TRECVID 2016 Workshop

WHU-NERCMS at TRECVID2016: Instance Search Task

Z. Wang, Y. Yang, S. Guan, C. Han, J. Lan, R. Shao, J. Wang, C. Liang

National Engineering Research Center for Multimedia Software, School of Computer,
Wuhan University
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences

November 14, 2016 NIST

Outline

1

Introduction
Problem and Motivation

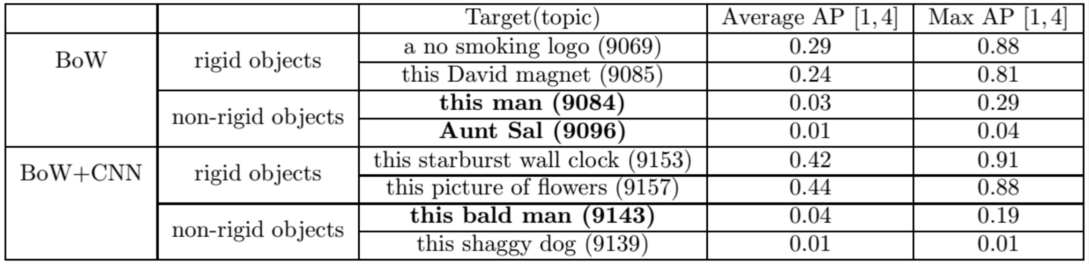
Proposed Approach
Framework and Details

Results
4 runs

Conclusion

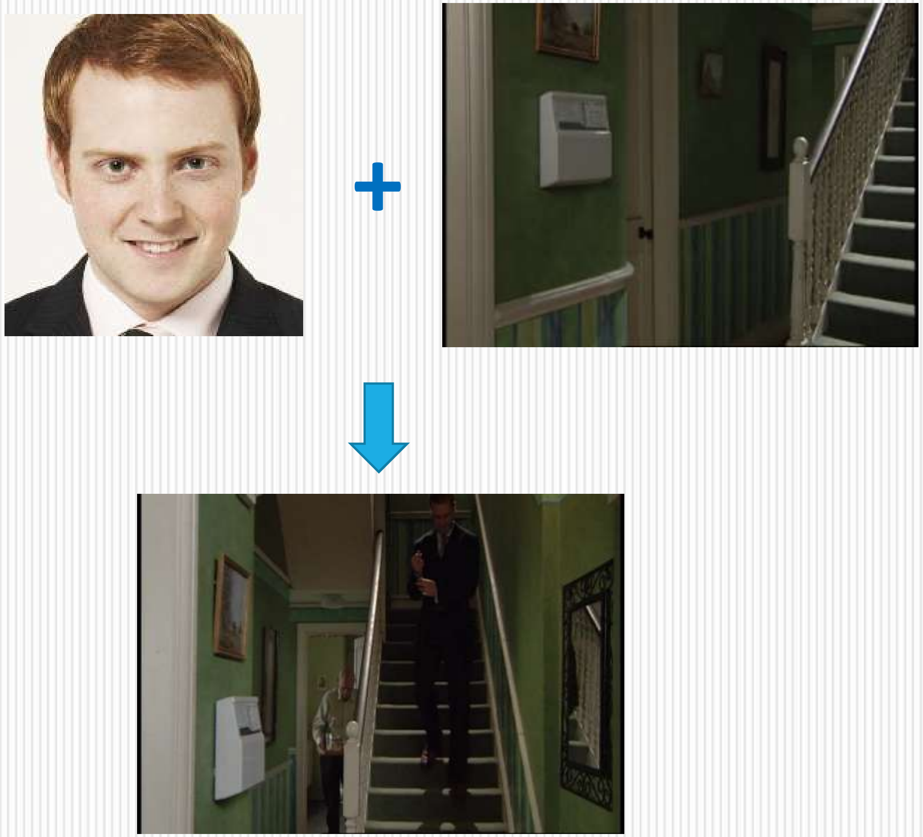
1 Introduction

Previous topics



		Target(topic)	Average AP [1, 4]	Max AP [1, 4]
BoW	rigid objects	a no smoking logo (9069)	0.29	0.88
		this David magnet (9085)	0.24	0.81
	non-rigid objects	**this man (9084)**	0.03	0.29
		Aunt Sal (9096)	0.01	0.04
BoW+CNN	rigid objects	this starburst wall clock (9153)	0.42	0.91
		this picture of flowers (9157)	0.44	0.88
	non-rigid objects	**this bald man (9143)**	0.04	0.19
		this shaggy dog (9139)	0.01	0.01

Topics in this year



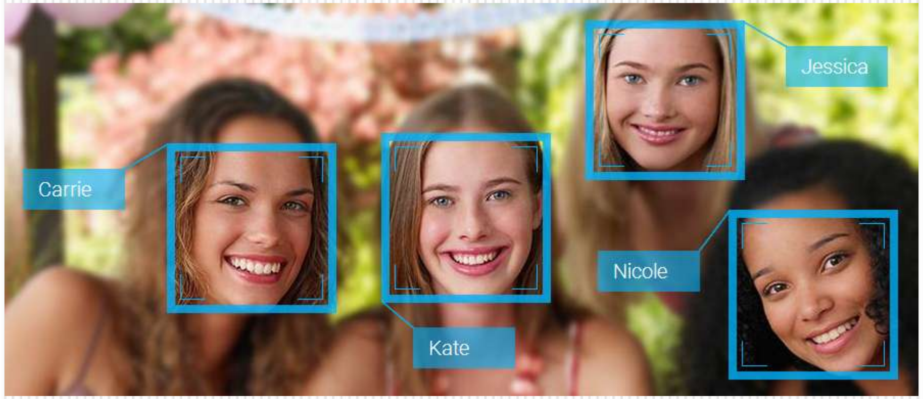
1 Introduction

How to find the specific person?

How to find the specific location?

How to fuse the person and scene results?

How to alleviate noise influence?

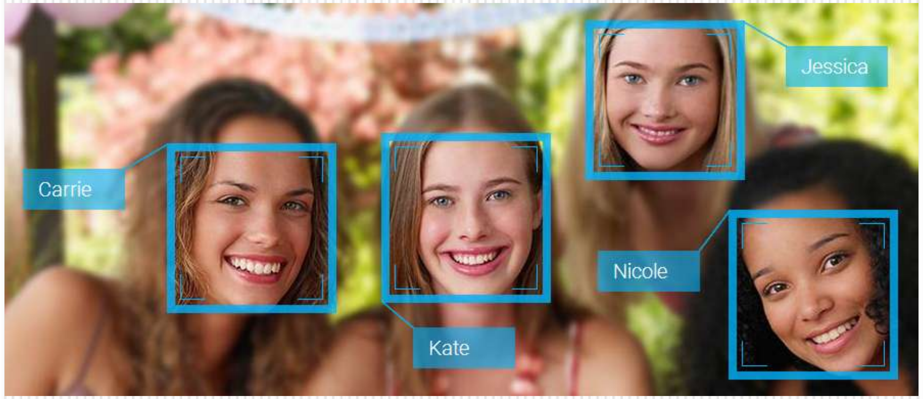


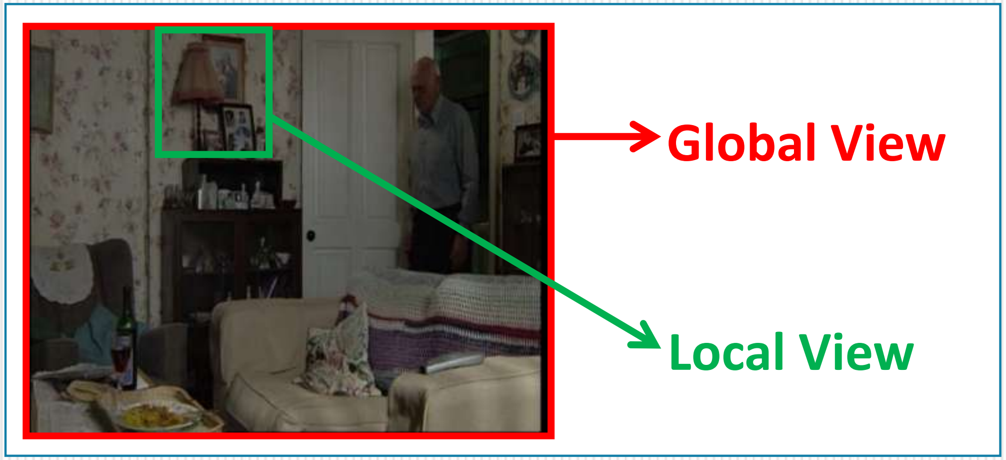
How to find the specific location?

