IARPA JANUS

Online Open World Face Recognition From Video Streams

ID:23202

Federico Pernici, Federico Bartoli, Matteo Bruni and Alberto Del Bimbo

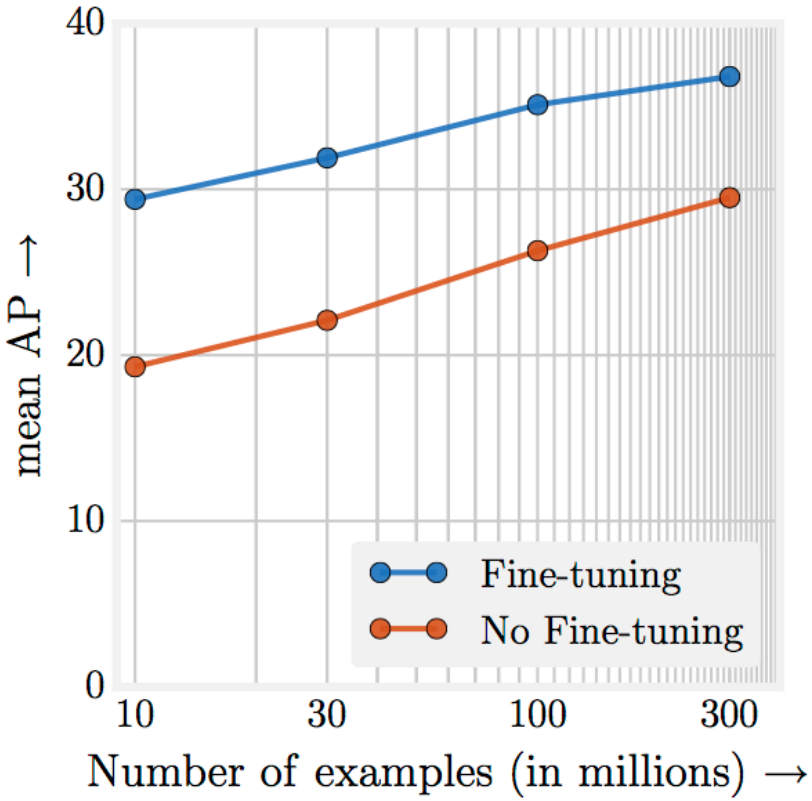
MICC - University of Florence - Italy

http://www.micc.unifi.it

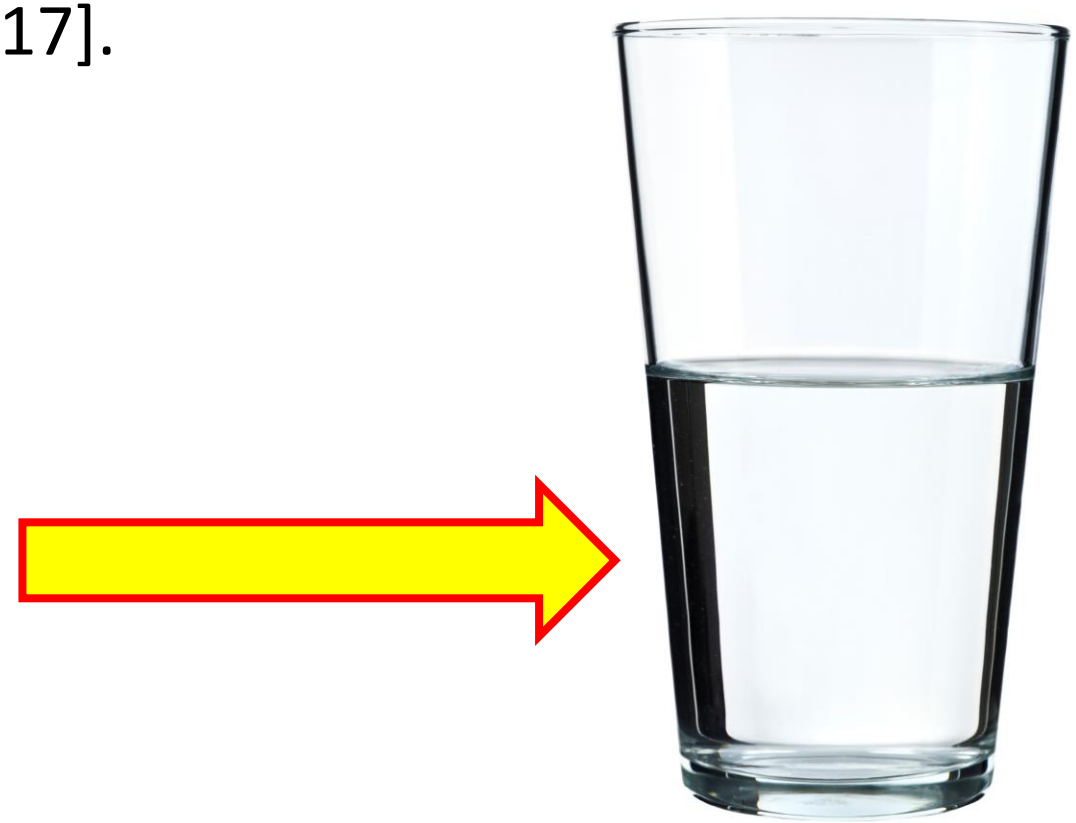



The effectiveness of data in Deep Learning

