Computer Vision for the Blind: a Comparison of Face Detectors in a Relevant Scenario

Marco De Marco Gianfranco Fenu
Eric Medvet Felice Andrea Pellegrino

Department of Engineering and Architecture
University of Trieste
Italy

Goodtechs, 30/11–1/12 2016, Venice (Italy)

http://machinelearning.inginf.units.it

Blindness

- Many assistants proposed to aid blind and visually impaired persons
- Some of them consists of a smart First Person Video (FPV) device, worn by the blind, for easing social interactions:
  - Is there anybody around?
  - How many people?
  - Is there someone I know?
  - Is there someone approaching me?

Blindness

- Many assistants proposed to aid blind and visually impaired persons
- Some of them consists of a smart First Person Video (FPV) device, worn by the blind, for easing social interactions:
  - Is there anybody around?
  - How many people?
  - Is there someone I know?
  - Is there someone approaching me?

Face Detection is an essential step: how effective are current detectors on real FPV images?

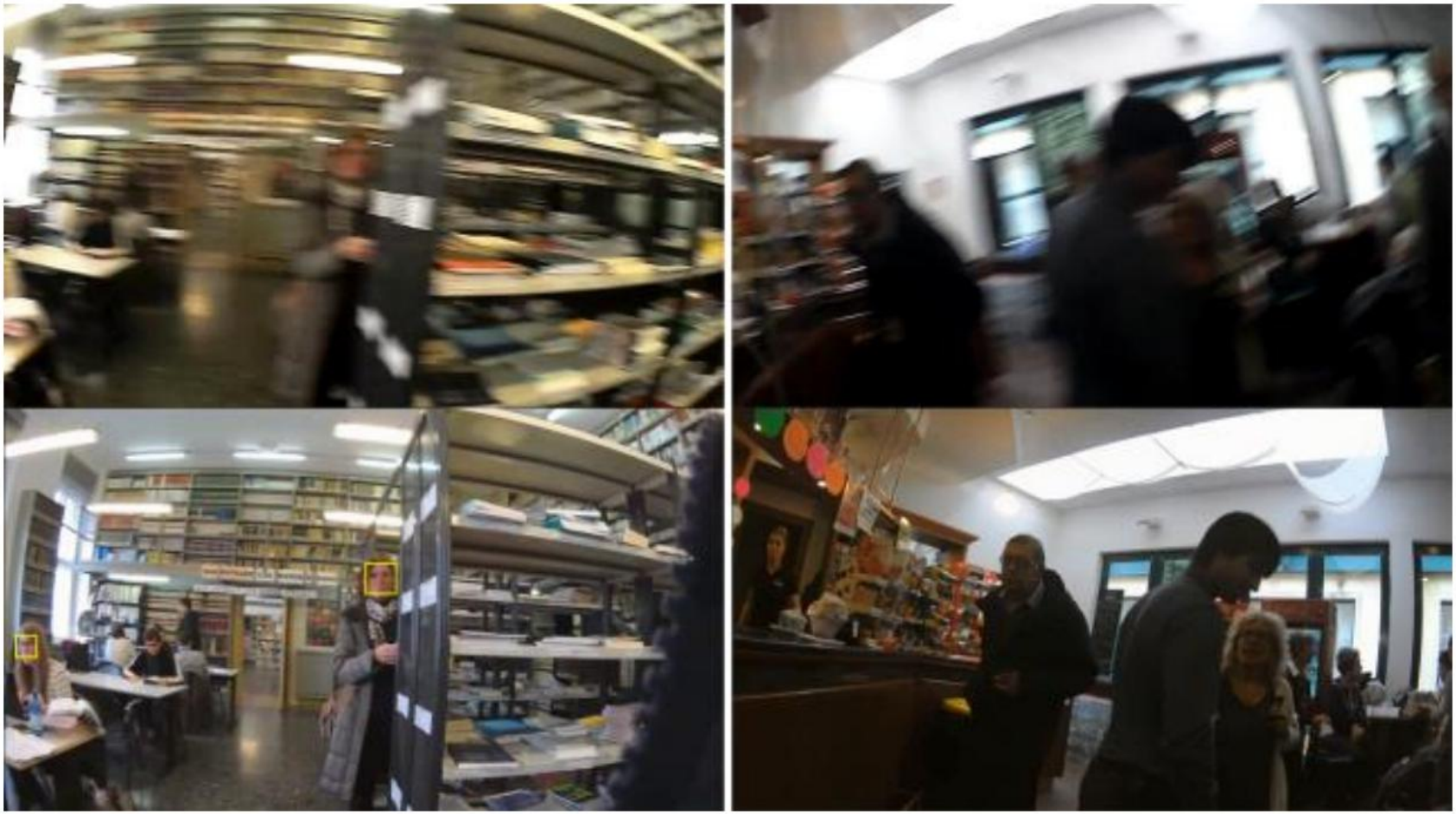
MACHINE LEARNING LAB

Real FPV images

- Motion blur (mannerism, loppy device)
- Suboptimal framing
- Rapidly varying light conditions
- Occlusions
- Distortion (wide-angle device)