How to fuse the person and scene results?

How to alleviate noise influence?

1 Introduction

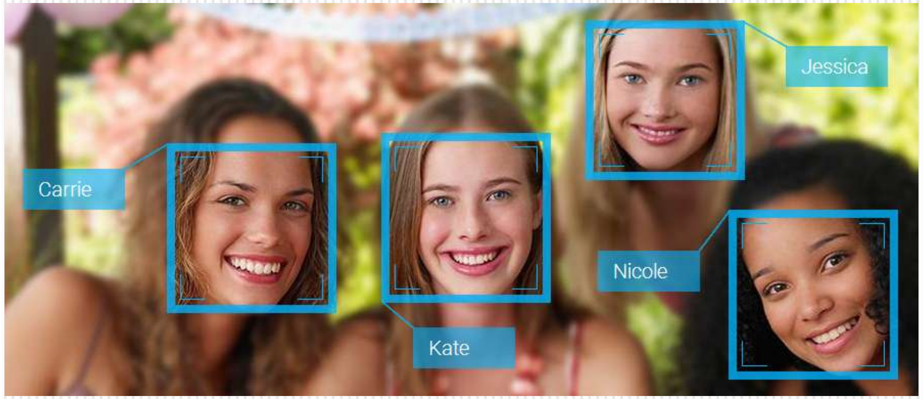


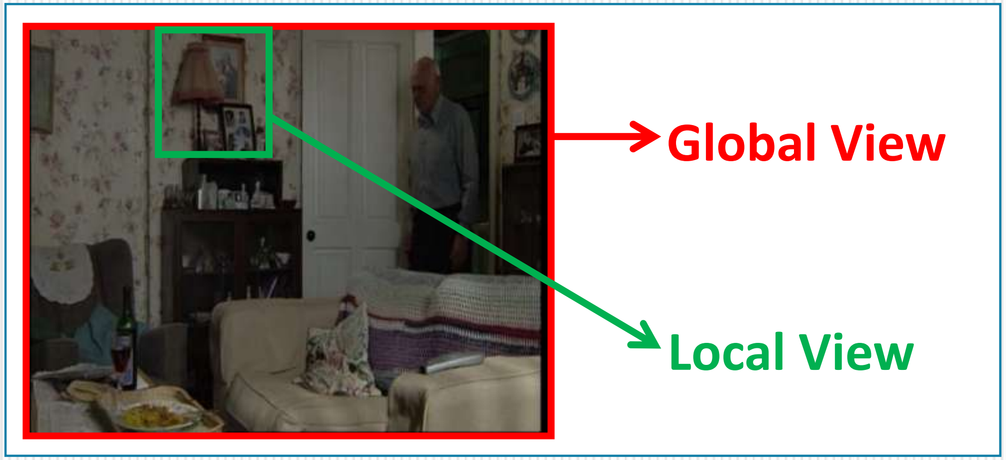


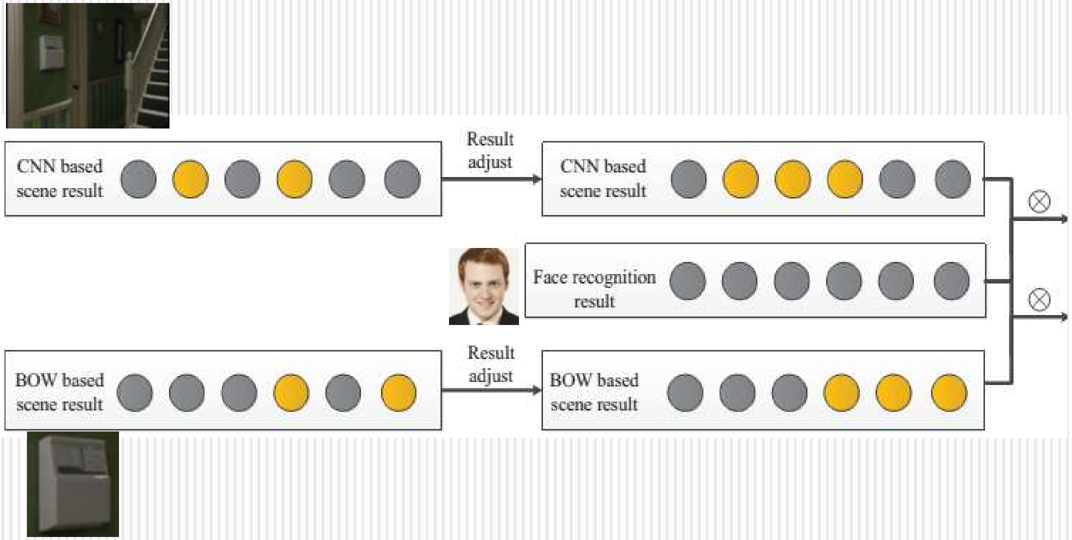
How to fuse the person and scene results?

How to alleviate noise influence?

