Semi-Supervised Evaluation of Face Recognition in Videos

V. Biaud, C. Herold, V. Despiegel, S. Gentric

stephane.gentric@morpho.com

IBPC 2014
April, 2nd

PURPOSE

- **Face recognition on still images is a mature topic**
  - Good performances on controlled data
  - Lots of databases available, well established metrics

- **Development of face recognition in video raises new issues, which requires dedicated data for training and evaluation**
  - Uncontrolled conditions in terms of pose, illumination, expression, resolution
  - How to make use of temporal, spatial and contextual information available on videos ?

- **Video labeling is a very tedious and time-consuming task**

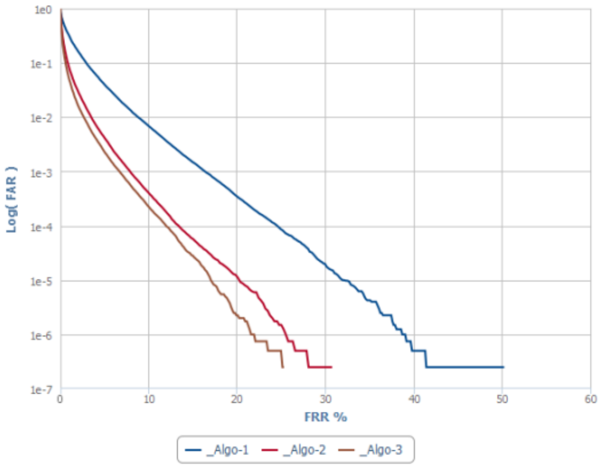
→ **how can we get around this ?**

OUTLINE

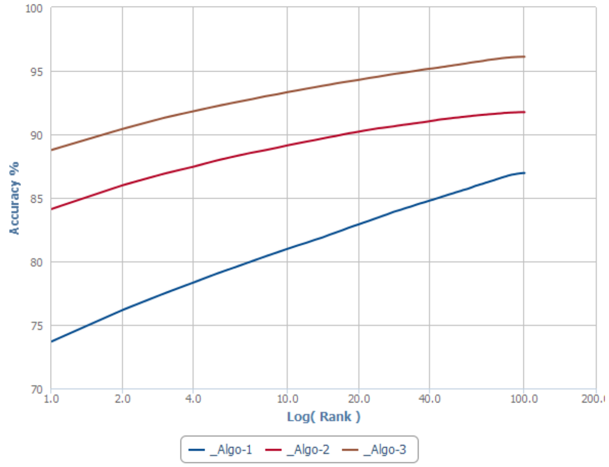
- **Introduction**
- Methodology
- Metric
- Results
- Conclusion

INTRODUCTION

- **For face recognition on still images, evaluation procedures are well defined**
  - Choose a database and **labeled faces with a unique ID**
  - For each algorithm,
    - Compute similarity scores for matching pairs and non-matching pairs
  - Plot standard curves: ROC, CMC



- **Comparison between algorithms can be done on databases representative of real-life scenarios**
  - ID document issuance
  - Mugshot images



INTRODUCTION

- **For face recognition in videos variability increases, making comparisons even more valuable**
  - Various face processing algorithms for detection, tracking, coding and comparison
  - Different scenarios: Mono/multi-camera, mono/multi-person, frame rate, illumination, etc.

- **How to evaluate the different face recognition algorithms ?**

- **Is it possible to evaluate algorithms without proper labeling, and if so what are the underlying assumption and bias ?**

OUTLINE

- Introduction
- **Methodology**
- Metric
- Results
- Conclusion

METHODOLOGY

- **Evaluation of various tracking strategies/various coding algorithms on specific video scenarios.**

- **What kind of ground truth information could we expect to have for next to no effort?**
  - Identities & boxes for each and every timestamp: extremely costly to generate
  - Identities & timestamps of presence in the video : unfortunately, not always available
  - List of persons that should/could appear in the video : nearly always

- **What kind of metric could we define?**

- **How fair would they be in term of algorithm comparison?**

METHODOLOGY

TV shows: large databases of videos, with a given set of actors

Additional data

- **Set of Actors:** prior information used to validate the algorithm results

➔ **Apply the face analysis process to the video** (face tracking, encoding). Output: one template per track.

➔ **Verify if the faces correspond to actors** (face comparison algorithm).

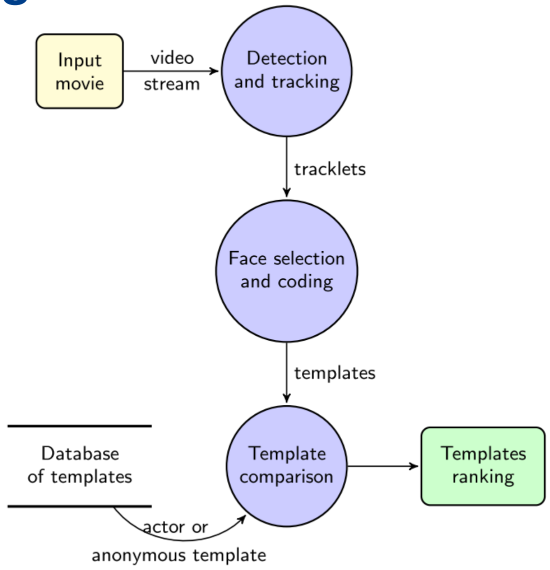
➔ No frame by frame verification (ID or face boxes) → no GT annotation needed. Global verification using biometry → based on the set of actors information