MACHINE LEARNING LAB

Real FPV images



MACHINE LEARNING LAB

Our work in brief

1. 4 relevant video sequences, manually annotated
2. 6 recent face detectors
3. experimental comparison
   - are detector effective?
   - what kind of faces do they struggle to detect?

Video sequences

Name	Resolution	Camera	Location	# frames	# faces
Coffee-shop	1280×720	GX9	Indoor	361	809
Library	1280×720	GX9	Indoor	361	1074
Office	1920×1080	CUBE	Indoor	558	206
Bus-stop	1920×1080	CUBE	Outdoor	448	1610

- acquired by a blind person (with all the privacy-related issues correctly addressed)
- two different worn devices (124° and 135°)
- many interactions

Manual annotation

For each frame, each face largest than 20 px

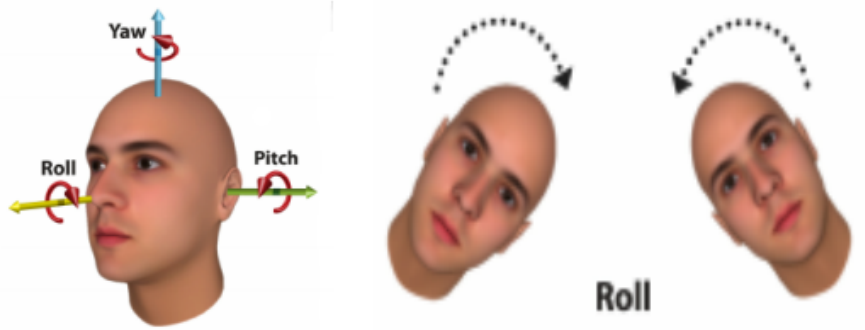
- bounding box (specific criteria)
- centers of eyes and mouth
- occlusion flag
- lateral flag

MACHINE LEARNING LAB

Faces features

Aimed at better characterizing detectors behavior:

- normalized bounding box area (NBBA): are smaller (farther) faces harder to detect?
- normalized distance from the center of the image (NDFC): are peripheral (distorted) faces harder to detect?
- root mean square contrast (RMSC) within the bounding box
- roll angle: are oblique faces harder to detect?
- occlusion: are occluded faces harder to detect?
- lateral (yaw): are lateral faces harder to detect?

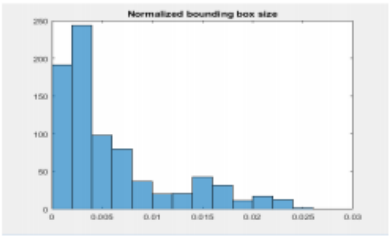


MACHINE LEARNING LAB

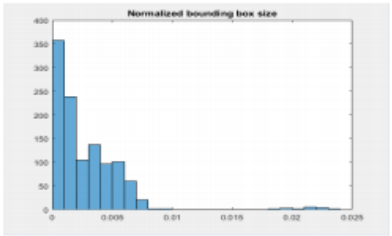
Faces features

We set a statically chosen threshold on each feature, assuming a trivial relation between feature and easyness of detection

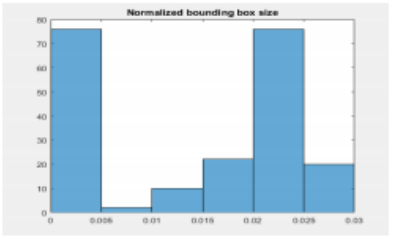
- e.g., NBBA $\leq \tau$ means small, hence harder to detect
- e.g., NDFC $\geq \tau$ means distorted, hence harder to detect