1 Introduction

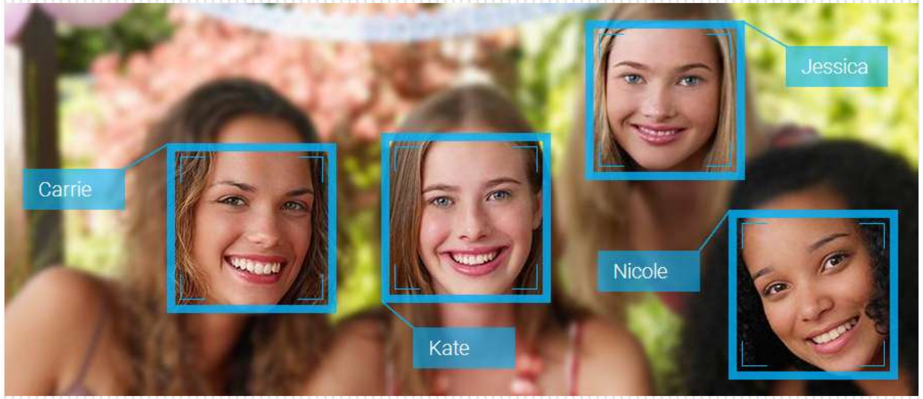


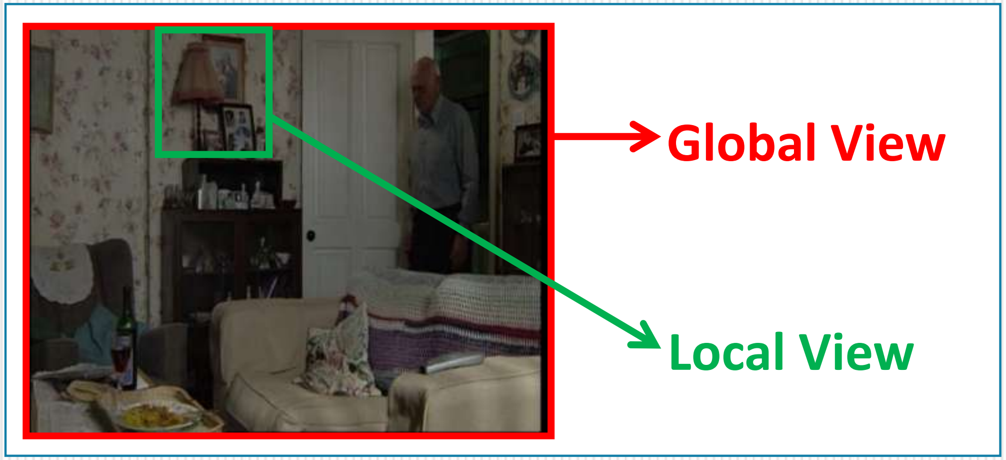




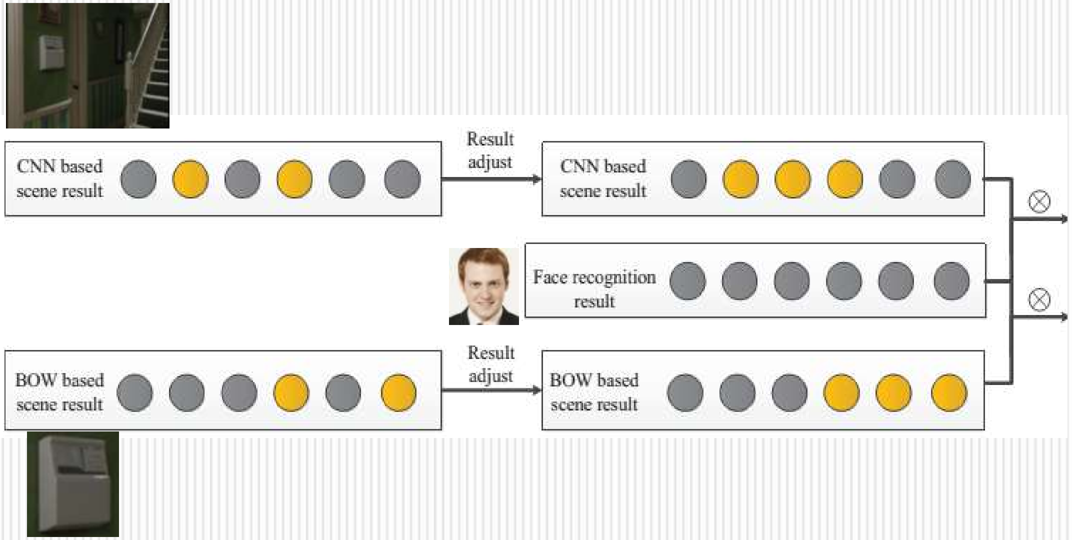
How to alleviate noise influence?