METHODOLOGY

➔ Extracted template is compared to a database containing:

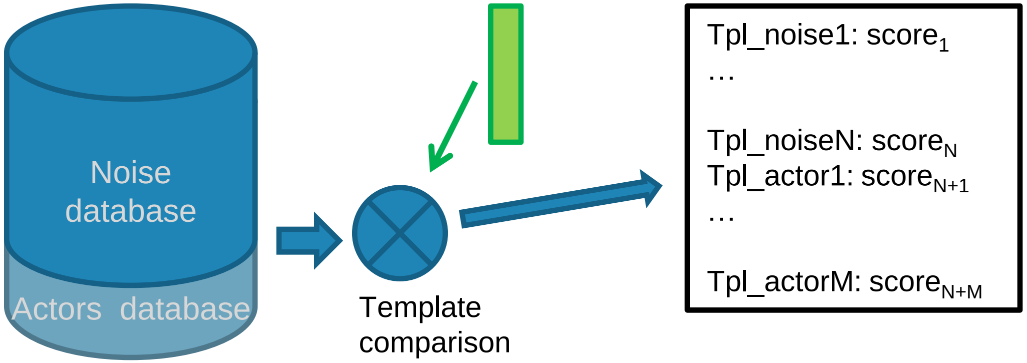
Manual work: generate this database

- **Actors database:** face images corresponding to most of the actors of the video. Mainly extracted from internet. Each actor can be represented multiple times.
- **Noise database:**
  - No image of the actors
  - Representative of the acquisition quality of the actors database.



➔ Generation of comparison scores

(all images are encoded to obtain a facial template)



Tpl_noise1: score$_1$
…

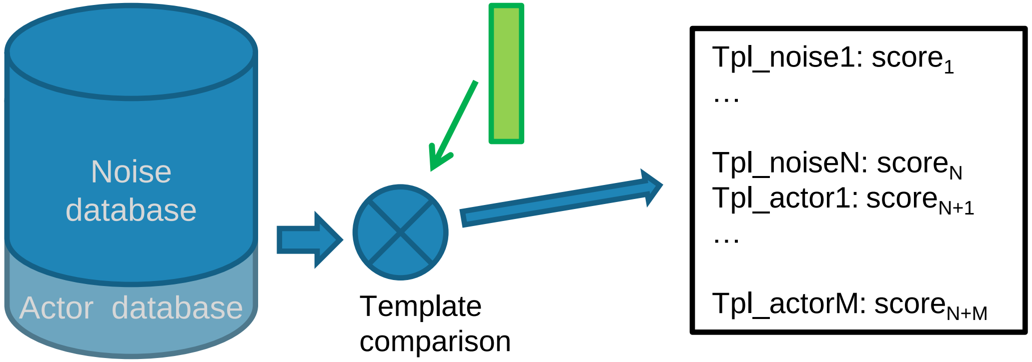
Tpl_noiseN: score$_N$
Tpl_actor1: score$_{N+1}$
…

Tpl_actorM: score$_{N+M}$

METHODOLOGY

➔ Database constitution

- Noise images have to be similar to actors images in terms of:
  - Ethnicity, gender, age
  - Illumination condition
  - Resolution
  - …



Tpl_noise1: $score_1$
…
Tpl_noiseN: $score_N$
Tpl_actor1: $score_{N+1}$
…
Tpl_actorM: $score_{N+M}$

- Proportion:
  - M actors, N noise images
  - If the noise images are similar to the actors images, the probability to match an outsider (not in the actor database) to an actor:

    p= M/(M+N)

.

OUTLINE

- ➔ Introduction
- ➔ Methodology
- ➔ **Metric**
- ➔ Results
- ➔ Conclusion

METRIC

- **From an operational point of view, the critical metric is the number of False Alarms.**
- **A bad threshold may swamp an operator with False Alarms, making the system useless.**
- **The targeted False Alarm Rate depends on the prior probability of finding a person of interest and the cost of processing a false alarm. (for example, in term of operator effort)**
- **The overall performance of the system also depends on the size of the watch list and on the number of persons passing in front of the camera.**
- **For all the 4 following metrics, we compute the false alarm rate in the same manner : per time units and against a watch list of the same size.**

