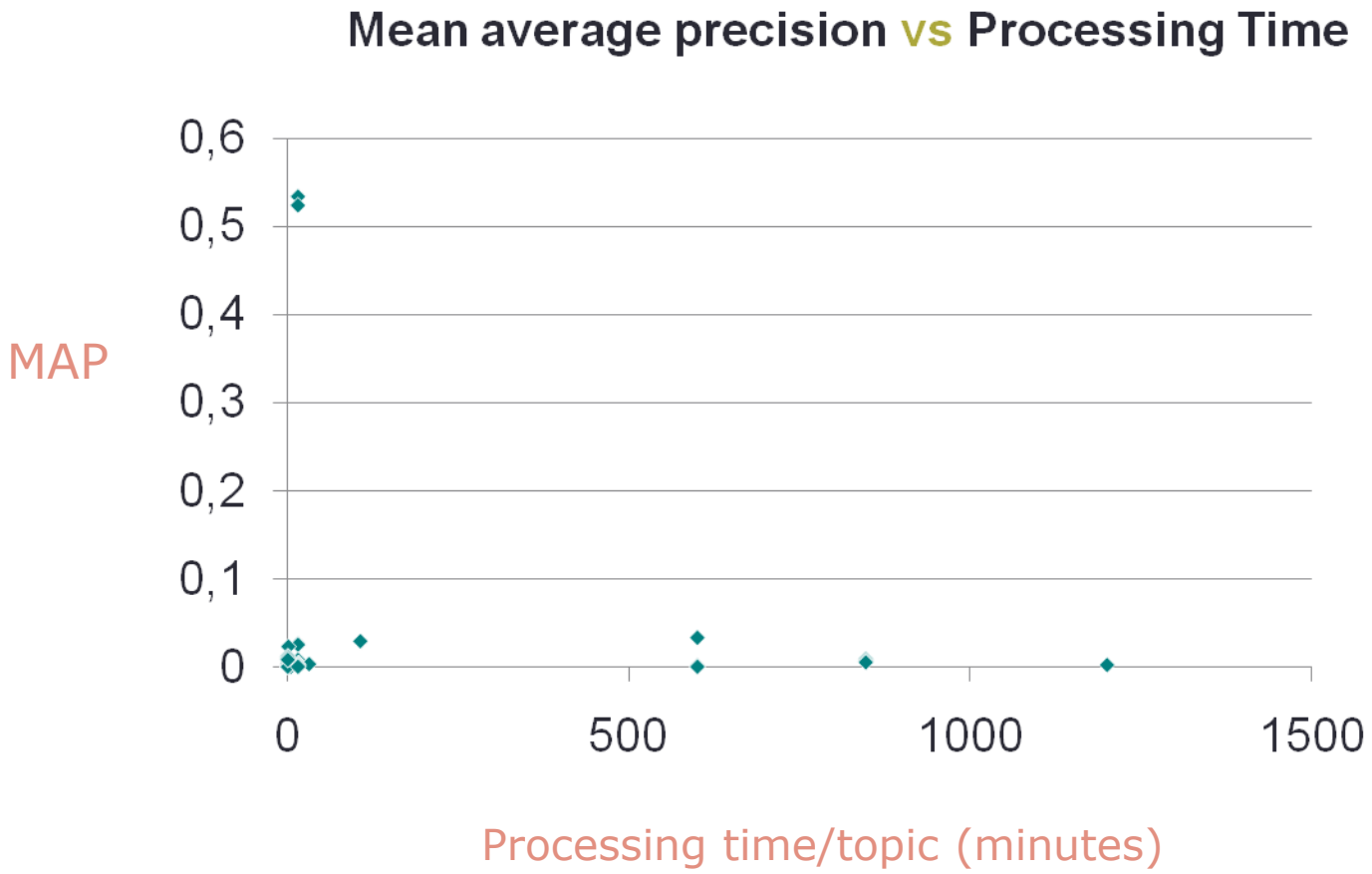
P/1 George W. Bush
P/2 George H. W. Bush
P/3 J. P. Balkenende
P/4 Bart Bosch
P/6 Prince Bernhard
P/8 Jeroen Kramer
P/11 Colin Powell
P/12 Midas Dekkers

C/5 Professor Fetze Alsvanouds
C/7 The Cook
C/9 Two old ladies, Ta en To
C/10 one of two officeworkers
C/14 Boy Zonderman