1 Introduction




Global View

Local View



Outdoor scene

Non face X

Outline

2

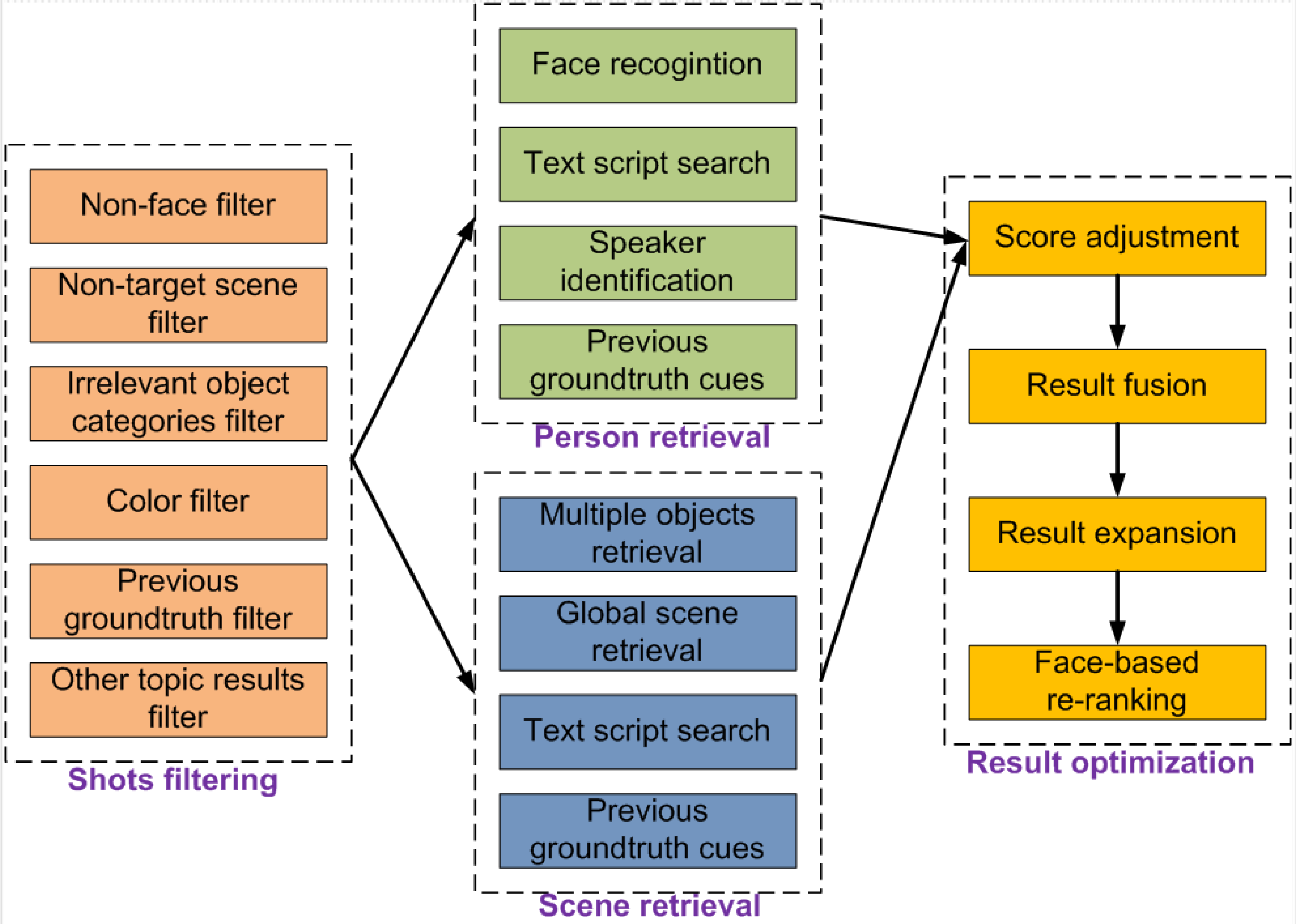
Introduction
Problem and Motivation

Proposed Approach
Framework and Details

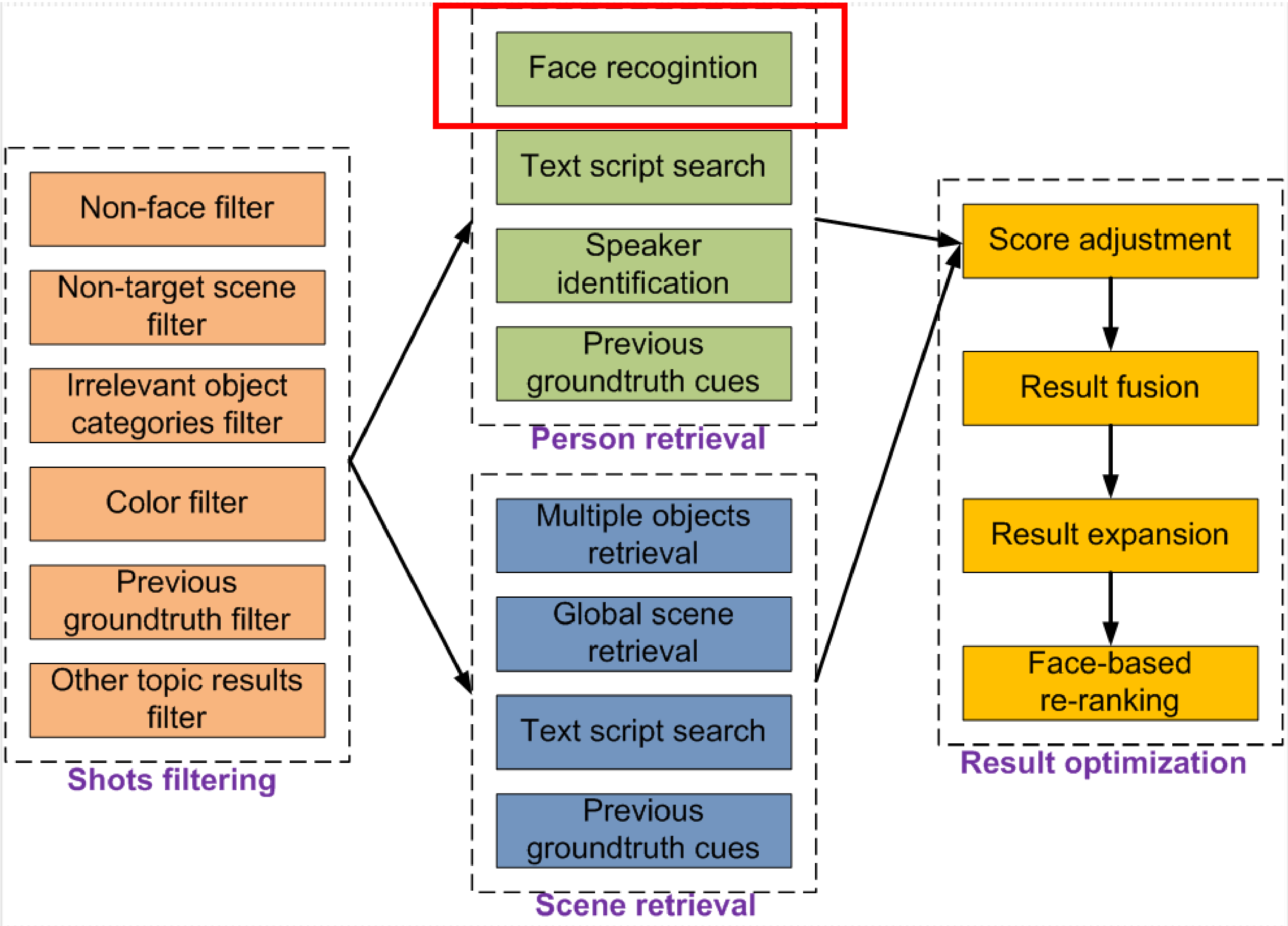
Results
4 runs

Conclusion

2 Proposed Approach



2 Proposed Approach



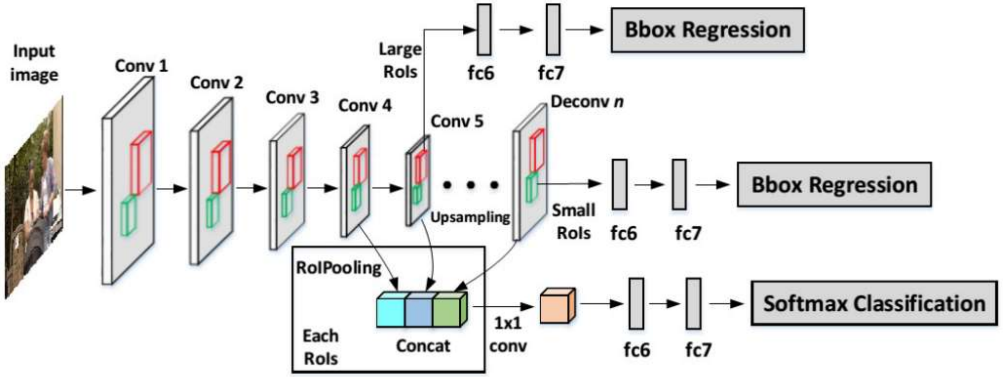
2 Proposed Approach – Face recognition

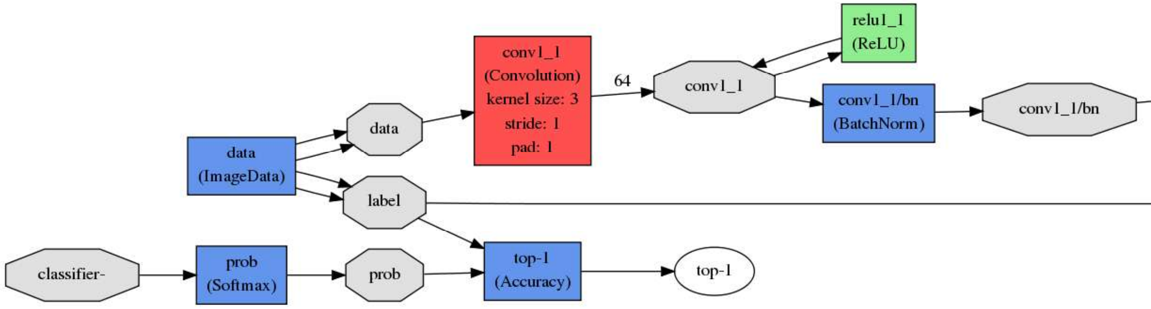
Face detection

- Scale-Adaptive Deconvolutional Regression face detection network
- Use the pretrained VGG16 model to initialize the network
- two regression layers + softmax layer

Face identification

- 9 convolutional layers, 5 pooling layers, 2 fully connected layer
- Softmax and triplet cost are combined
- Trained in our collected IVA-WebFace with 80 thousand identities and each has about 500-800 face images.



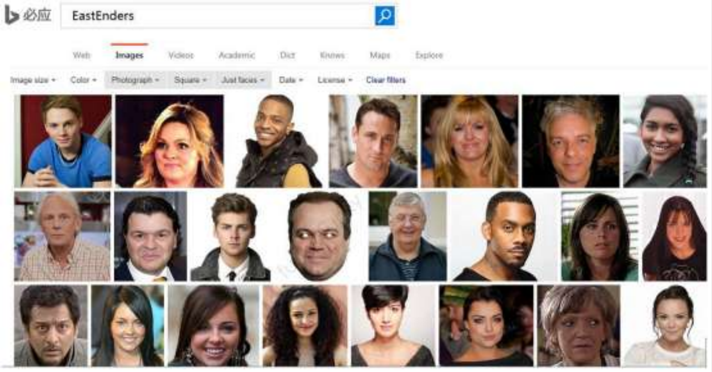


Y. Zhu, J. Wang, C. Zhao, H. Guo and H. Lu. Scale-adaptive Deconvolutional Regression Network for Pedestrian Detection, ACCV, 2016.
Haiyun Guo, et al. Multi-View 3D Object Retrieval with Deep Embedding Network, ICIP, 2016.