- Performance increases linearly with orders of magnitude of training data [Chen2017].



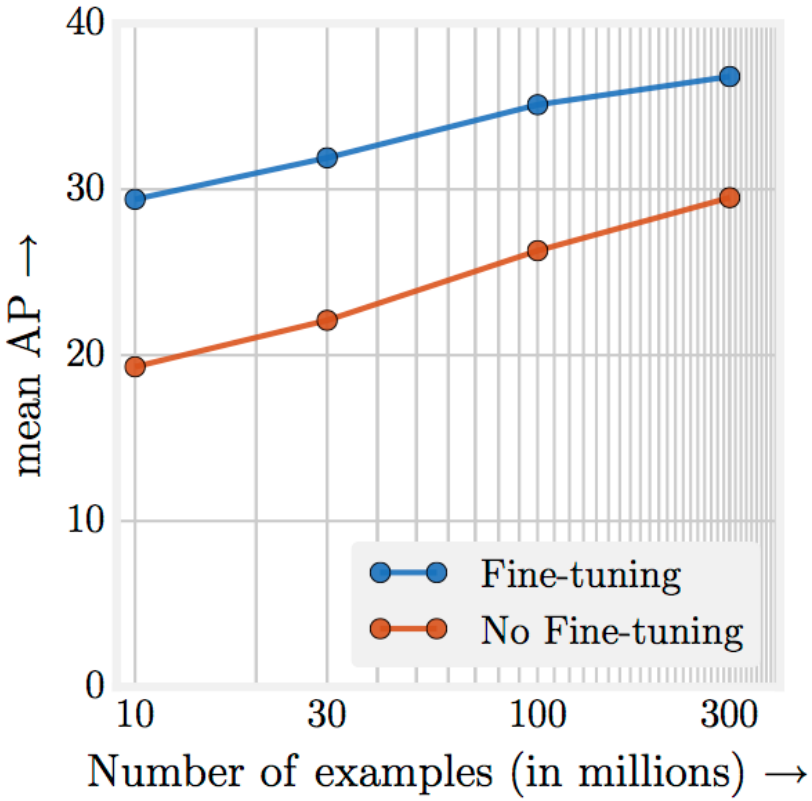
(Log scale)



[Sun2017: Revisiting the Unreasonable Effectiveness of Data ICCV2017]

However…

- Linear improvement in performance requires **exponential number** of labelled examples.



(Log scale)

[Sun2017: Revisiting the Unreasonable Effectiveness of Data ICCV2017]