METRIC

(A) with a complete labeling

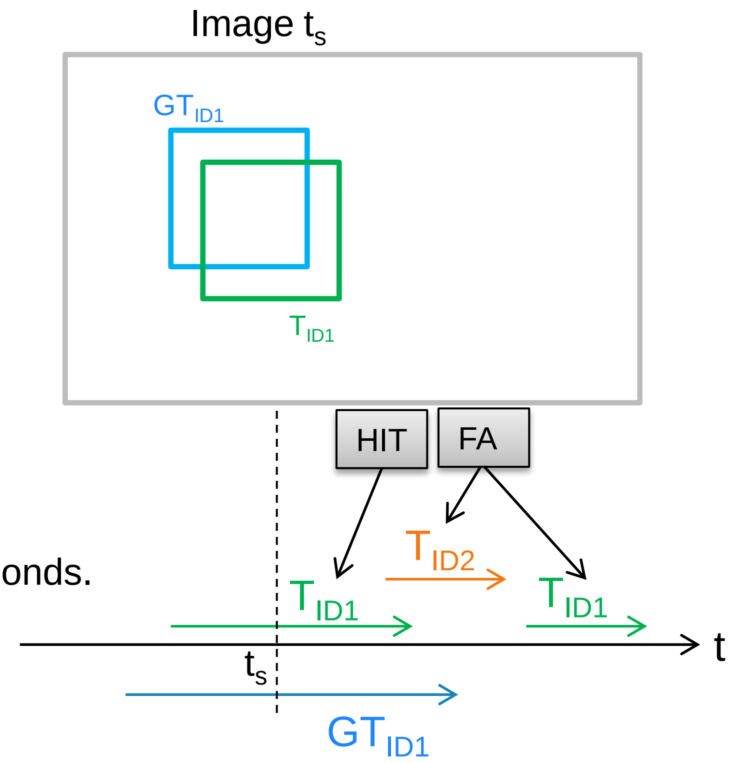
ground truth	evaluated results	metric

full tracks
(ID + timestamps + boxes)
	Tracks (timestamps + boxes)
	with candidate list
		FAR = nb false alarms / hour
		Accuracy = nb HIT / nb GT

- **A candidate track and a GT track are associated when :**
  - At least one frame in common where boxes overlap
- **A HIT is a candidate :**
  - with a score above the threshold.
  - with a track associated with a ground truth track of the same ID

  We count a **maximum of one HIT per GT track**.
- **A False Alarm is a candidate :**
  - with a score above the threshold.
  - That is not a HIT

  We count a maximum of one FA and per face in the gallery per 30 seconds.



METRIC

(B) with a partial labeling

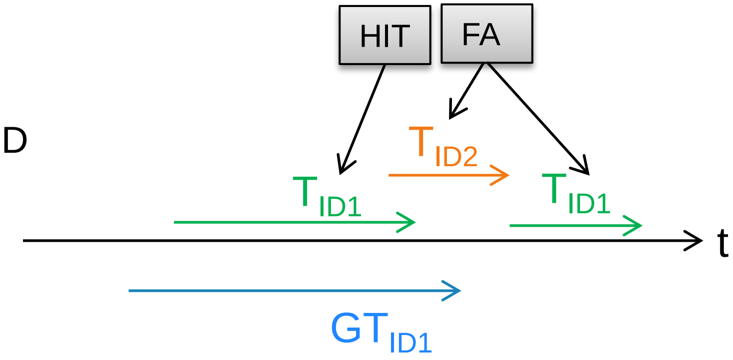
ground truth	evaluated results	metric

presence tracks
(ID + timestamps)
	Tracks (timestamps)
	with candidate list
		FAR = nb alarms / hour
		Accuracy = nb HIT / nb GT

- **A candidate track and a GT track are associated when :**
  - There is at least one frame in common

- **A HIT is a candidate :**
  - with a score above the threshold.
  - with a track associated with a ground truth track of the same ID

  We count a **maximum of one HIT per GT track.**



- **A False Alarm is a candidate :**
  - with a score above the threshold.
  - That is not a HIT

  We count a maximum of one FA and per face in the gallery per 30 seconds.

- **Bias :**
  - Position of a hit is not checked : with multiple faces in the video at the same time, in rare cases, a false alarm can be counted as a hit

METRIC

(C) with one person per video

ground truth	evaluated results	metric

One person per video	candidate lists	FAR = nb alarms / hour
		Accuracy = nb HIT / nb video

- **A HIT is a candidate :**
  - with a score above the threshold.
  - with the ID of the video

  We count a maximum of one HIT per video.

- **A false Alarm is a candidate :**
  - with a score above the threshold.
  - That is not a HIT

  We can have multiple false alarms per video

- **Bias :**
  - Tracking Algorithms can be adapted to this simple case
  - Representative of specific scenarios.

METRIC

(D) semi-supervised metric : using only a set of actors

ground truth	evaluated results	metric

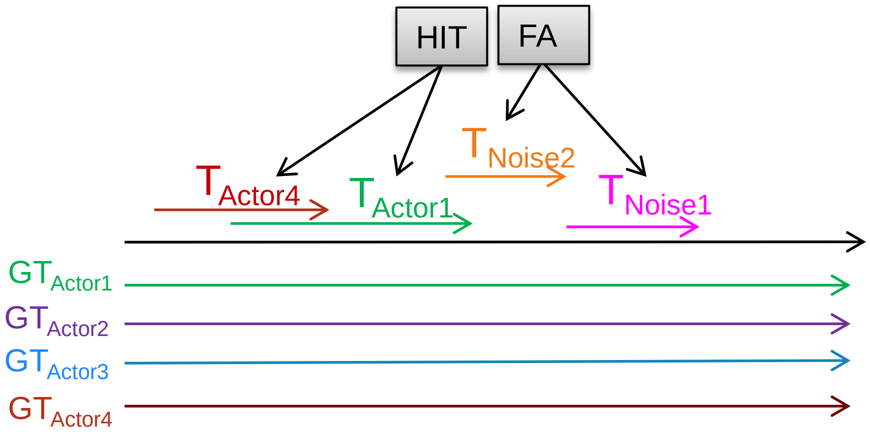
Set of actors	Tracks with candidate list	FAR = nb alarms / hour
		Accuracy = nb HIT

- **A HIT is a candidate :**
  - with a score above the threshold.
  - with the ID of an actor

  We count a maximum of one HIT per actor per 30 seconds.