2 Proposed Approach – Face recognition

Face library

- Search the keyword EastEnders in Bing
- Our own face library includes 815 face images



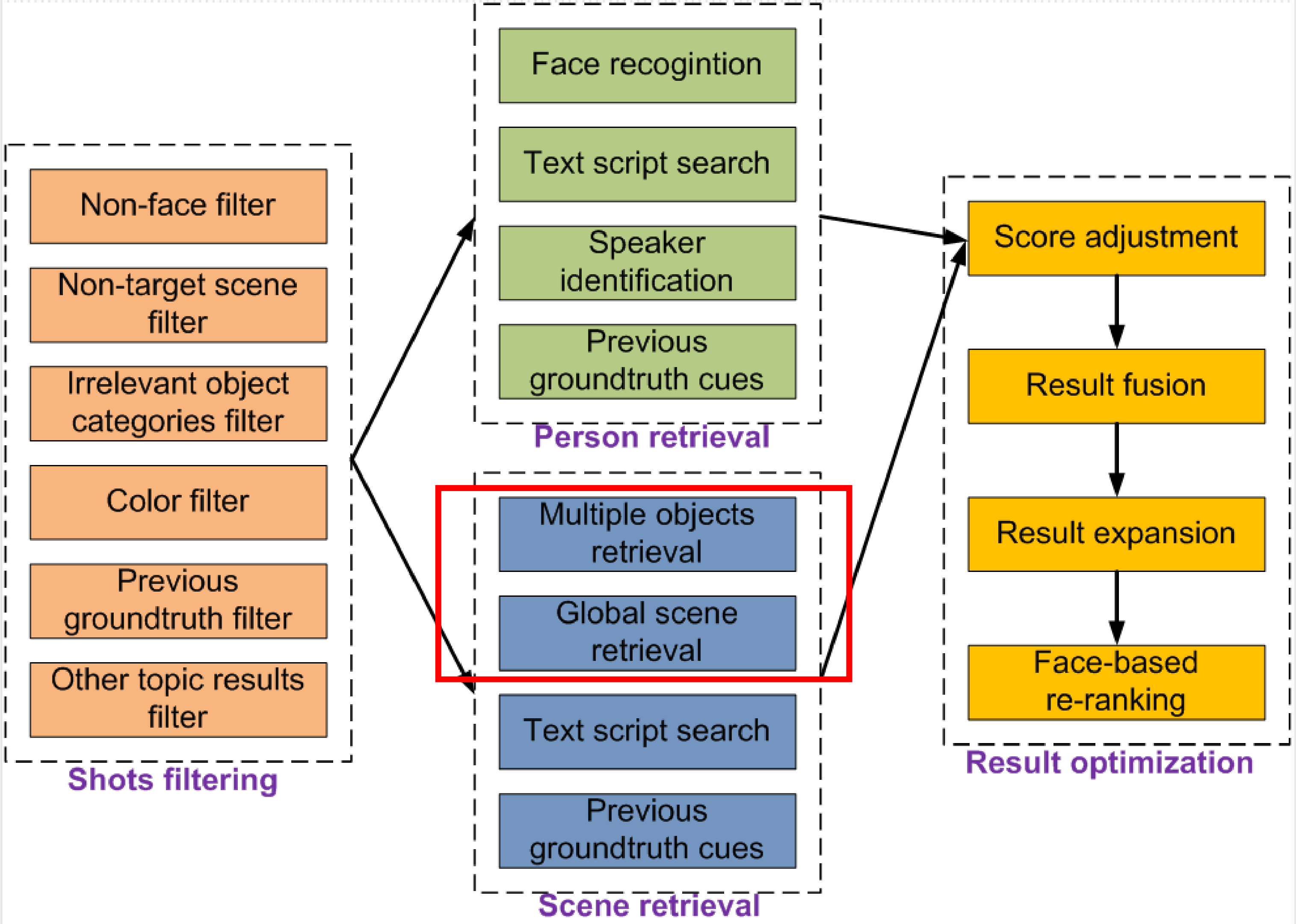
815

Name	Faces with diferent angles in the face library
Brad	
Dot	
Fatboy	
Jim	
Pat	
Patrick	
Stacy	

2 Proposed Approach – Face recognition

DEMO

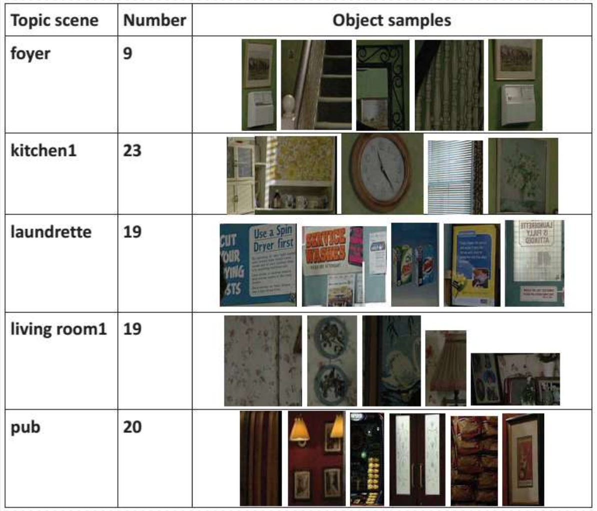
2 Proposed Approach



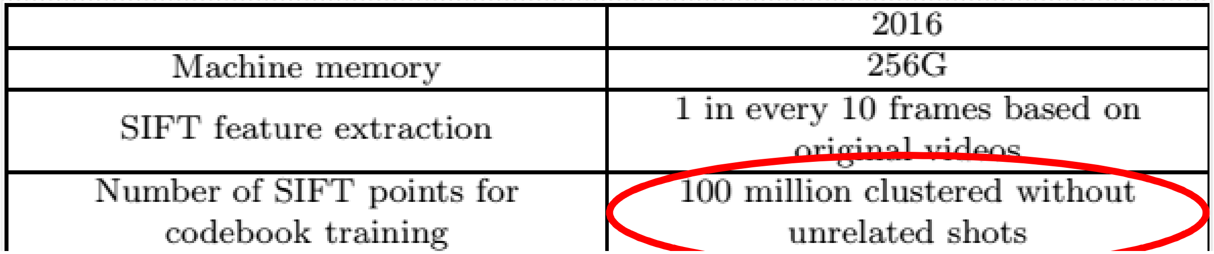
2 Proposed Approach – Local View + Global View

Multiple objects retrieval

- Through identifying typical objects in a certain topic scene, we can seek out shots of this scene indirectly



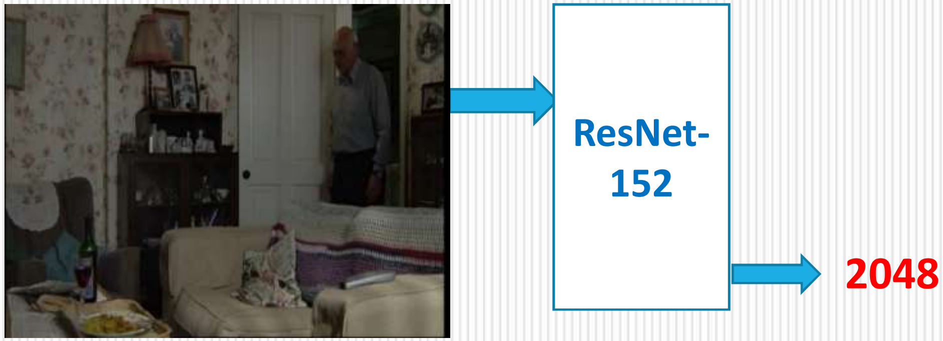
Topic scene	Number	Object samples
foyer	9	
kitchen1	23	
laundrette	19	
living room1	19	
pub	20	