The cost of annotation

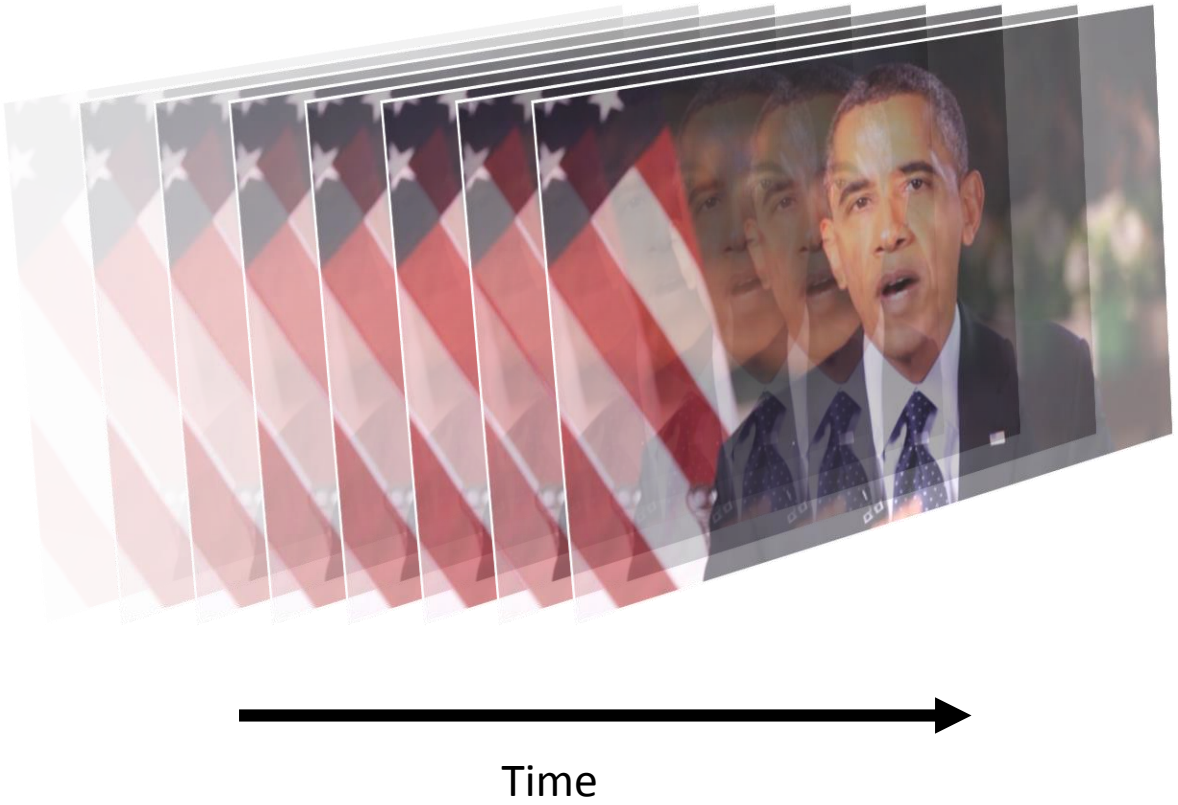
- The **cost of annotation** remains the most critical fact in Supervised Learning.

- Crowdsourcing...
  - 1M images with 1000 categories at 1 cent per question $10M.
  - ImageNet used several heuristics (e.g., hierarchy of labels) to reduce the space of questions, reducing the cost to the order of $100K

Learning from video streams

An attracting alternative:

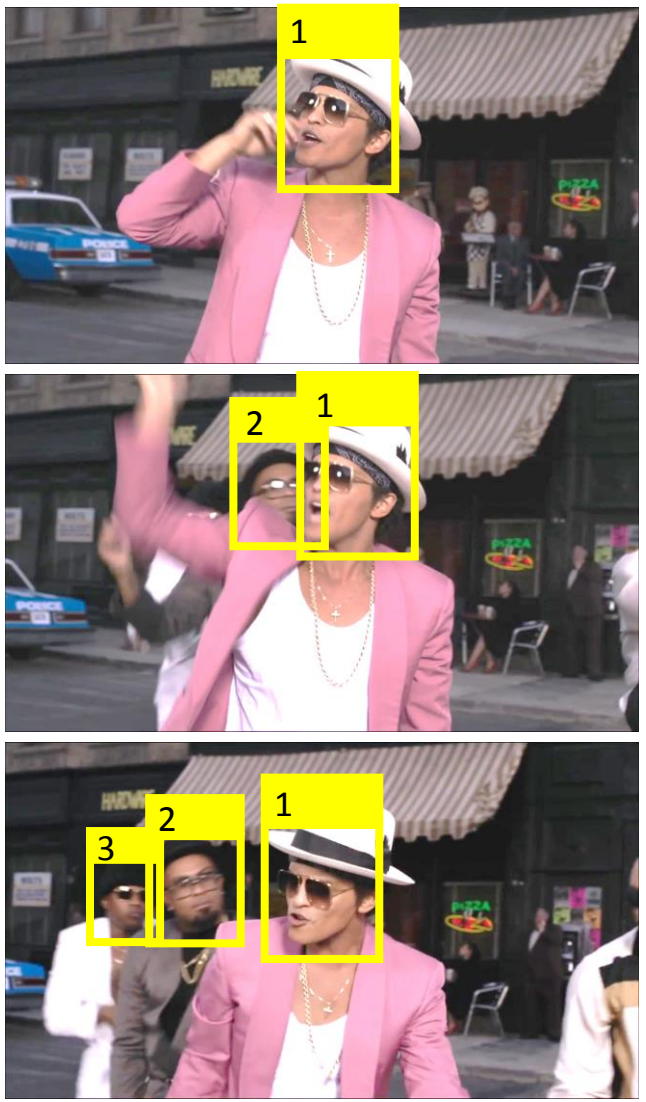
- learn objects appearance from **video streams** with no supervision, both exploiting
  - the **large quantity** of video available in the Internet and
  - the fact that **adjacent** video **frames** contain **semantically similar** information (weak supervision).