- **A false Alarm is a candidate :**
  - with a score above the threshold.
  - with an ID coming from of the noise database

  We count a maximum of one HIT per subject per 30 seconds



- **Bias :**
  - No absolute accuracy
  - False Alarm against other actors are counted as hit.

METRIC

- **This metric is by construction a relative metric**
  - Its aims is to compare algorithms (coding, tracking strategies) not to give absolute figures.
  - As for more classical metrics on video, there are a number of unseen characteristics of the video that have a big impact on performances (are the actors frontal in the video, what is the number of persons, is the camera moving, is the illumination uniform, how compressed is the video …)
- **In order to validate this new metric for algorithm comparison, we have compared different algorithms with different metrics :**
  - Our semi supervised metric (D)
  - Metric with partial labeling (B)

OUTLINE

- ➔ Introduction
- ➔ Methodology
- ➔ Metric
- ➔ **Results**
- ➔ Conclusion

RESULTS

➔ Algorithms :

- Detection and Tracking algorithms
  - TR 0 : Basic tracking
  - TR 1 : 3D face tracking
  - TR 2 : Real time tracking

- Feature Extraction and Matching algorithms :
  - FE 1 : Direct encoding.
  - FE 2 : Use of a 3D morphable model.

RESULTS

➔ Databases :

- UK Home Office CAST
  - Ground truth available
  - 10 hours, HD video, different surveillance scenarios
  - set : 100 actors

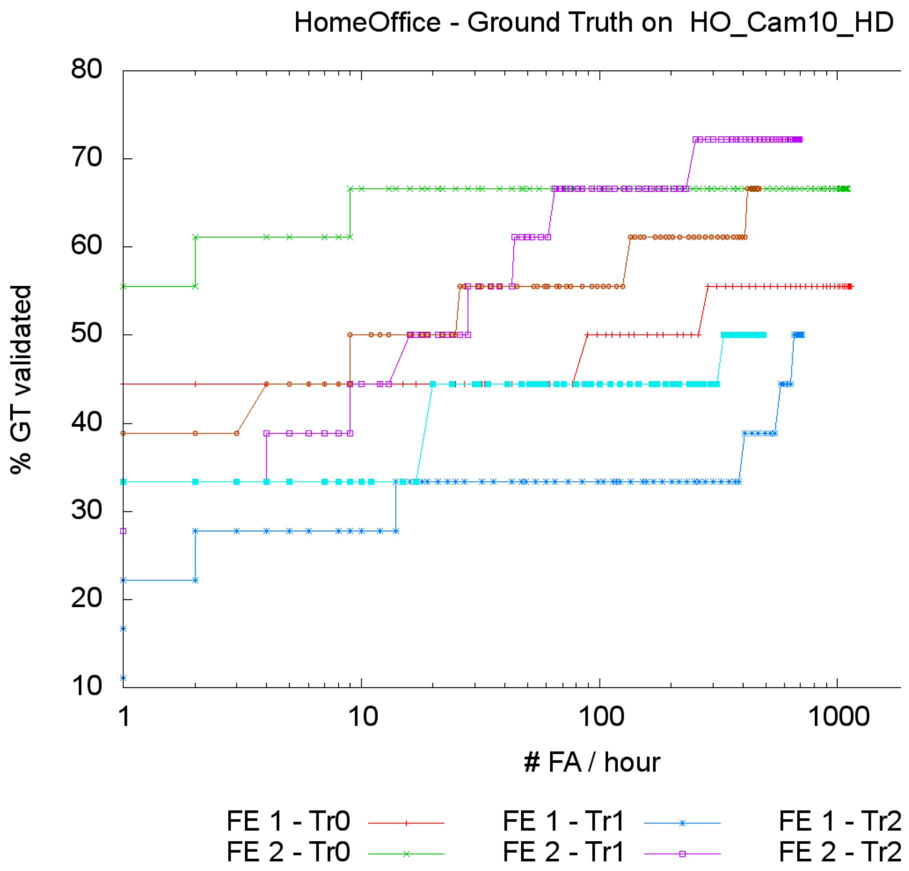
- Prison Break :
  - seasons 1 to 4, 77 hours of videos.
  - set : 20 actors