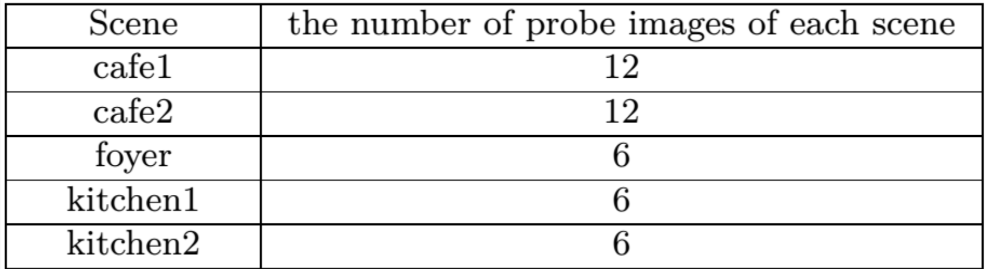


	2016
Machine memory	256G
SIFT feature extraction	1 in every 10 frames based on original videos
Number of SIFT points for codebook training	100 million clustered without unrelated shots

Global scene retrieval

- Global feature: the output of the fully connected layer
- ResNet-152 model pre-trained by Facebook AI Research



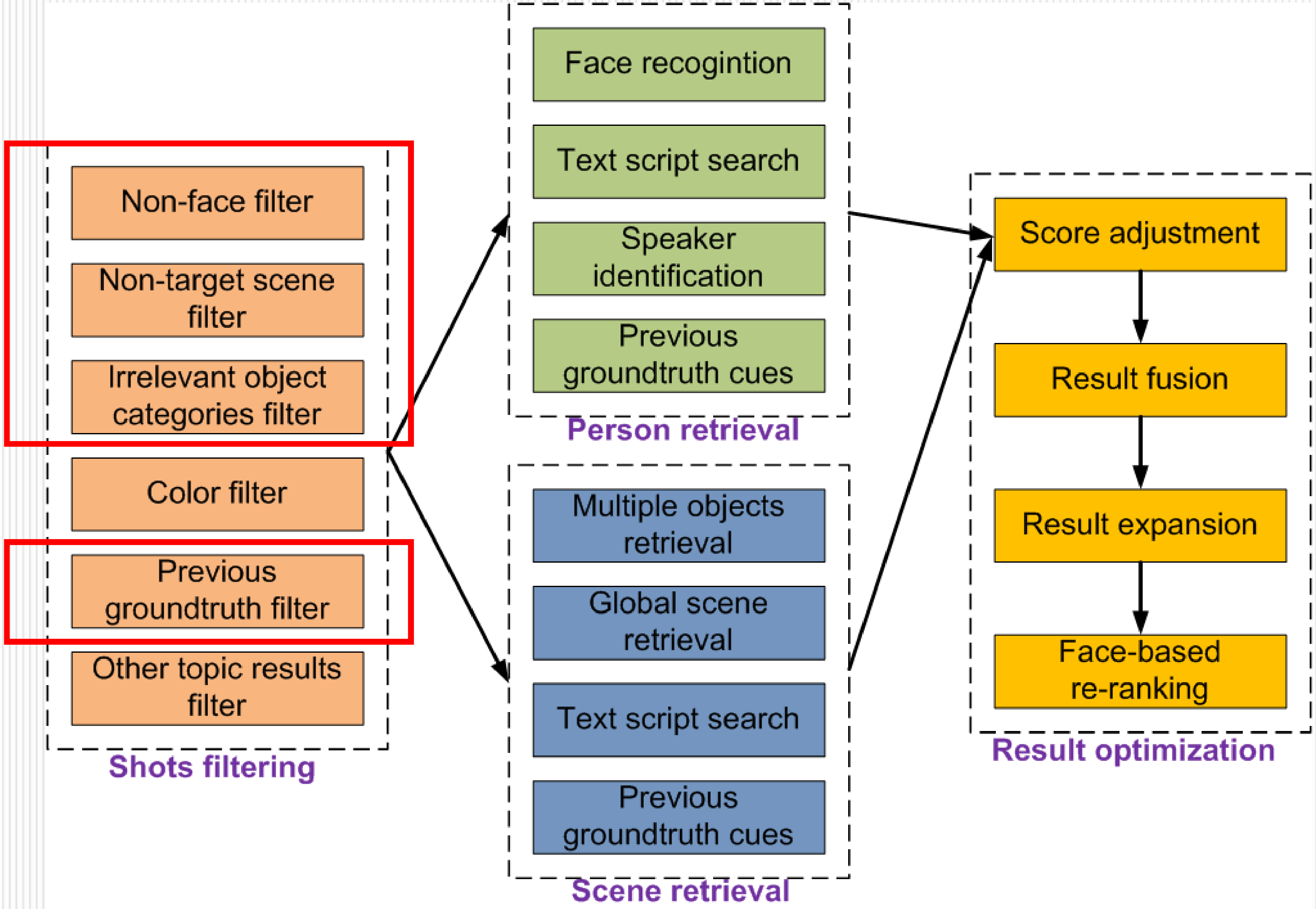


Scene	the number of probe images of each scene
cafe1	12
cafe2	12
foyer	6
kitchen1	6
kitchen2	6

2 Proposed Approach – Local View + Global View

DEMO

2 Proposed Approach



2 Proposed Approach - Filtering

Non-target face filter

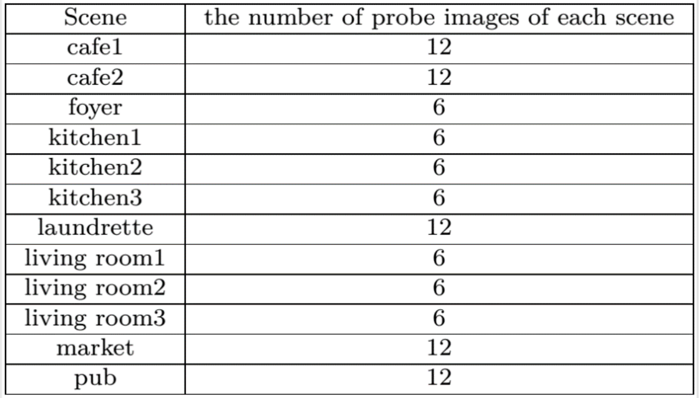
- **217,894** shots are deleted
- 851 ground truth shots deleted
- 822 of them are recovered with expanding shots
- Up to **46%** of original video shots are filtered

(a) shot209_497 (b) shot33_2216

Due to non-front and occlusion, some ground truth shots are filtered by mistake.

Non-target scene filter

- Global feature: the output of the fully connected layer
- ResNet-152 model pre-trained by Facebook AI Research
- We filter **5592** shots



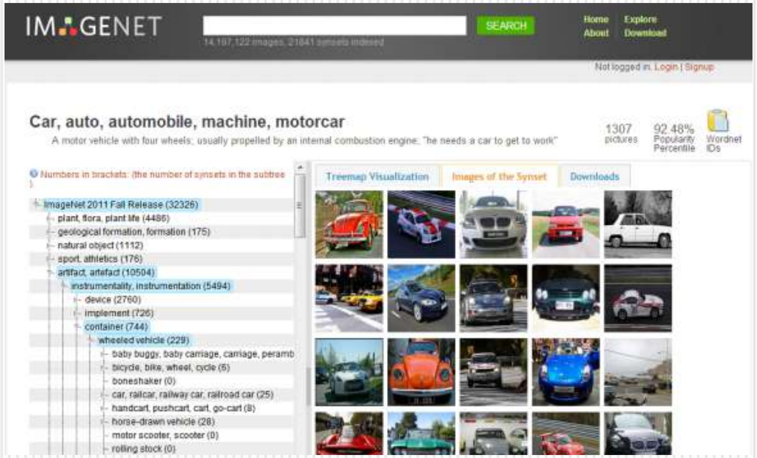
Scene	the number of probe images of each scene
cafe1	12
cafe2	12
foyer	6
kitchen1	6
kitchen2	6
kitchen3	6
laundrette	12
living room1	6
living room2	6
living room3	6
market	12
pub	12

