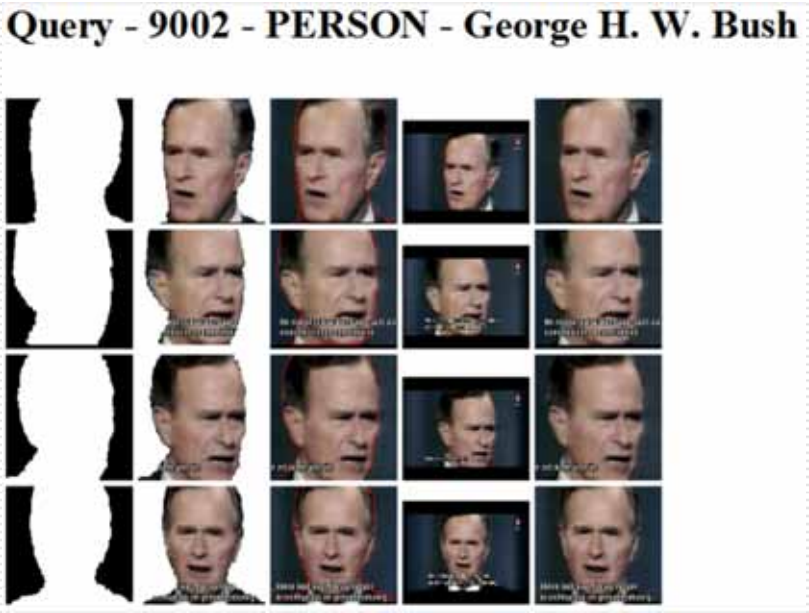
Baseline Approach for Instance Search Task: Local Region-based Face Matching and Regional Combination of Local Features

Duy-Dinh Le, Sebastien Poullot, and Shin'ichi Satoh

National Institute of Informatics, JAPAN

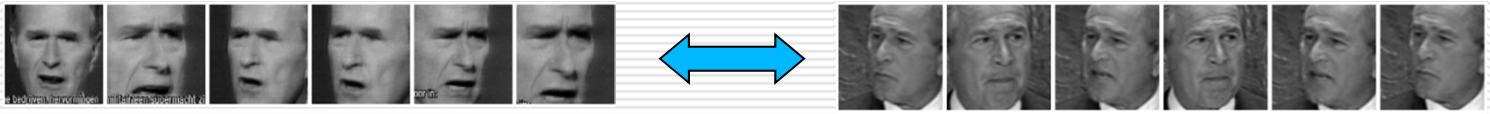
Task Overview

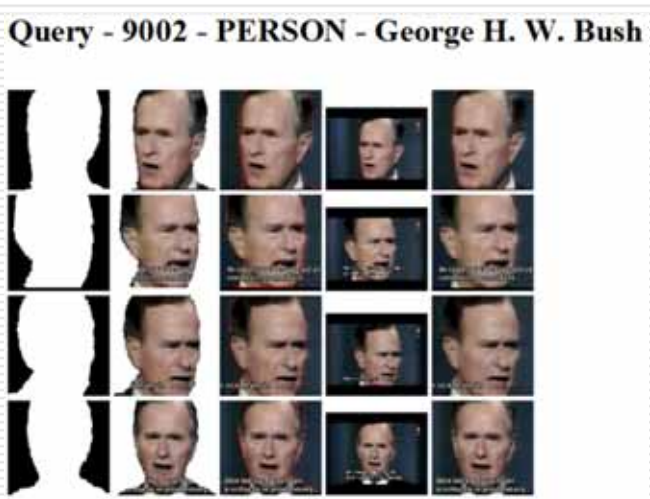
- "Given a collection of queries that delimit a person, object, or place entity in some example video, locate for each query the 1000 shots most likely to contain a recognizable instance of the entity." *(cf. TRECVID guideline)*.
- Examples for one query
  - ~5 frame images.
  - mask of an inner region of interest.
  - the inner region against a grey background.
  - the frame image with the inner region region outlined in red.
  - a list of vertices for the inner region region the target type: PERSON, CHARACTER, LOCATION, OBJECT.



Challenges – PERSON (1/2)

- Large variations in poses, sizes, facial expressions, illuminations, aging, complex background, etc.
- Examples
  - George **H.** W. Bush vs George W. Bush.