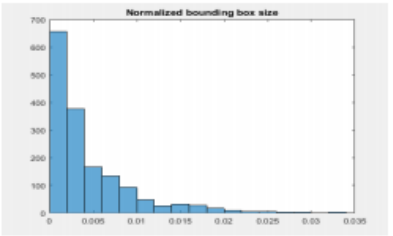
Coffee-shop



Library



Office



Bus-stop

Contenders

Recent, with available implementation

- Viola-Jones (VJ), from Matlab Computer Vision Toolbox
- GMS Vision, from Android JDK + OpenCV for frame grabbing
- Normalized Pixel Difference (NPD), authors' code
- Pixel Intensity Comparison (PICO), authors' code
- Face-Id, deep learning, both detection and recognition, authors' code
- Visage, commercial solution for 2D/3D face identification, demo tool

All with default parameters (fairness)

Detector assessment

Computed on a sequence:

- true positives (TP): *number* of detected faces
- false positives (FP): *number* of detections which are not faces
- false negatives (FN): *number* of undetected faces

Cast as:

- precision, ratio of detected faces among all detections: $\frac{\text{TP}}{\text{TP+FP}}$
- recall, ratio of detected faces among all faces: $\frac{\text{TP}}{\text{TP+FN}}$
- false positives per frame (FPPF): $\frac{\text{FP}}{n_f}$
  - meaningful for video: how often a wrong detection occurs?

Comparable among sequences

MACHINE LEARNING LAB

Detected vs. undetected face

On a single frame:

- zero or more detections d_i (regions deemed to contain a face)
- zero or more faces g_i (manually annotated bounding boxes)

How to decide/count TP, FP, FN?

MACHINE LEARNING LAB

Detected vs. undetected face

On a single frame:

- zero or more detections d_i (regions deemed to contain a face)
- zero or more faces g_i (manually annotated bounding boxes)

How to decide/count TP, FP, FN?

1. for each i, j, compute Intersection to Union Areas Ratio
   $\text{IUAR}(d_i, g_j) = \frac{\text{area}(d_i \cap g_j)}{\text{area}(d_i \cup g_j)}$
2. find best matches (using Hungarian algorithm)
3. decide
   - $\text{IUAR}(d_i, g_j) > 0.5$, d_i is a TP
   - $\text{IUAR}(d_i, g_j) \leq 0.5$, d_i is a FP
   - g_j is not assigned to any d_i, g_j is a FN

MACHINE LEARNING LAB

Results: general

Method	Sequence	Precision	Recall	FPPF
Viola-Jones	Coffee-shop	0.129	0.367	5.543
	Library	0.140	0.267	4.867
	Office	0.031	0.709	8.197
	Bus-stop	0.222	0.725	9.158
	Average	0.132	0.513	7.196
GMS	Coffee-shop	0.364	0.015	0.058
	Library	1.000	0.004	0.000
	Office	0.387	0.141	0.082
	Bus-stop	0.202	0.020	0.290
	Average	0.284	0.021	0.114
NPD	Coffee-shop	0.228	0.305	2.319
	Library	0.159	0.222	3.504
	Office	0.256	0.583	0.625
	Bus-stop	0.687	0.747	1.221
	Average	0.376	0.489	1.735

Method	Sequence	Precision	Recall	FPPF
PICO	Coffee-shop	0.337	0.121	0.535
	Library	0.030	0.003	0.266
	Office	0.538	0.413	0.131
	Bus-stop	0.202	0.020	0.290
	Average	0.589	0.160	0.238
Face-Id	Coffee-shop	0.143	0.001	0.017
	Library	0.889	0.007	0.003
	Office	–	0.0	0.0
	Bus-stop	1.000	0.001	0.000
	Average	0.611	0.003	0.004
Visage	Coffee-shop	0.043	0.002	0.125
	Library	0.045	0.001	0.058
	Office	0.137	0.068	0.158
	Bus-stop	0.072	0.006	0.286
	Average	0.087	0.007	0.163

Results: general

Method	Sequence	Precision	Recall	FPPF
Viola-Jones	Coffee-shop	0.129	0.367	5.543
	Library	0.140	0.267	4.867
	Office	0.031	0.709	8.197
	Bus-stop	0.222	0.725	9.158
	Average	0.132	0.513	7.196
GMS	Coffee-shop	0.364	0.015	0.058
	Library	1.000	0.004	0.000
	Office	0.387	0.141	0.082
	Bus-stop	0.202	0.020	0.290
	Average	0.284	0.021	0.114
NPD	Coffee-shop	0.228	0.305	2.319
	Library	0.159	0.222	3.504
	Office	0.256	0.583	0.625
	Bus-stop	0.687	0.747	1.221
	Average	0.376	0.489	1.735