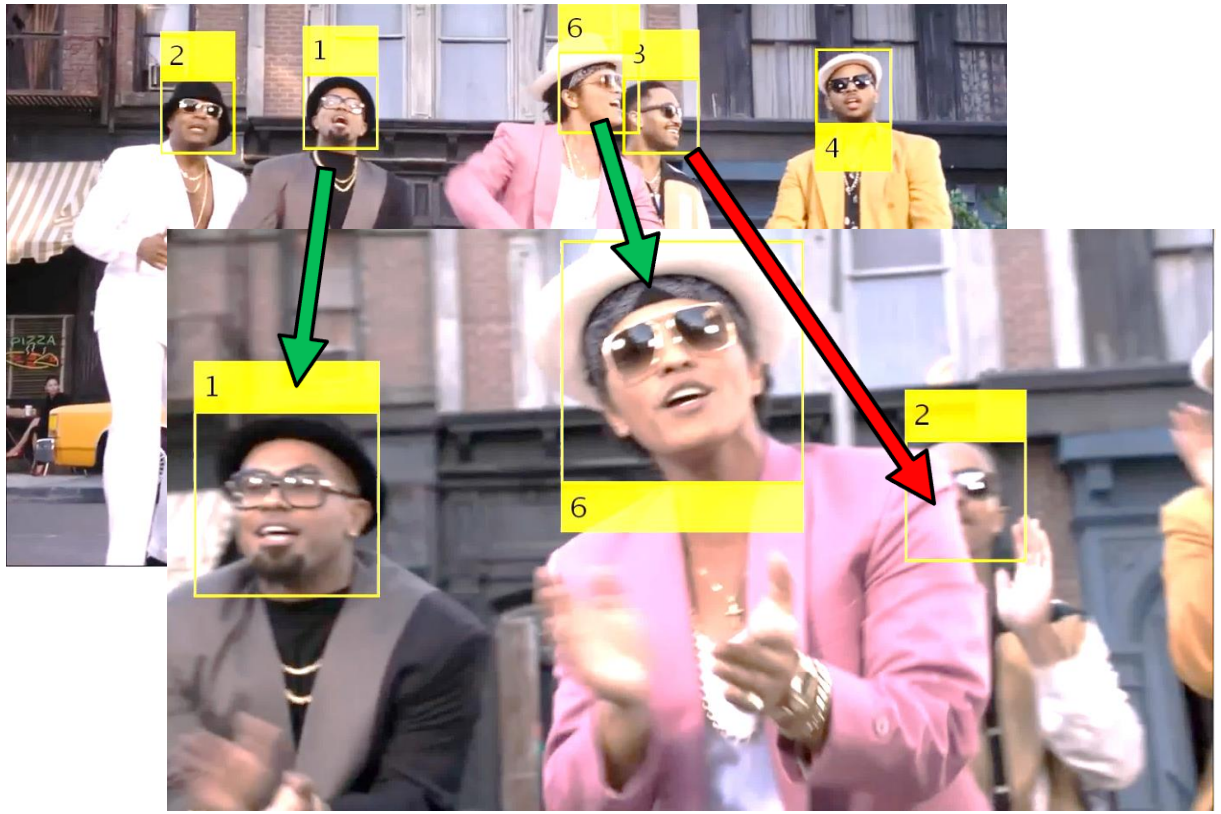
Practical Problem…

- Online **Open World** Face Recognition from video streams
  - It is not possible to predict a priori how many face objects to recognize (i.e. the number of **classes is unknown**).
  - The system must be able to **detect known/unknown** classes.
  - There are no labels.
  - The system must be able to **add** the detected **unknown classes** to the model (Open World).
  - The system cannot be retrained from scratch (it must be works forever).

- The problem appears to present a daunting challenge for deep learning (*catastrophic forgetting*).



Problem details...