(a) *living room3* (b) *kitchen3*

2 Proposed Approach - Filtering

Irrelevant object categories filter

- 37 categories about vehicles, such as ambulance, minibus and police van
- 52 categories only appear outdoor, such as hippopotamus, Indian elephant and castle
- We totally delete **19,244** shots



Car, auto, automobile, machine, motorcar

A motor vehicle with four wheels; usually propelled by an internal combustion engine: "he needs a car to get to work"

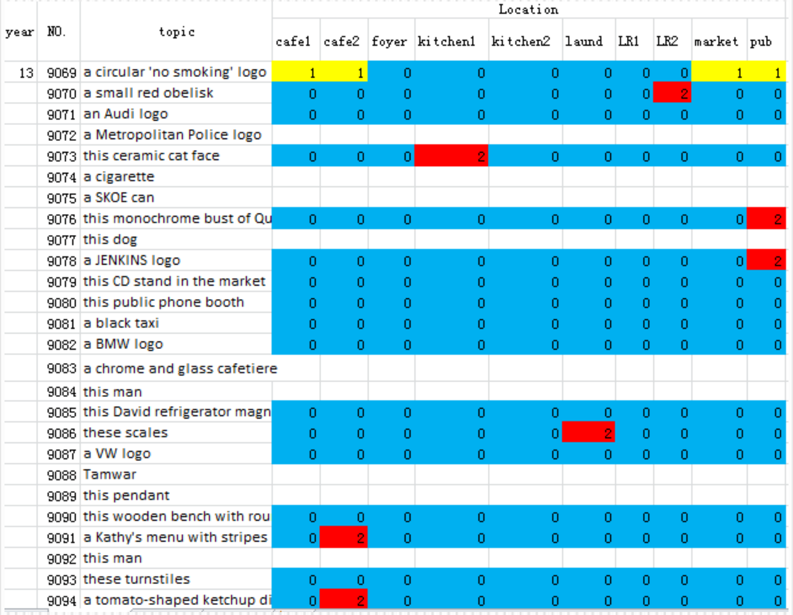
Numbers in brackets: (the number of synsets in the subtree)

- ImageNet 2011 Fall Release (32326)
  - plant, flora, plant life (4486)
  - geological formation, formation (175)
  - natural object (1112)
  - sport, athletics (176)
  - artifact, artefact (10504)
    - instrumentality, instrumentation (5494)
      - device (2760)
      - implement (726)
      - container (744)
        - wheeled vehicle (229)
          - baby buggy, baby carriage, carriage, peramb
          - bicycle, bike, wheel, cycle (6)
          - boneshaker (0)
          - car, railcar, railway car, railroad car (25)
          - handcart, pushcart, cart, go-cart (8)
          - horse-drawn vehicle (28)
          - motor scooter, scooter (0)
          - rolling stock (0)

http://imagenet.stanford.edu/synset?wnid=n03417042

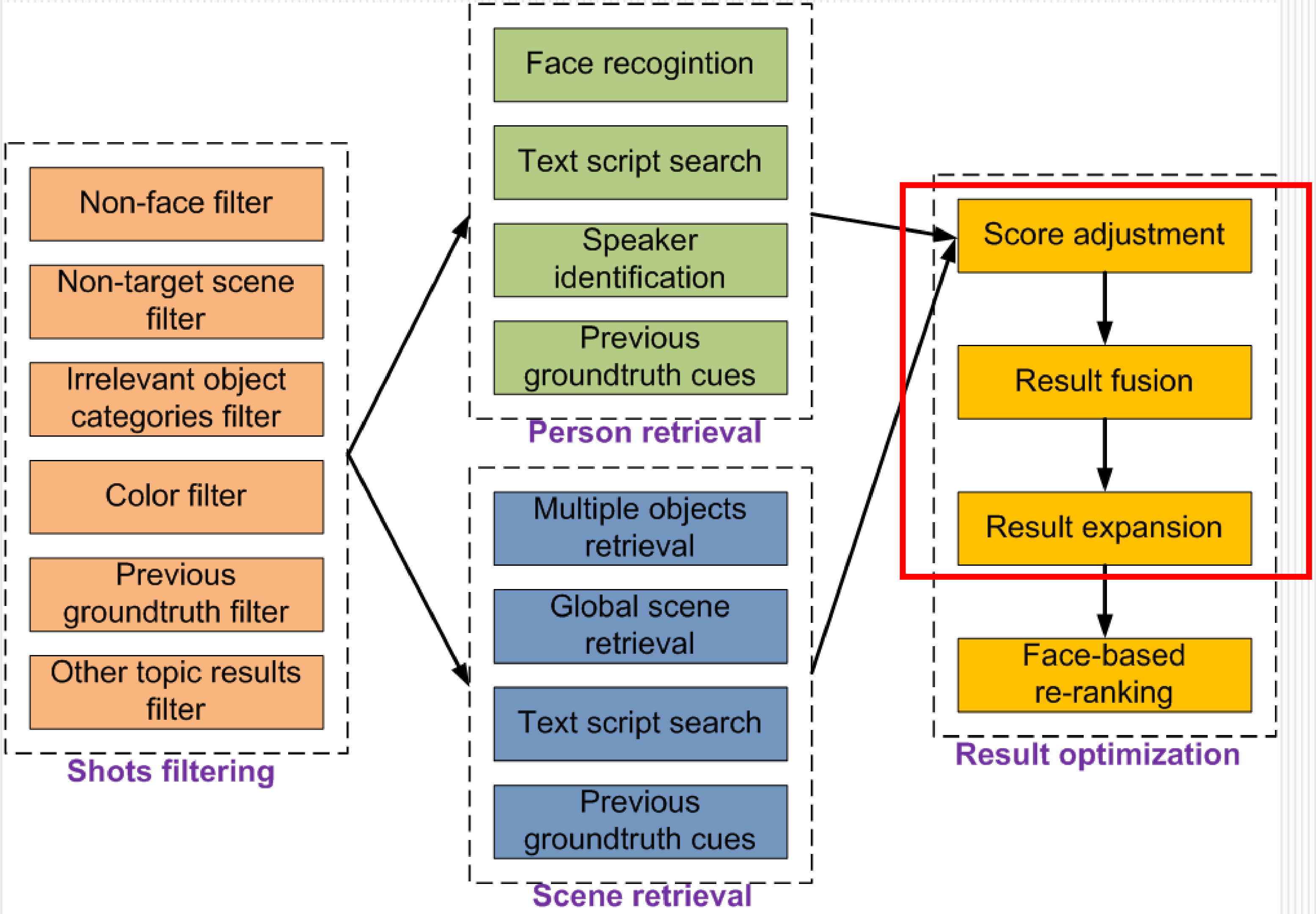
Previous groundtruth filter

- Some landmark objects only appear in a specific location.
- Some objects must not be contained in the topics of this year.
- We filter **12,006** shots