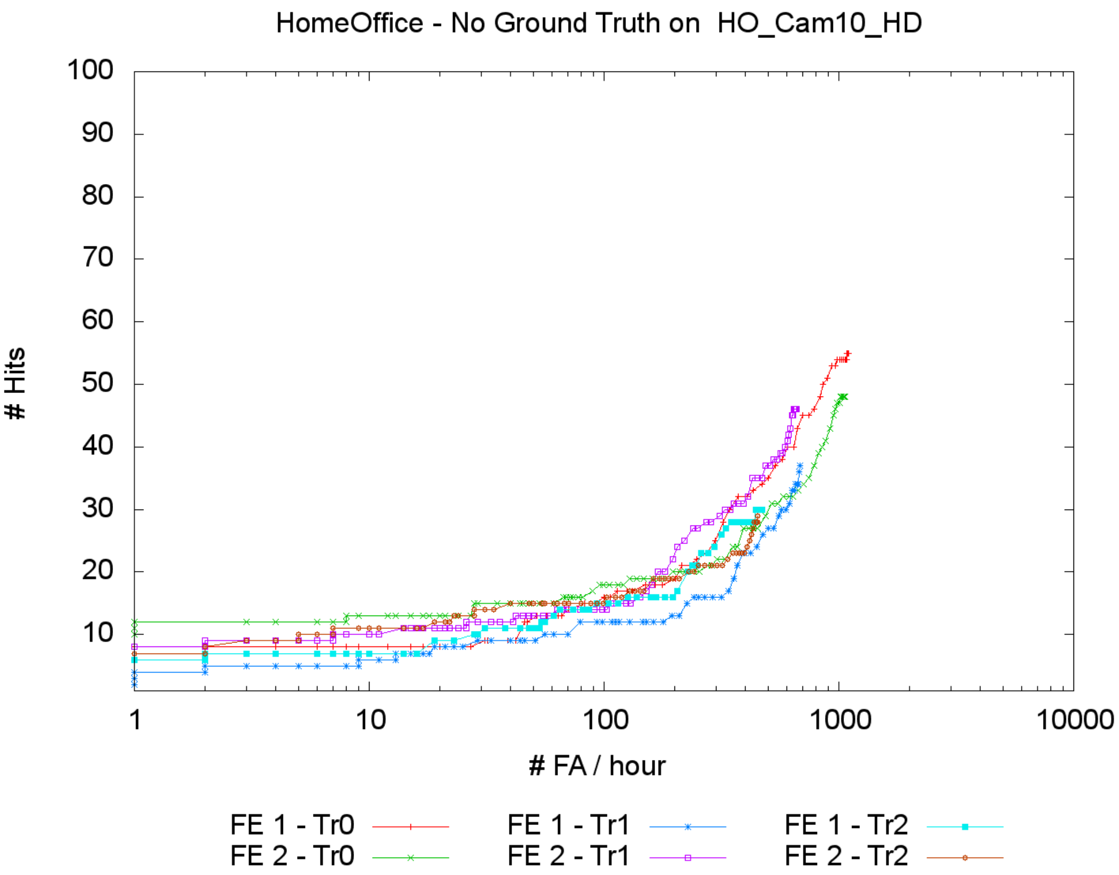
RESULTS

➔ Samples of video

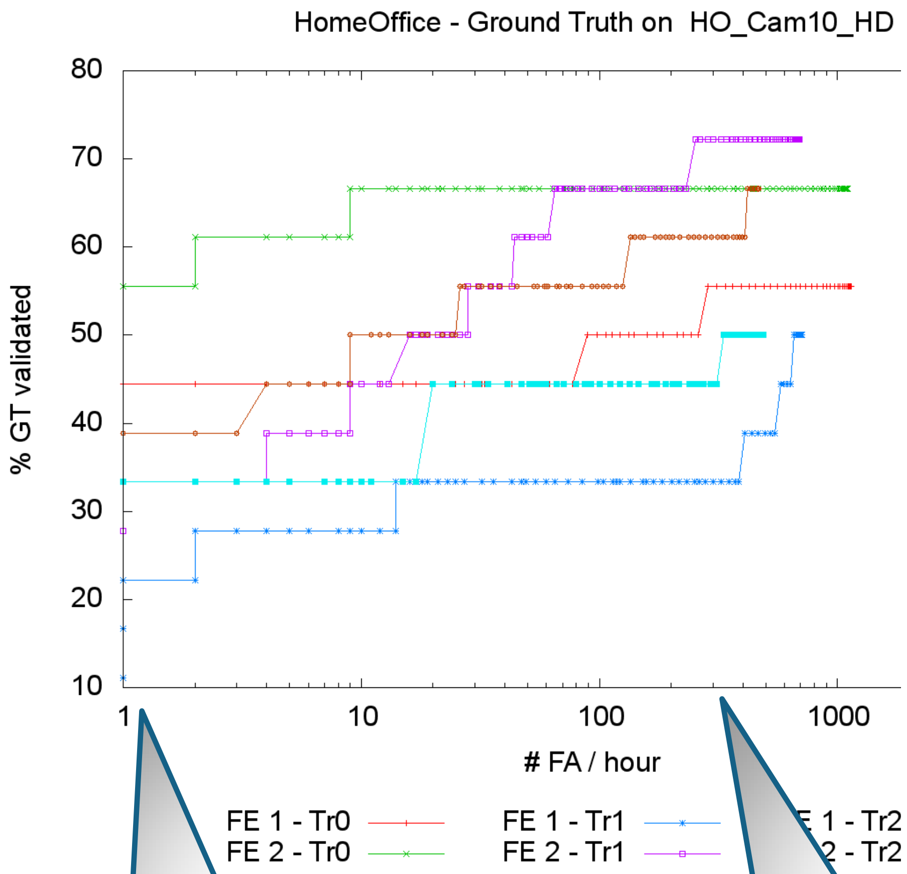
▪ “Grey’s Anatomy” with basic tracking	video
▪ UK Home Office CAST with basic tracking	video
▪ “Prison Break” with 3D tracking	video
▪ “Caméra Café” with 3D tracking	video
▪ UK Home Office CAST with 3D tracking	video
▪ UK Home Office CAST : A Hit from Ground Truth	video