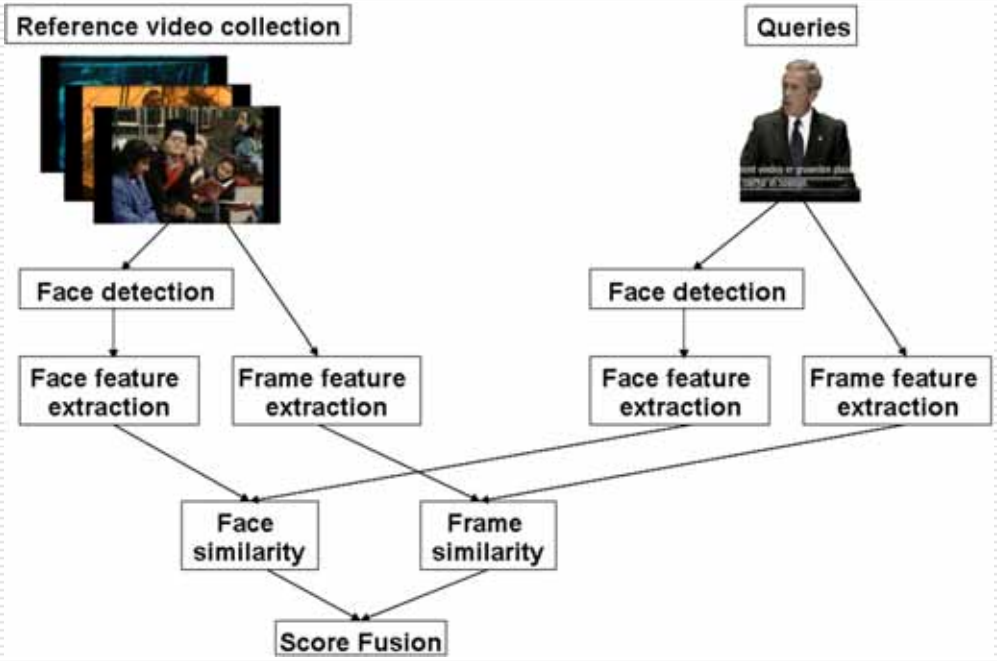


Challenges – OBJECT (2/2)

- Large variations in orientations, sizes, deformations, etc.
- Examples

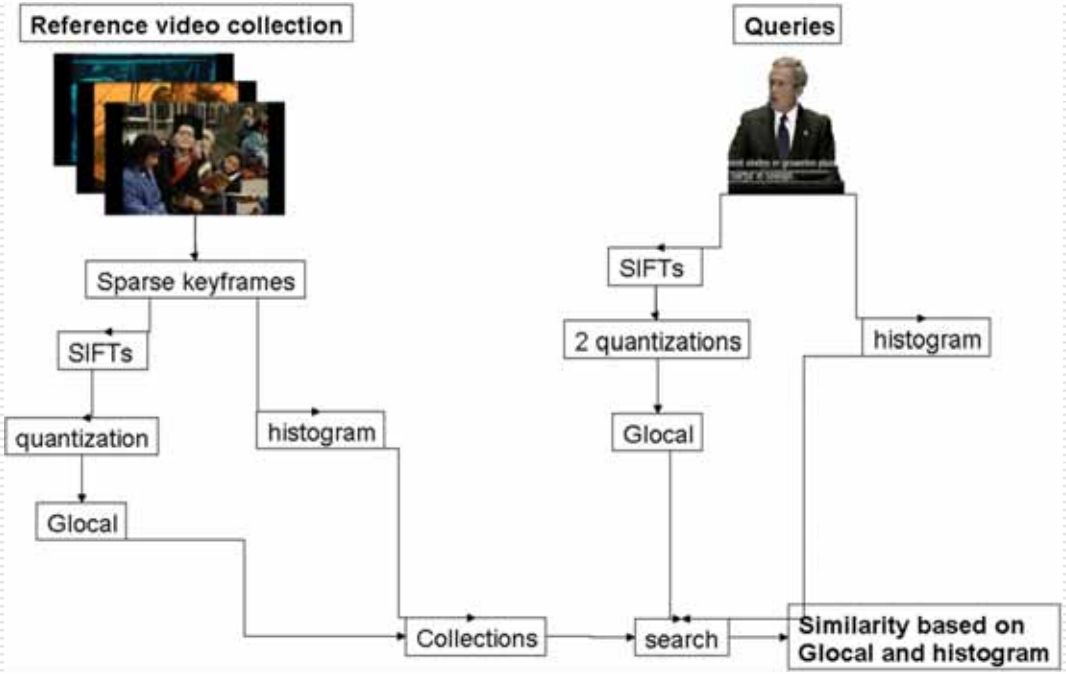
Baseline Approach – Overview (1/2)

- System 1:
  - Different treatments for different query types: PERSON, CHARACTER vs OBJECT, LOCATION.
  - Face representation: local region-based feature.
  - Frame representation: SIN task features → global + local features.