Method	Sequence	Precision	Recall	FPPF
PICO	Coffee-shop	0.337	0.121	0.535
	Library	0.030	0.003	0.266
	Office	0.538	0.413	0.131
	Bus-stop	0.202	0.020	0.290
	Average	0.589	0.160	0.238
Face-Id	Coffee-shop	0.143	0.001	0.017
	Library	0.889	0.007	0.003
	Office	–	0.0	0.0
	Bus-stop	1.000	0.001	0.000
	Average	0.611	0.003	0.004
Visage	Coffee-shop	0.043	0.002	0.125
	Library	0.045	0.001	0.058
	Office	0.137	0.068	0.158
	Bus-stop	0.072	0.006	0.286
	Average	0.087	0.007	0.163

- All detectors perform poorly on average

Results: general

Method	Sequence	Precision	Recall	FPPF
Viola-Jones	Coffee-shop	0.129	0.367	5.543
	Library	0.140	0.267	4.867
	Office	0.031	0.709	8.197
	Bus-stop	0.222	0.725	9.158
	Average	0.132	0.513	7.196
GMS	Coffee-shop	0.364	0.015	0.058
	Library	1.000	0.004	0.000
	Office	0.387	0.141	0.082
	Bus-stop	0.202	0.020	0.290
	Average	0.284	0.021	0.114
NPD	Coffee-shop	0.228	0.305	2.319
	Library	0.159	0.222	3.504
	Office	0.256	0.583	0.625
	Bus-stop	0.687	0.747	1.221
	Average	0.376	0.489	1.735

Method	Sequence	Precision	Recall	FPPF
PICO	Coffee-shop	0.337	0.121	0.535
	Library	0.030	0.003	0.266
	Office	0.538	0.413	0.131
	Bus-stop	0.202	0.020	0.290
	Average	0.589	0.160	0.238
Face-Id	Coffee-shop	0.143	0.001	0.017
	Library	0.889	0.007	0.003
	Office	–	0.0	0.0
	Bus-stop	1.000	0.001	0.000
	Average	0.611	0.003	0.004
Visage	Coffee-shop	0.043	0.002	0.125
	Library	0.045	0.001	0.058
	Office	0.137	0.068	0.158
	Bus-stop	0.072	0.006	0.286
	Average	0.087	0.007	0.163

- All detectors perform poorly on average
- Best is NPD on Bus-stop, but with 1.2 FPPF!

MACHINE LEARNING LAB

Results: general

Method	Sequence	Precision	Recall	FPPF
Viola-Jones	Coffee-shop	0.129	0.367	5.543
	Library	0.140	0.267	4.867
	Office	0.031	0.709	8.197
	Bus-stop	0.222	0.725	9.158
	Average	0.132	0.513	7.196
GMS	Coffee-shop	0.364	0.015	0.058
	Library	1.000	0.004	0.000
	Office	0.387	0.141	0.082
	Bus-stop	0.202	0.020	0.290
	Average	0.284	0.021	0.114
NPD	Coffee-shop	0.228	0.305	2.319
	Library	0.159	0.222	3.504
	Office	0.256	0.583	0.625
	Bus-stop	0.687	0.747	1.221
	Average	0.376	0.489	1.735

Method	Sequence	Precision	Recall	FPPF
PICO	Coffee-shop	0.337	0.121	0.535
	Library	0.030	0.003	0.266
	Office	0.538	0.413	0.131
	Bus-stop	0.202	0.020	0.290
	Average	0.589	0.160	0.238
Face-Id	Coffee-shop	0.143	0.001	0.017
	Library	0.889	0.007	0.003
	Office	–	0.0	0.0
	Bus-stop	1.000	0.001	0.000
	Average	0.611	0.003	0.004
Visage	Coffee-shop	0.043	0.002	0.125
	Library	0.045	0.001	0.058
	Office	0.137	0.068	0.158
	Bus-stop	0.072	0.006	0.286
	Average	0.087	0.007	0.163

- All detectors perform poorly on average
- Best is NPD on Bus-stop, but with 1.2 FPPF!
- Clear trade-off between precision (FPPF) and recall
  - differences among detectors (e.g., Face-Id vs. VJ)