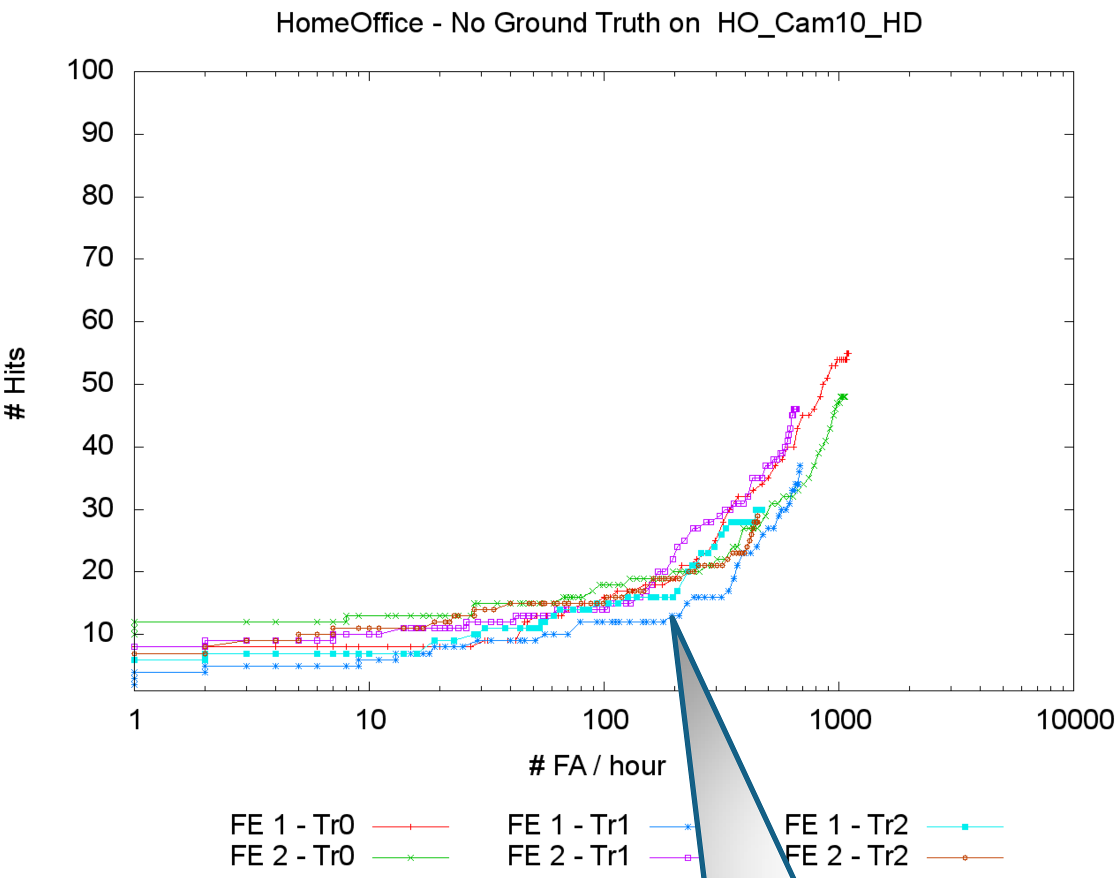
RESULTS


HomeOffice - Ground Truth on HO_Cam10_HD


HomeOffice - No Ground Truth on HO_Cam10_HD