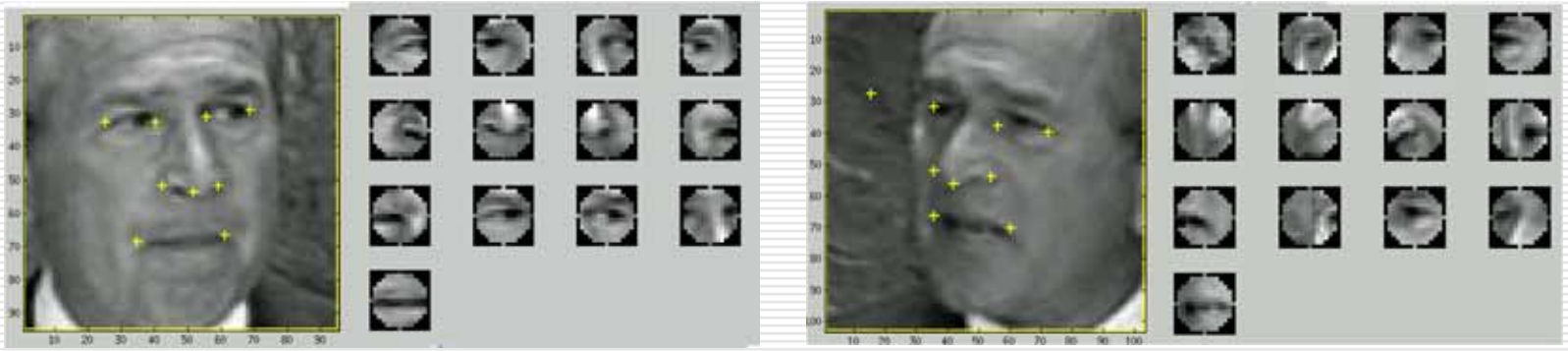
Baseline Approach – Overview (2/2)

- System 2:
  - General treatment for all queries.
  - Focus on the mask of query examples.
  - Region representation: CCD task features: regional combination of local features.



Feature Representation – System 1 (1/2)

- Face feature
  - Frontal faces are detected by NII's face detector (similar to Viola-Jones face detector).
  - Pixel intensity inside 15x15 circular regions corresponding to 13 facial points (9 facial feature points are detected, 4 more facial feature points [1] are inferred from these 9 points) → 13x149 = 1,937 dimensions. (using code provided by VGG – Oxford, UK) [2].
  - Local binary patterns feature extracted from 5x5 grid, 30 bins → 5x5x30 = 750 dimensions.

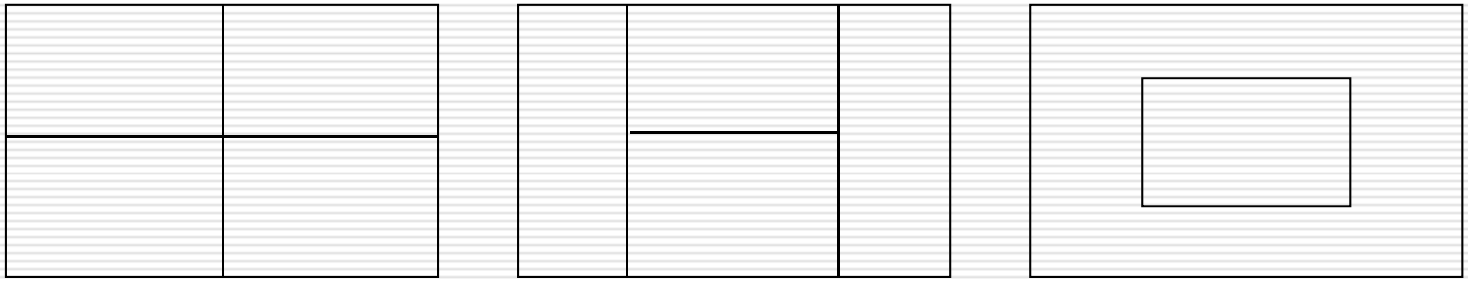


(1) the centers of the eyes, a point between the eyes, and the center of the mouth.

(2) http://www.robots.ox.ac.uk/~vgg/research/nface/

Feature Representation – System 1 (2/2)

- Global feature – SIN task
  - Color moments: 5x5 grid, HSV space → 5x5x3x3 = 225 dimensions.
  - Local binary patterns: 5x5 grid, 30 bins → 5x5x30 = 750 dimensions.
- Local feature
  - 10 predefined regions.
  - BoW of SIFT descriptors extracted from keypoints detected by HARHES keypoint detector.
  - 738 words x 10 regions = 7,380 dims.