Results: recall w.r.t. features

Method	NBBA		NDFC		Roll		RMSC		L/NL		O/NO	
	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	NL	L	NO	O
Face-Id	0.001	0.001	0.001	0.002	0.001	0.000	0.000	0.002	0.001	0.001	0.002	0.000
GMS	0.006	0.039	0.004	0.041	0.041	0.002	0.006	0.039	0.044	0.001	0.043	0.002
NPD	0.304	0.160	0.122	0.342	0.443	0.009	0.277	0.188	0.441	0.024	0.441	0.023
PICO	0.054	0.143	0.046	0.151	0.190	0.005	0.093	0.104	0.196	0.001	0.187	0.010
Viola-Jones	0.364	0.149	0.148	0.366	0.491	0.004	0.325	0.189	0.486	0.027	0.475	0.038
Visage	0.002	0.018	0.002	0.018	0.019	0.000	0.003	0.017	0.019	0.001	0.020	0.000

MACHINE LEARNING LAB

Results: recall w.r.t. features

Method	NBBA		NDFC		Roll		RMSC		L/NL		O/NO	
	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	NL	L	NO	O
Face-Id	0.001	0.001	0.001	0.002	0.001	0.000	0.000	0.002	0.001	0.001	0.002	0.000
GMS	0.006	0.039	0.004	0.041	0.041	0.002	0.006	0.039	0.044	0.001	0.043	0.002
NPD	0.304	0.160	0.122	0.342	0.443	0.009	0.277	0.188	0.441	0.024	0.441	0.023
PICO	0.054	0.143	0.046	0.151	0.190	0.005	0.093	0.104	0.196	0.001	0.187	0.010
Viola-Jones	0.364	0.149	0.148	0.366	0.491	0.004	0.325	0.189	0.486	0.027	0.475	0.038
Visage	0.002	0.018	0.002	0.018	0.019	0.000	0.003	0.017	0.019	0.001	0.020	0.000

- occluded/lateral/oblique (roll) faces are much harder to detect

MACHINE LEARNING LAB

Results: recall w.r.t. features

Method	NBBA		NDFC		Roll		RMSC		L/NL		O/NO	
	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	NL	L	NO	O
Face-Id	0.001	0.001	0.001	0.002	0.001	0.000	0.000	0.002	0.001	0.001	0.002	0.000
GMS	0.006	0.039	0.004	0.041	0.041	0.002	0.006	0.039	0.044	0.001	0.043	0.002
NPD	0.304	0.160	0.122	0.342	0.443	0.009	0.277	0.188	0.441	0.024	0.441	0.023
PICO	0.054	0.143	0.046	0.151	0.190	0.005	0.093	0.104	0.196	0.001	0.187	0.010
Viola-Jones	0.364	0.149	0.148	0.366	0.491	0.004	0.325	0.189	0.486	0.027	0.475	0.038
Visage	0.002	0.018	0.002	0.018	0.019	0.000	0.003	0.017	0.019	0.001	0.020	0.000

- occluded/lateral/oblique (roll) faces are much harder to detect
- larger faces (NBBA) are easier to detect, except with NPD and Viola-Jones
  - detectors parameters
- contrast eases detection, except with NPD and Viola-Jones

MACHINE LEARNING LAB

Results: recall w.r.t. features

Method	NBBA		NDFC		Roll		RMSC		L/NL		O/NO	
	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	$< \tau$	$\geq \tau$	NL	L	NO	O
Face-Id	0.001	0.001	0.001	0.002	0.001	0.000	0.000	0.002	0.001	0.001	0.002	0.000
GMS	0.006	0.039	0.004	0.041	0.041	0.002	0.006	0.039	0.044	0.001	0.043	0.002
NPD	0.304	0.160	0.122	0.342	0.443	0.009	0.277	0.188	0.441	0.024	0.441	0.023
PICO	0.054	0.143	0.046	0.151	0.190	0.005	0.093	0.104	0.196	0.001	0.187	0.010
Viola-Jones	0.364	0.149	0.148	0.366	0.491	0.004	0.325	0.189	0.486	0.027	0.475	0.038
Visage	0.002	0.018	0.002	0.018	0.019	0.000	0.003	0.017	0.019	0.001	0.020	0.000

- occluded/lateral/oblique (roll) faces are much harder to detect
- larger faces (NBBA) are easier to detect, except with NPD and Viola-Jones
  - detectors parameters
- contrast eases detection, except with NPD and Viola-Jones
- unclear impact of NDFC: easier if far from the center
  - further investigation needed

MACHINE LEARNING LAB

Concluding remarks and future work

Considered detectors perform poorly in this scenario

- change scenario: e.g., detection of approaching faces
- use video, rather than a set of still images

MACHINE LEARNING LAB

Thanks!