RESULTS



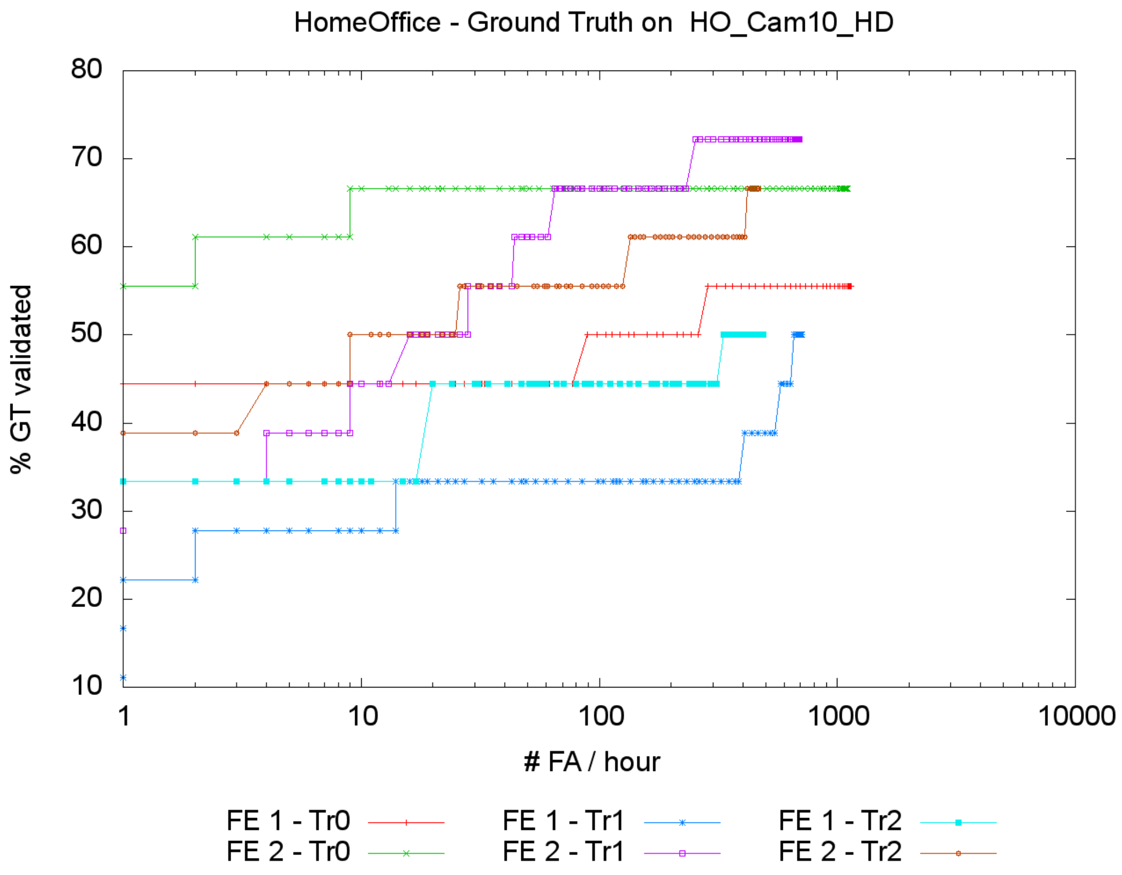


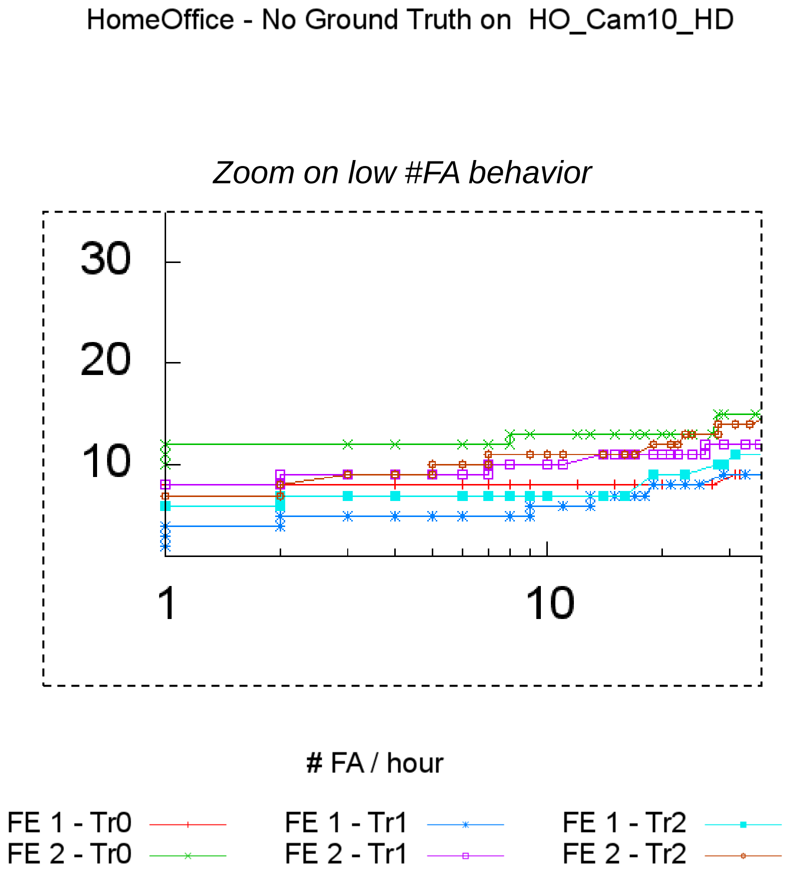
High matching threshold

Low matching threshold :