Retrieval Strategy – System 1

- For PERSON queries, extract frontal faces and face descriptors.
- Extract frame descriptors for all query examples and keyframes in the reference database (50 keyframes/shot).
- Compute similarity between query examples and keyframes using the face descriptors and the frame descriptors. The similarities are
  - L1, L2 for the face descriptors and the global features.
  - HIK for the local feature.
  - No indexing technique is used to boost the speed.
- Compute the similarity score for one query and one shot
  - Pick the minimum score among pairs between query examples and the keyframes of the input shot.
- Fusion the scores of face descriptors and frame descriptors
  - Normalize scores using sigmoid function.
  - Linear combination of weighted scores
    - Very high weight for the face descriptor: $w_{face} = 300$. → Focus on FACE.
    - Low weight for the frame descriptors: $w_{frame\_i} = 1$.

Feature Representation – System 2

- Query
  - Focus on mask of query examples.
  - Extract Sift(DoG) features and synthesis Glocal features on a 2048 words vocabulary.
  - Take normalized RGB histogram of the area.
    → 2 descriptors for each query example.
- Reference database
  - Extract low rate KF (0.4 per second).
  - Extract Sift(DoG) features and synthesis Glocal features on a 2048 words vocabulary.
  - Take normalized RGB histogram of the area.
    → 2 descriptors for each keyframe.

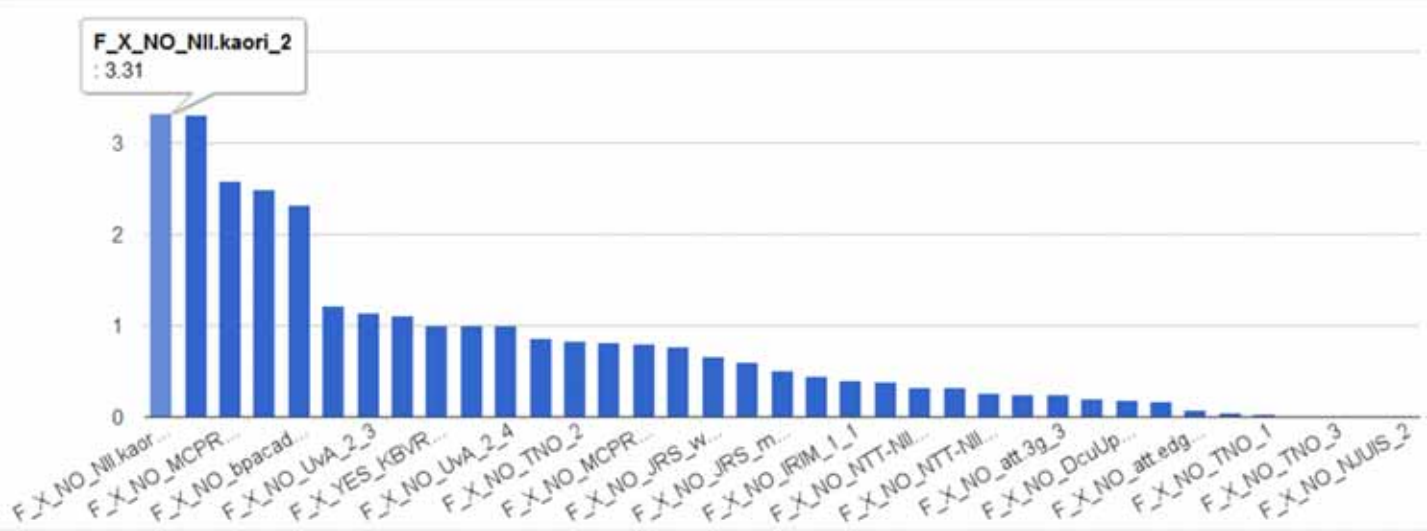
Retrieval Strategy – System 2

- Compute similarity between query example descriptors and keyframe ones. The similarities are
  - Dice coefficient for Glocal.
  - L1 for RGB histograms.
- Simply added together for 1 query example.
- All similarity scores of the query examples are added for each keyframe.

Results(*) – System 1 (1/2)

- L1 is the most suitable choice for similarity measure.
- Good face feature brings good result.