year	NO.	topic	Location									
			cafe1	cafe2	foyer	kitchen1	kitchen2	laund	LR1	LR2	market	pub
13	9069	a circular 'no smoking' logo	1	1	0	0	0	0	0	0	1	1
	9070	a small red obelisk	0	0	0	0	0	0	0	2	0	0
	9071	an Audi logo	0	0	0	0	0	0	0	0	0	0
	9072	a Metropolitan Police logo										
	9073	this ceramic cat face	0	0	0	2	0	0	0	0	0	0
	9074	a cigarette										
	9075	a SKOE can										
	9076	this monochrome bust of Qu	0	0	0	0	0	0	0	0	0	2
	9077	this dog										
	9078	a JENKINS logo	0	0	0	0	0	0	0	0	0	2
	9079	this CD stand in the market	0	0	0	0	0	0	0	0	0	0
	9080	this public phone booth	0	0	0	0	0	0	0	0	0	0
	9081	a black taxi	0	0	0	0	0	0	0	0	0	0
	9082	a BMW logo	0	0	0	0	0	0	0	0	0	0
	9083	a chrome and glass cafetiere										
	9084	this man										
	9085	this David refrigerator magn	0	0	0	0	0	0	0	0	0	0
	9086	these scales	0	0	0	0	0	2	0	0	0	0
	9087	a VW logo	0	0	0	0	0	0	0	0	0	0
	9088	Tamwar										
	9089	this pendant										
	9090	this wooden bench with rou	0	0	0	0	0	0	0	0	0	0
	9091	a Kathy's menu with stripes	0	2	0	0	0	0	0	0	0	0
	9092	this man										
	9093	these turnstiles	0	0	0	0	0	0	0	0	0	0
	9094	a tomato-shaped ketchup di	0	2	0	0	0	0	0	0	0	0

2 Proposed Approach



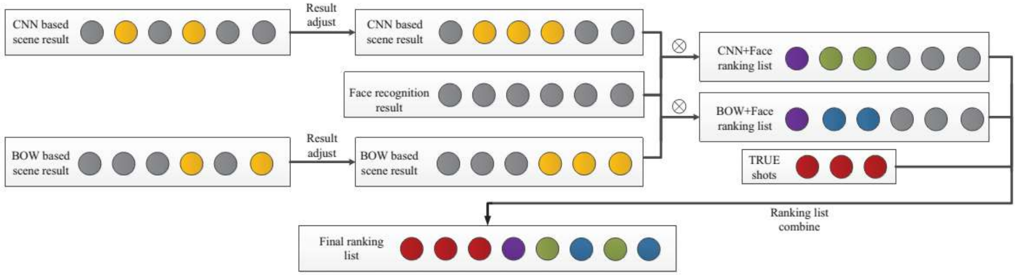
Score adjustment and Result expansion

- The scene in TV series is likely to be blocked by the person, which causes the similarity scores of such shots are not high.
- we find high-score shots with high slope of the score curve, and adjust those missed low-score shots among adjacent high-score shots.

(a) shot192_1199 (b) shot192_1201 (c) shot192_1205

Result fusion

- three score vectors which have values from 0 to 1



Outline

3

Introduction
Problem and Motivation

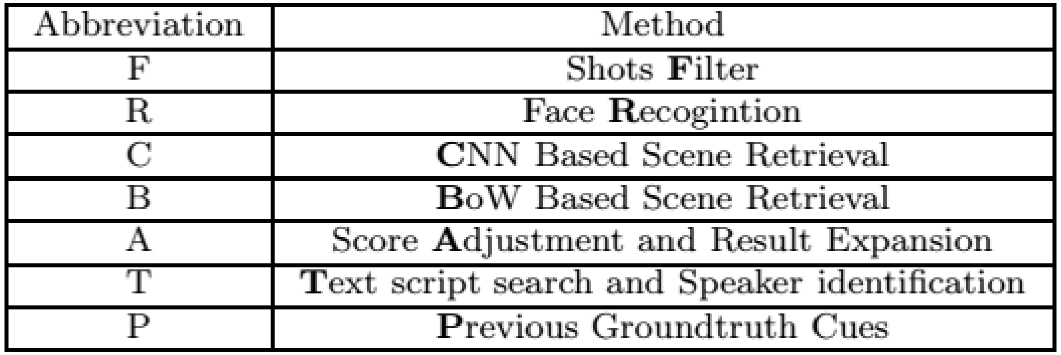
Proposed Approach
Framework and Details

Results
4 runs

Conclusion

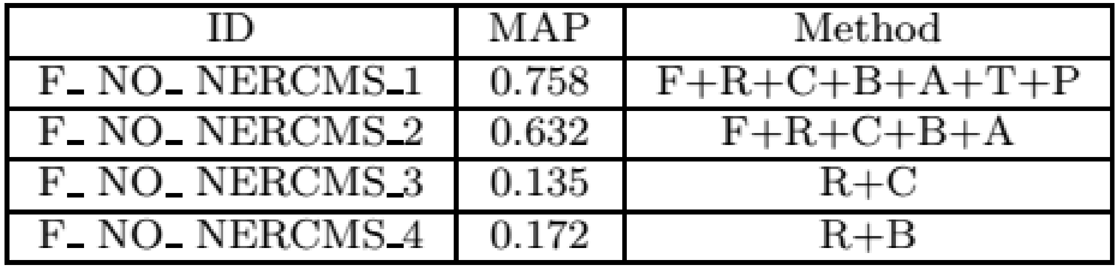
3 Results

Description of our methods



Abbreviation	Method
F	Shots **F**ilter
R	Face **R**ecogintion
C	**C**NN Based Scene Retrieval
B	**B**oW Based Scene Retrieval
A	Score **A**djustment and Result Expansion
T	**T**ext script search and Speaker identification
P	**P**revious Groundtruth Cues

Results of our submitted 4 runs



ID	MAP	Method
F_ NO_ NERCMS_1	0.758	F+R+C+B+A+T+P
F_ NO_ NERCMS_2	0.632	F+R+C+B+A
F_ NO_ NERCMS_3	0.135	R+C
F_ NO_ NERCMS_4	0.172	R+B

Outline

4

Introduction

Problem and Motivation

Proposed Approach

Framework and Details

Results

4 runs

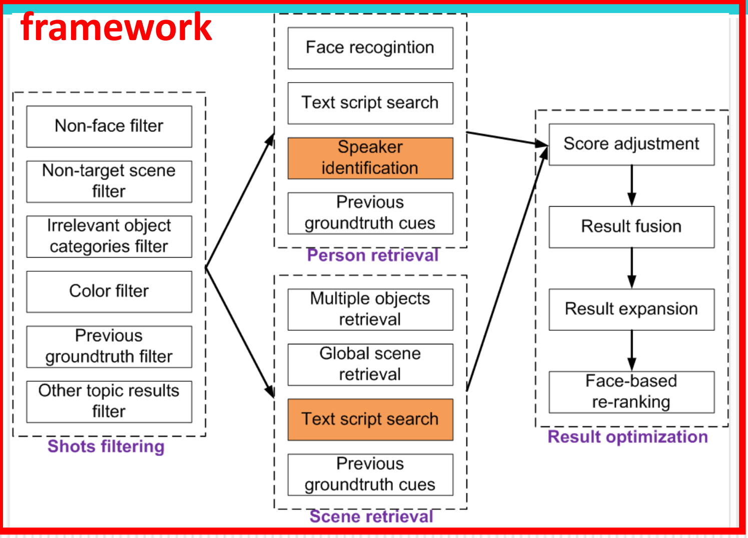
Conclusion

4 Conclusion

1	**Specific person:**	**Face recognition + Face library**
2	**Specific scene:**	**Local view (BoW) + Global view (CNN)**
3	**Result combination:**	**Score adjustment + Results expansion**
4	**Shots filter:**	**Non face + Outdoor scene + Groundtruth**

THANKS

2 Proposed Approach



Text script retrieval and Speaker identification

- Text script: for the target person Jim, the retrieval keywords are Brads, Stace, Stacey, Bradley, Dot, because they are family
- 412 audio library: target persons-6 voice segments of each person, the rest 93 persons-4 voice segments of each person
- MFCC feature of all voice segment

(a) shot5_1269 (b) shot10_279

(a) shot63_1614 (b) shot76_147