O/13 IKEA logo on clothing
O/15 black robes with white bibs
O/16 zebra stripes on pedestrian crossing
O/17 KLM Logo
O/19 Kappa Logo
O/20 Umbro Logo
O/21 tank
O/22 Willem Wever van

L/18=interior of Dutch parliament

Evaluation


Mean average precision vs Processing Time
MAP
0,6
0,5
0,4
0,3
0,2
0,1
0
0
500
1000
1500
Processing time/topic (minutes)

Overview of submissions

Beijing University of Post and Telecommunications

- Features:
  - analyzed region of interest (not full sample image)
  - Focus on face recognition
  - Additional features:
    - HSV hist, Gabor Wavelet, Edge hist, HSV correlogram, proportion B/W, body color
- Experiments
  - Compared different fusion strategies
  - for one run used a web image (9002,9003,9011,9012,9014) as sample (improved p/3)
  - top result for c/5 and c/9

Dublin City University and UPC

- segmentation into a hierarchy of regions (200 segments per frame)
- visual codebook for each topic
- search / detect
  - traverse the segment hierarchy for each test frame
  - classify each segment with SVM
  - smart pruning using aggregated feature vector of subtree
- conclusions
  - no conclusive results yet due to software bugs

Hungarian Academy of Sciences (JUMAS)

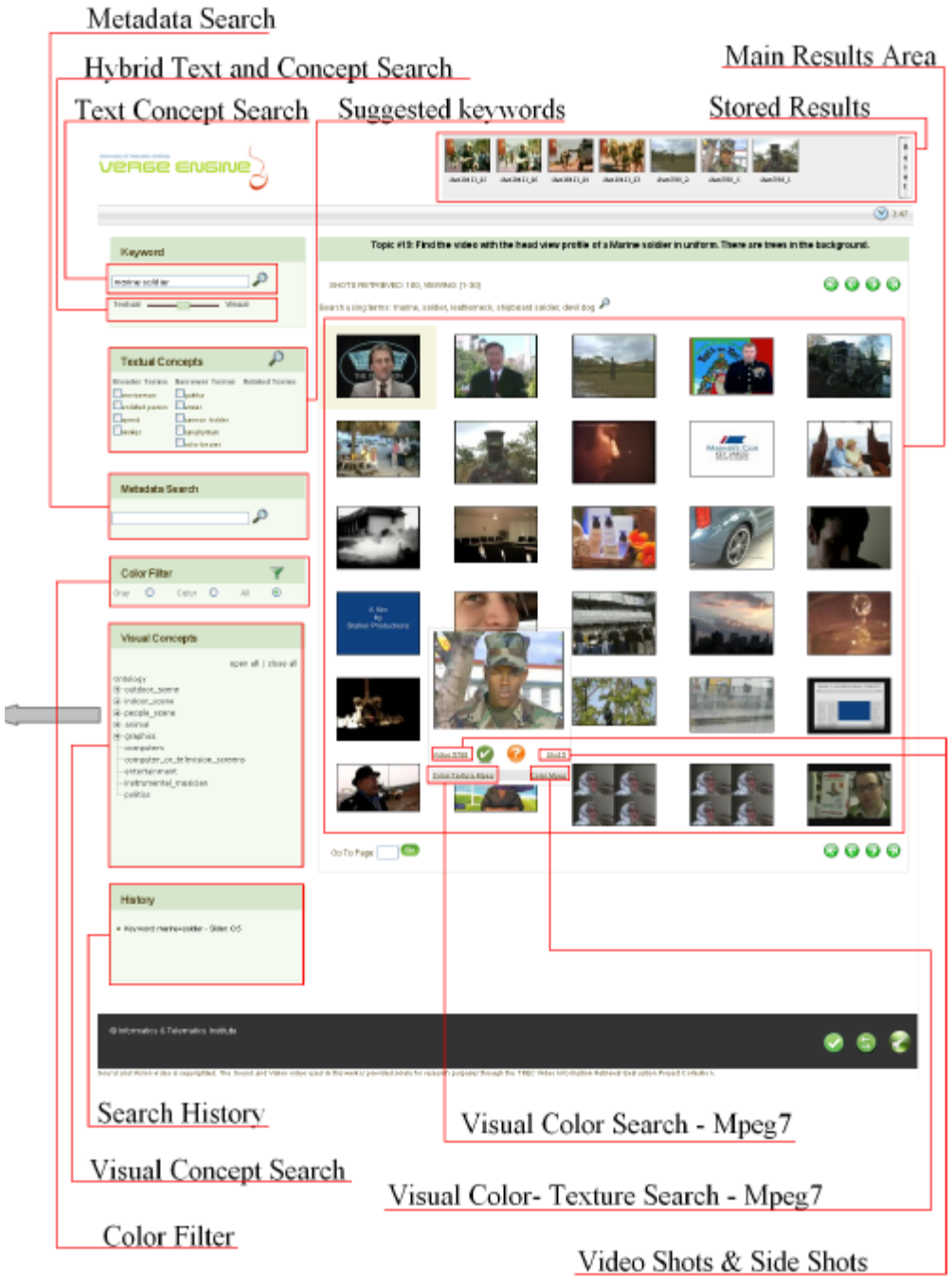
- text search in ASR transcript, no visual analysis
- top result on Balkenende and GHW Bush topics