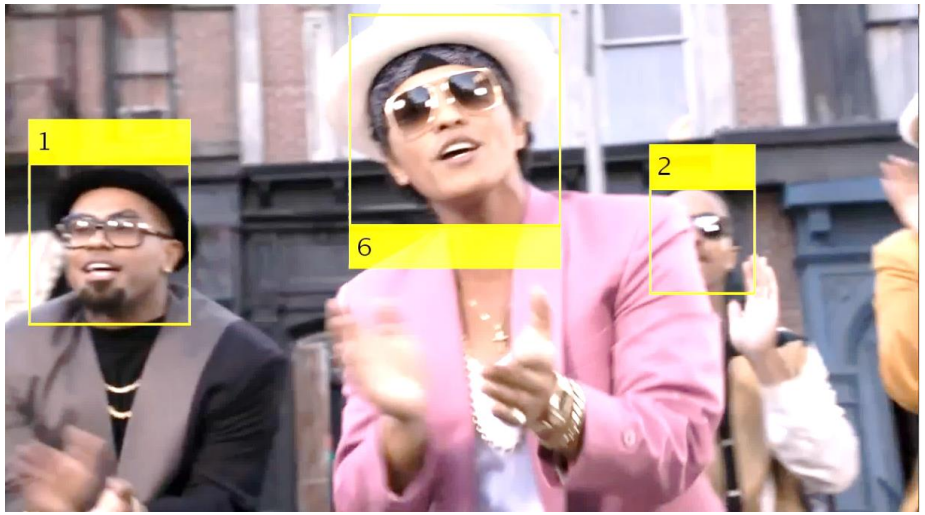


Unconstrained videos are typically made of shots

- New face identities...
- Wrong identity associations...
- False positives... (not a novel class)

Problem details

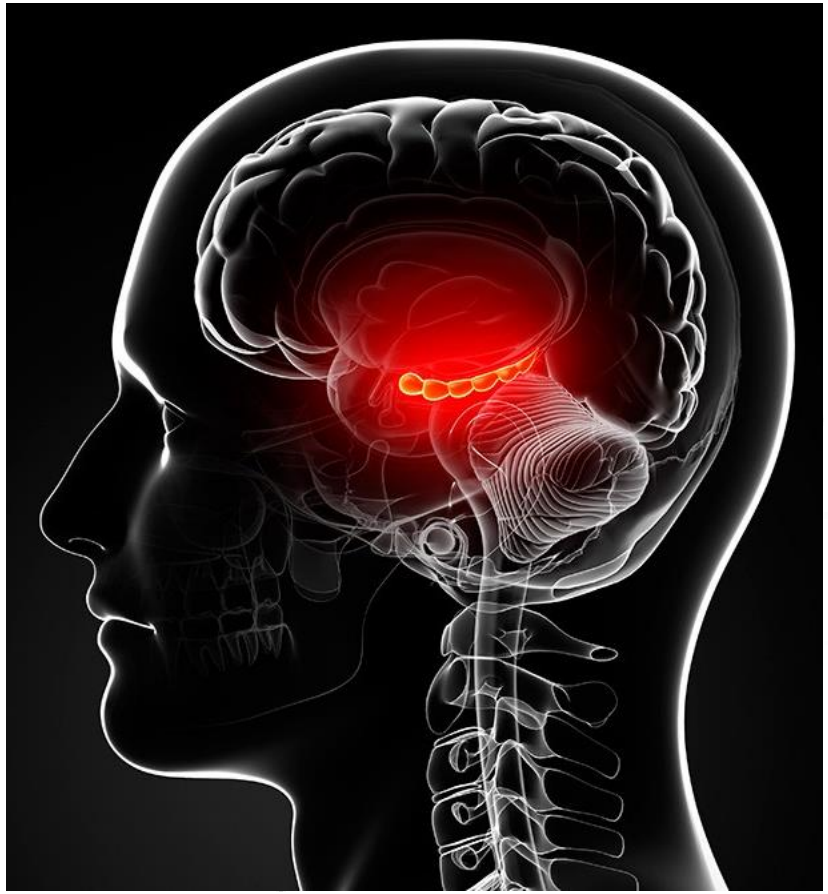
- The Learner operates in two steps.
  - First, it automatically labels the data in the next frame.
  - Second, it uses this labeled data to train the classifier.
- Errors may introduce noisy labels (wrong identities).
- **Noisy labels** may impair irreversibly the learning process as time advance.