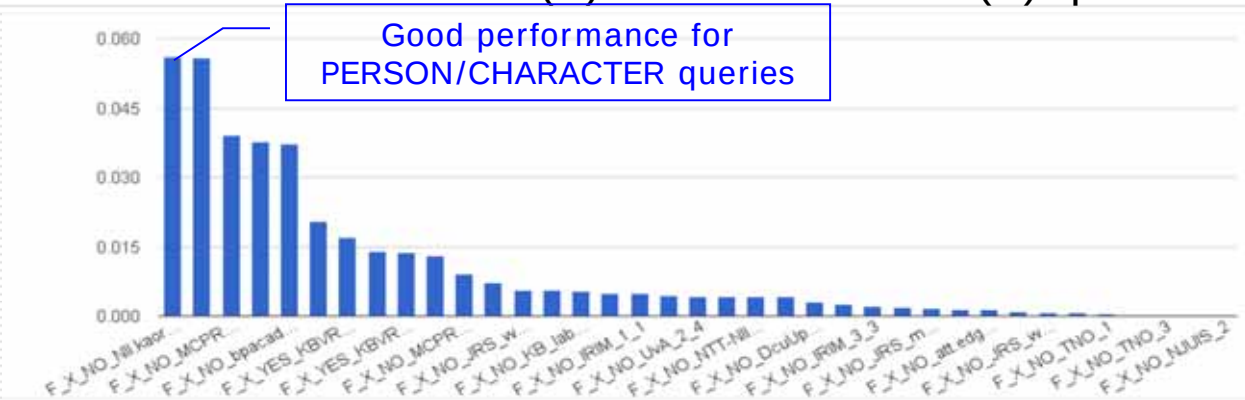
	L1	L2	HIK
Face.vgg	2.87	2.14	
Face.lbp	0.04	0.03	0.04
Local.harhes			0.1
Global.cm	0.24	0.23	
Global.lbp	0.07	0.04	0.08
Fusion all	3.31		



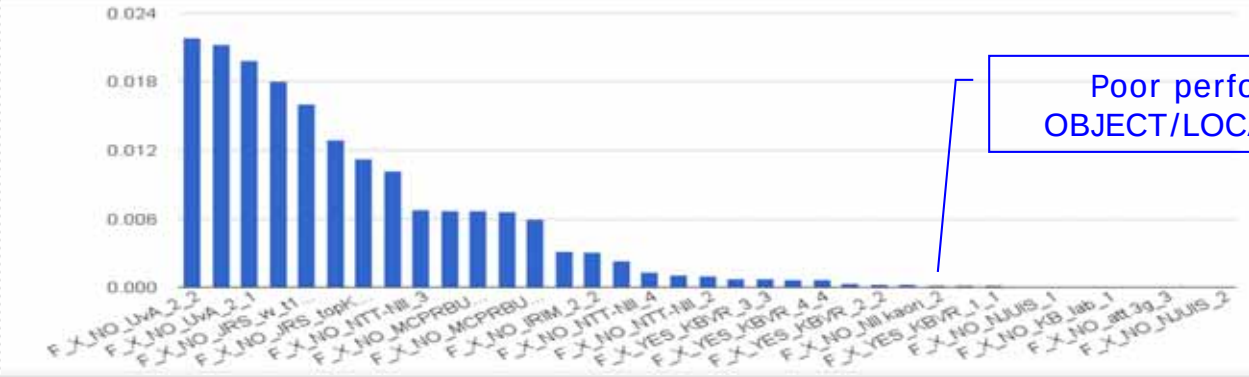
(*) http://satoh-lab.ex.nii.ac.jp/users/ledduy/nii-trecvid/ins-tv10/ins-tv10.php → view query examples, groundtruth, and ranked lists.

Results – System 1 (2/2)

- Performance for PERSON(8) and CHARACTER(5) queries → 13 queries.



- Performance for OBJECT(8) and LOCATION(1) queries → 9 queries.