For low threshold, statistically some non-actors are counted as HIT

RESULTS



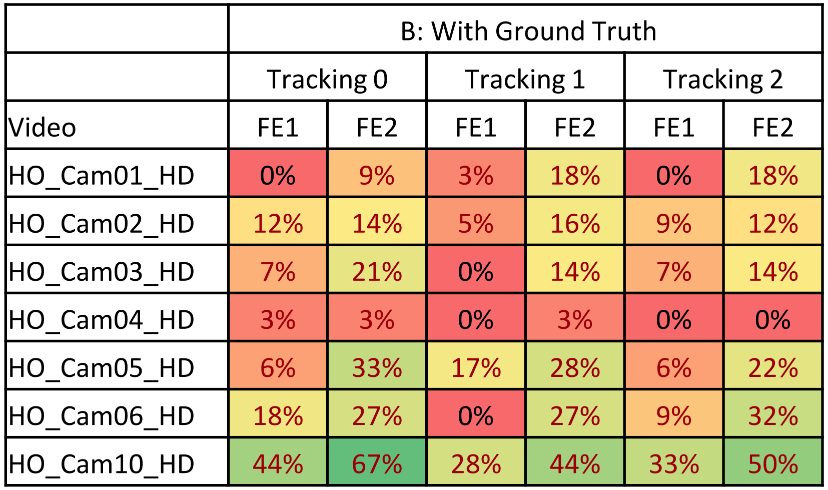


→ Similar tendencies can be noticed between the evaluated algorithms on low #FA/hour range.

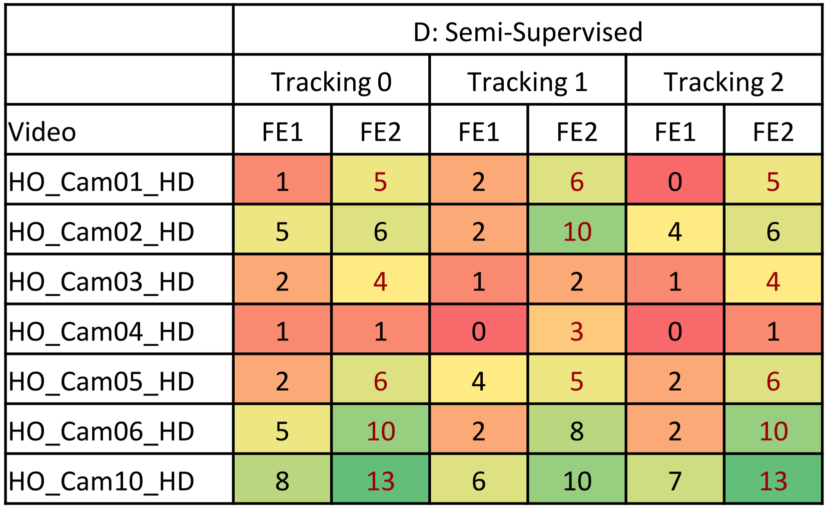
RESULTS

Synthetic tables on CAST videos

- Performances at 10 False Alarms per Hour



	B: With Ground Truth					
	Tracking 0		Tracking 1		Tracking 2	
Video	FE1	FE2	FE1	FE2	FE1	FE2
HO_Cam01_HD	0%	9%	3%	18%	0%	18%
HO_Cam02_HD	12%	14%	5%	16%	9%	12%
HO_Cam03_HD	7%	21%	0%	14%	7%	14%
HO_Cam04_HD	3%	3%	0%	3%	0%	0%
HO_Cam05_HD	6%	33%	17%	28%	6%	22%
HO_Cam06_HD	18%	27%	0%	27%	9%	32%
HO_Cam10_HD	44%	67%	28%	44%	33%	50%



	D: Semi-Supervised					
	Tracking 0		Tracking 1		Tracking 2	
Video	FE1	FE2	FE1	FE2	FE1	FE2
HO_Cam01_HD	1	5	2	6	0	5
HO_Cam02_HD	5	6	2	10	4	6
HO_Cam03_HD	2	4	1	2	1	4
HO_Cam04_HD	1	1	0	3	0	1
HO_Cam05_HD	2	6	4	5	2	6
HO_Cam06_HD	5	10	2	8	2	10
HO_Cam10_HD	8	13	6	10	7	13

Even with a relatively small number of actors, as a first order, the two metrics allow a fair and equivalent comparison of the different algorithms.

RESULTS

➔ Performances on Prison Break

- Noise database: LFW → faces under variable pose

Prison Break: No GT FAR=10FA/h

Tracking 0		Tracking 1	
FE1	FE2	FE1	FE2
1641	2114	1535	2221

- **Tracking 0:** limited to frontal poses.
  Tracking 1: robust to non-frontal pose → more tracks (x2 compared to Tracking 0) → more potential FA.

- **FE1:** Input face directly encoded. Risk of pose matching with non frontal faces of the noise database if input are non-frontal (case Tracking 1 – FE1).
  FE2: Fit a 3DMM to rectify the pose to improve the face comparison.

CONCLUSION

➔ Our method

- Evaluation available over large sets of videos
- No manual labeling needed.
  Requirement: a set of face images corresponding to the actors.
- Comparison of different face algorithms (tracking and coding) under controlled False Alarm Rate.
- Small bias to be careful about, despite being a low cost yet efficient first approximation.

➔ In the future

- Use “Hannah and her Sisters” video (Ground truth available)
- Exhaustive Internal Evaluation of algorithms on non-annotated video data.
- Ground truth information automatic generation on images for algorithm training.

THANK YOU