IRIM

- Region Based Similarity Search: codebook of visual words based on image regions
- features (per cell / grid for efficiency)
  - HSV histogram (n=1000), Wavelet histogram (n=100), MPEG-7 edge histogram
- fusion: concatenation of features, matching : overlap of codewords
- conclusion:
  - HSV only performed best
  - used complete query frame (context helped on average. eg Bush @ White House and for logos on shirts)

ITI-CERTH

- interactive runs (15min)
- representation
  - full frame
  - segmentation using k-means with connectivity constraint
- features: text, HLF concepts, fusion
- conclusion
  - experiment: effect of using segmentation module
  - no significant difference, some topics improve, others decrease
  - no analysis which interaction module



Joanneum

- features
  - faces represented by Gabor wavelets
  - bag of features (mix of local desriptors)
  - SIFT
  - HOG (Histogram of Oriented Gradients)
  - region covariance descriptor
  - everything computed offline
- fusion methods
  - unweighted (max, score add)
  - weighted
- conclusions
  - single features better than fusion
  - weighting improves fusion slightly
  - top results P/1 (GWB), O/22 (van)(different runs)

KB Video Retrieval

- task generic system using 400 concept detectors (LSCOM)
- ontology based concept expansion
- low level features: edge, color, texture, local descriptors
- conclusion
  - intended as a baseline run for the INS task
  - top result George HW Bush

University of Amsterdam

- approach: INS ~ concept detection
  - SIFT and RGB SIFT visual descriptors
- SVM classifier trained on positive and 50 random dissimilar frames
- conclusion: approach is less suited for person and character queries, competitive for object and location queries

National Institute of Informatics

- features
  - local descriptor around facial points
  - local descriptor : 2 codebooks of quantized SIFT descriptors (2048/16384)
  - global: color histogram
- fusion methods
  - linear interpolation of normalized scores
- conclusions
  - face specific features make a difference
  - top result for Colin Powell and Bart Bosch

TNO

- features
  - codebook of 4096 clusters of SURF keypoints (BoF) (sparse sampling)
  - codebook of 512 clusters of SURF keybpoints (based on all query samples)
  - COTS face detector
  - segmented query was used
- conclusions
  - face detector ineffective
  - query specific codebook more effective
  - maybe filter out subtitles (cover large part of codebook)
  - top result p/12

5 more participants:

- **Notebook papers not yet available**
  - AT&T Labs – Research
  - Nanjing University
  - NTT Communication Science Laboratories-NII
  - Tokushima University
  - Xi'an Jiaotong University

Observations (1)

- Task was very hard
- Resolution of sample region of interest is important
- No clear idea yet what is the best strategy for segmenting frames, types of features, use of context, how to select codebook etc.
- Need error analysis of spiky results, why do systems score mostly zeroes for the majority of topics?
- Efficiency has not been the focus for some of the systems

Observations (2)

- Several sites would like more training examples, why not extract more frames from video (tracking...)
- Not clear how much was gained from context (no specific contrastive runs reported), some sites report that context helped
- Just like early HLF years, type specific approaches seem to have some advantage

Questions / Remarks

- Should we allow external training/sample data?
  - then we move in the direction of HLF
  - different run conditions
- Are there advantages to model the task in a detection framework?
  - in some use case video collections, no natural boundaries exist
  - we could use existing metrics for detection accuracy, detection cost etc.
- Sound and Vision video linking use case:
  - "only *central* entities should be identified and linked"
- More during the panel…