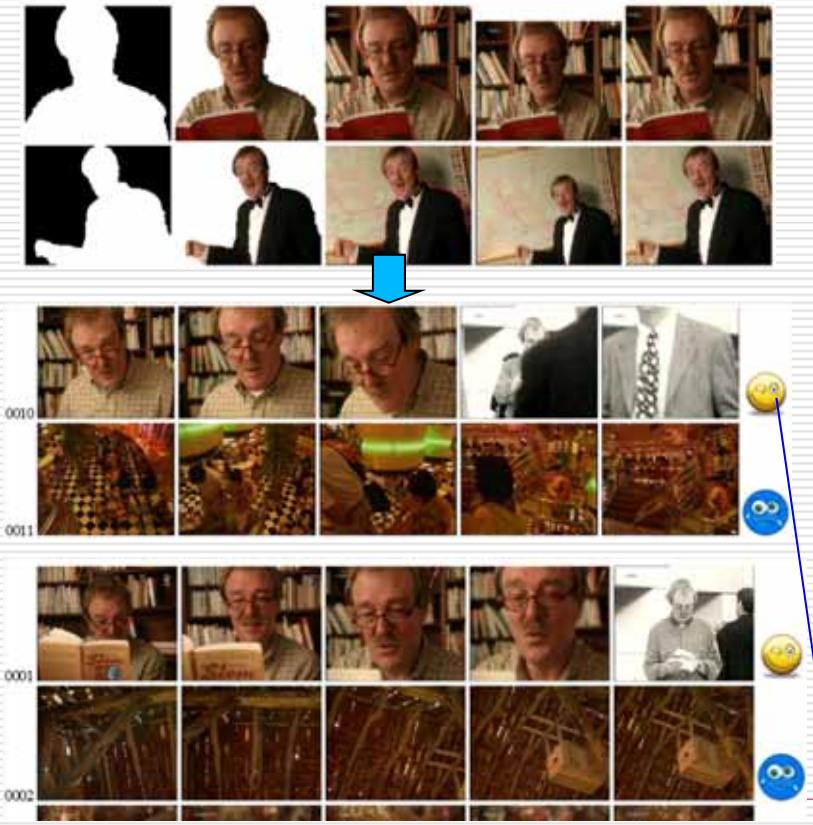
Some Results – System 1

- System-1: Fusion helps to improve the performance
- Only face descriptor:8 - 15 - 18 - 20
- Fusion: 7 - 11 - 17 - 19

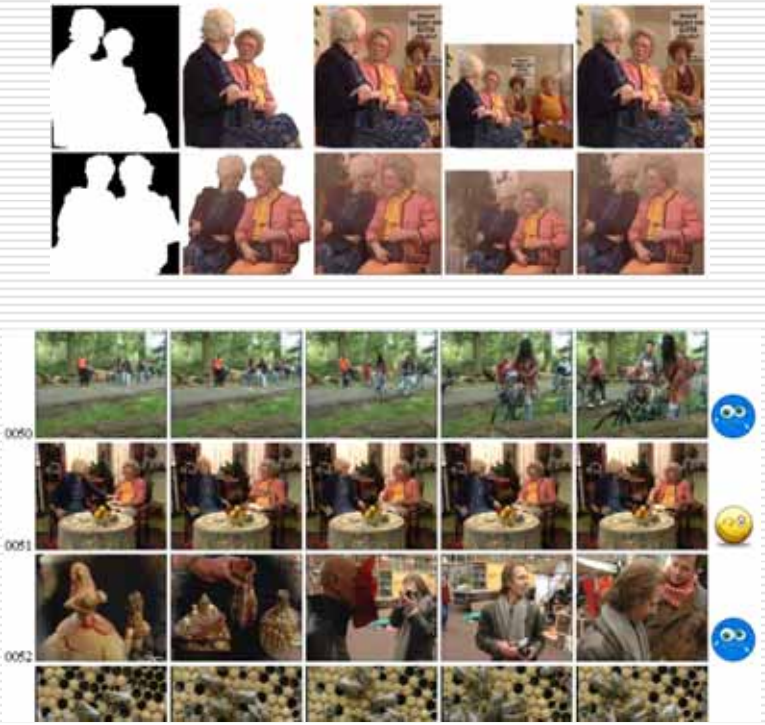
Some Results – System 1

- **Color moments feature** → good performance for PERSON queries

Query - 9012 - PERSON - Midas Dekkers.



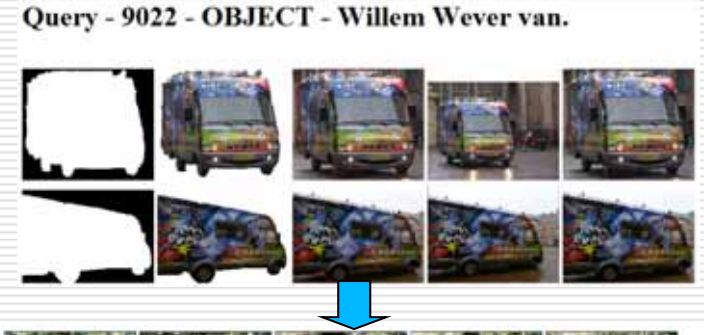
Query - 9009 - CHARACTER - Two old ladies, Ta en To. List.



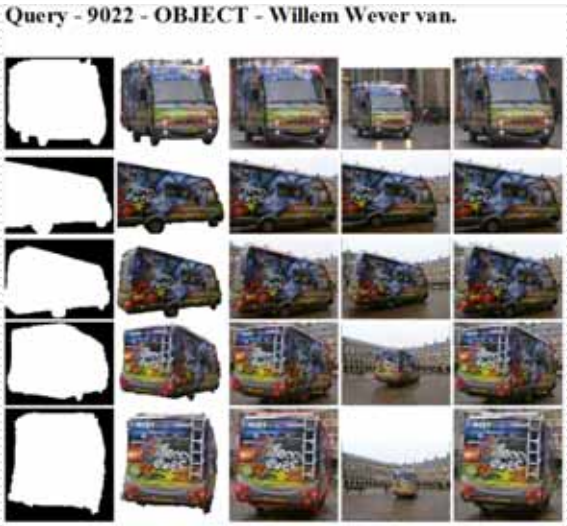
Rank 1, and 10

Some Results – System 1

- Local feature → **HIK might not be suitable** similarity measure since it is easy to bias in favor of images with complex texture.



Some Results – System 2



Willem(query22) rank 240
Glocal only

Willem(query22) rank 48
Glocal only

Some Results – System 2

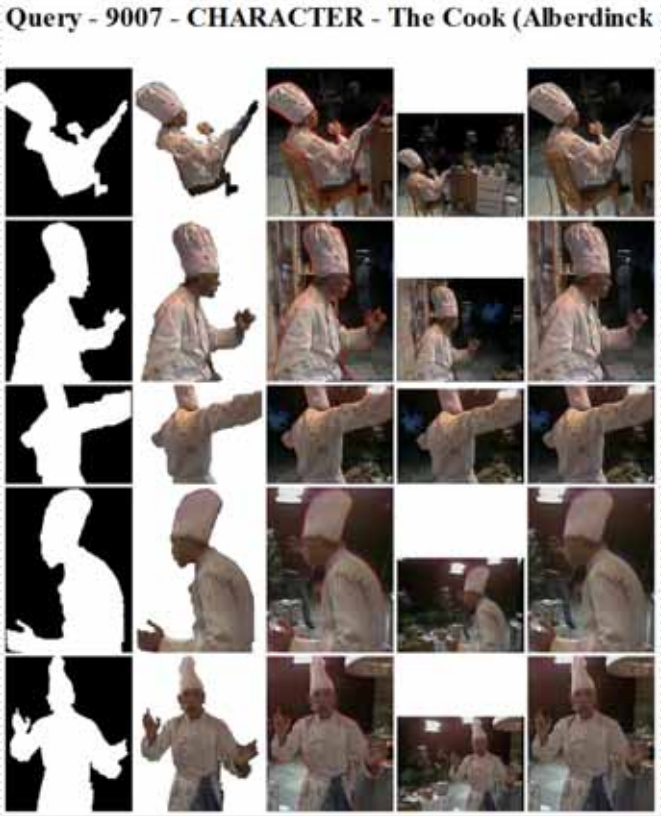
Query12
Glocal+RBG

Rank 14

Rank 62

Rank 92

Some Results – System 2



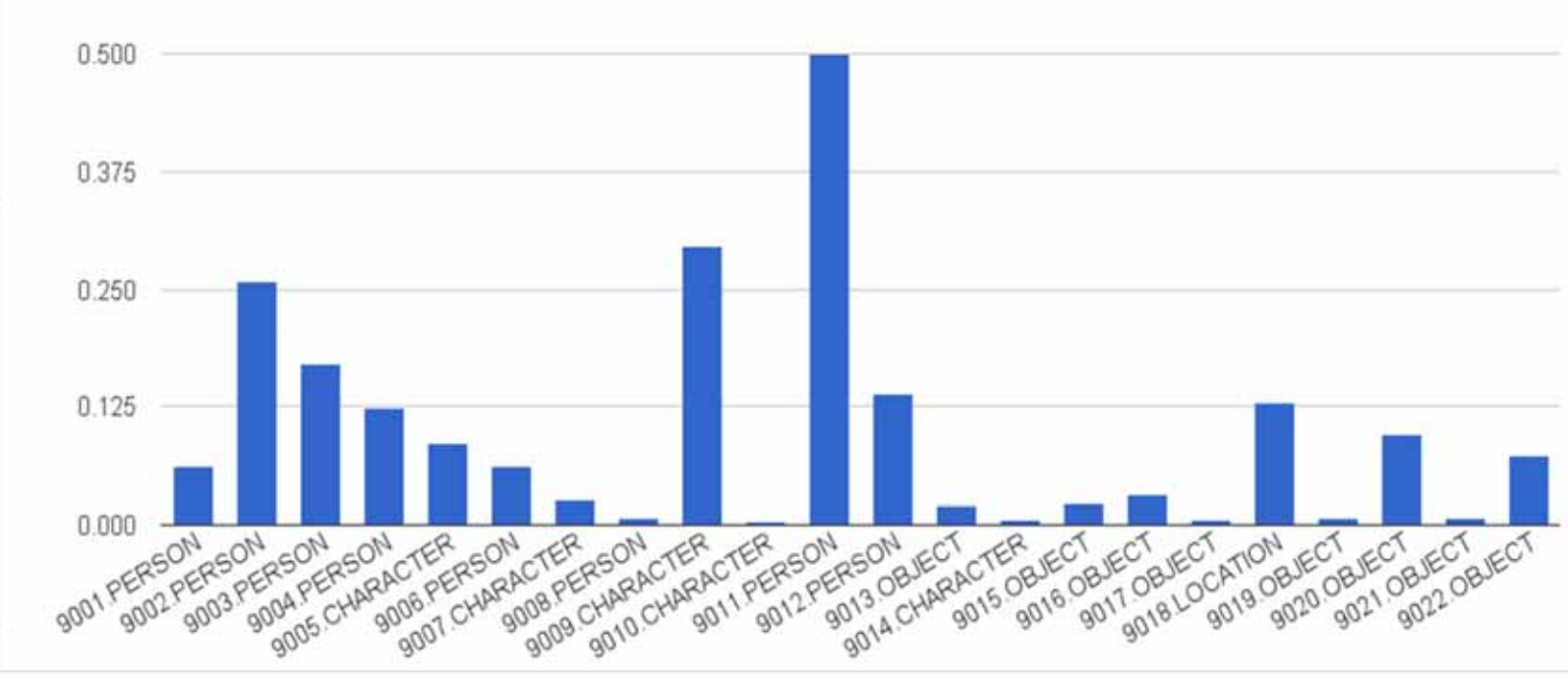
Query 9007
Glocal+RGB

Rank 86

Rank 123

Discussions

- For PERSON and CHARACTER queries, the (max) performance is usually high.
- Current face matching technique only handles frontal faces. More efforts should be made to handle multi-view faces.



Discussions - 1

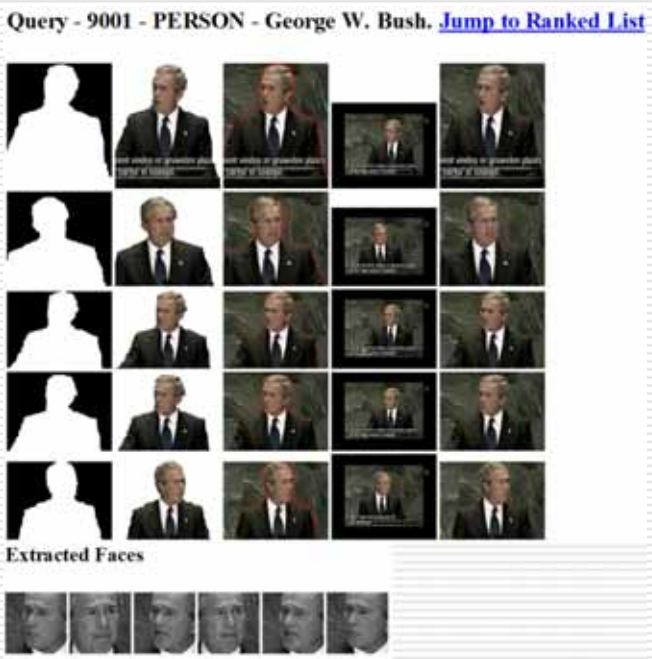
- Fusion of different features for different object types helps to improve the performance. However, how to efficiently fuse is questionable. Our approach is quite ad-hoc.
- Appropriate similarity measure should be carefully selected.
- Dense sampling in keyframe extraction is an important factor.

Discussions - 2

- Bad quality of queries is damageable for local feature.
- Color moments feature is simple, but can achieve reasonable result. In some cases, it outperforms local features.
- How to deal with scale and comparison to images from reference database.

Demo – 1

- URL: http://satoh-lab.ex.nii.ac.jp/users/ledduy/nii-trecvid/ins-tv10/ins-tv10.php
- Username/password: trecvid/niitrec.
- Functions: view query examples, ground truth, and ranked lists of runs.



Demo - 2

- Result page

Irrelevant

Relevant

Thank you and Question