Our solution: exploit a Memory module

- The appearance in video streams typically evolves over time:
  - **Data** can **no longer** be assumed as independent and identically distributed (**i.i.d.**)

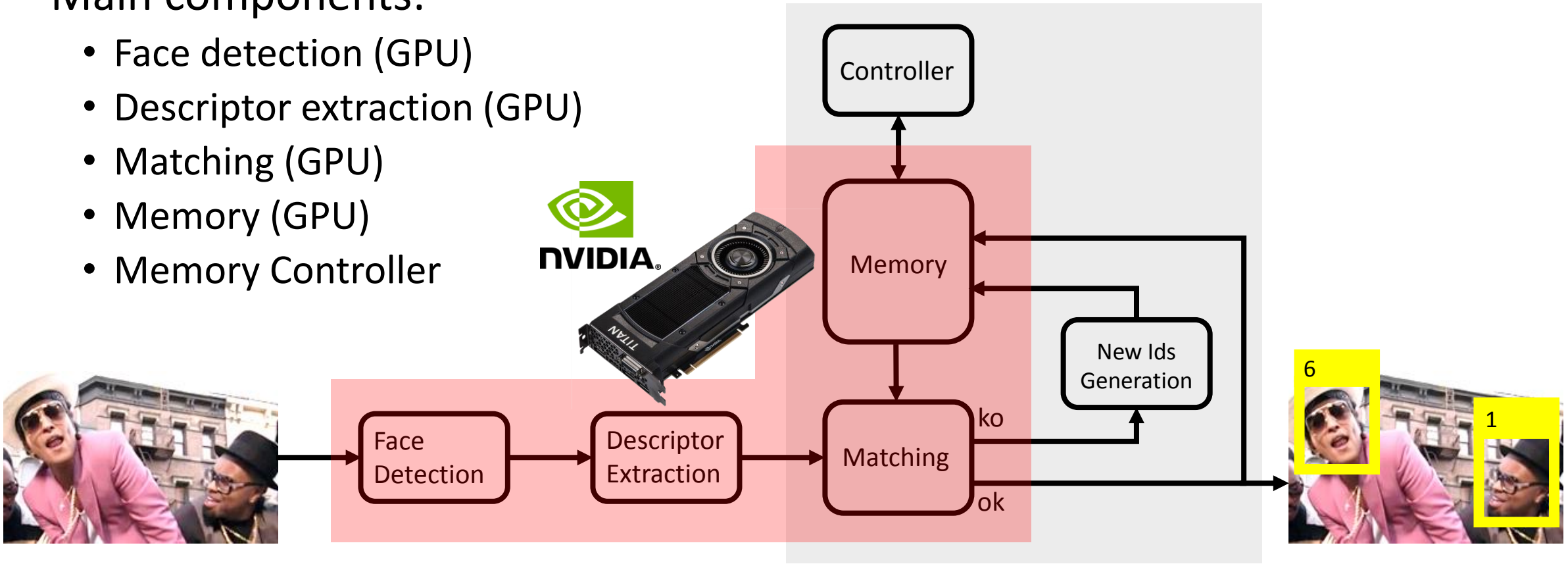
- **Store** the past experience in a **memory** module (i.e. Hippocampus) [Schaul2015].
  - If appearances are never forgotten (Infinite Memory), it is possible to limit the non stationary effects [Cornuéjols2006].
  - This also makes it possible to mix more and less recent information.



[Schaul2015: Prioritized Experience Replay]

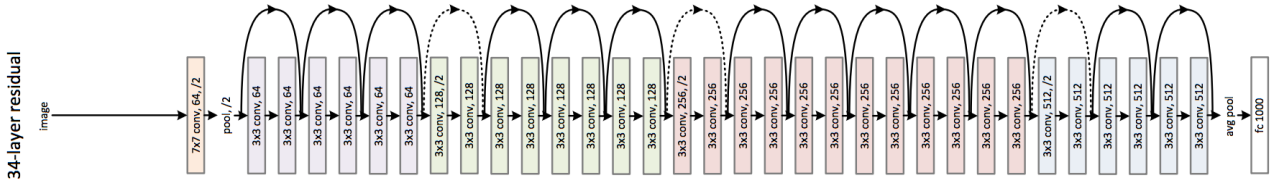
System Overview

- Main components:
  - Face detection (GPU)
  - Descriptor extraction (GPU)
  - Matching (GPU)
  - Memory (GPU)
  - Memory Controller

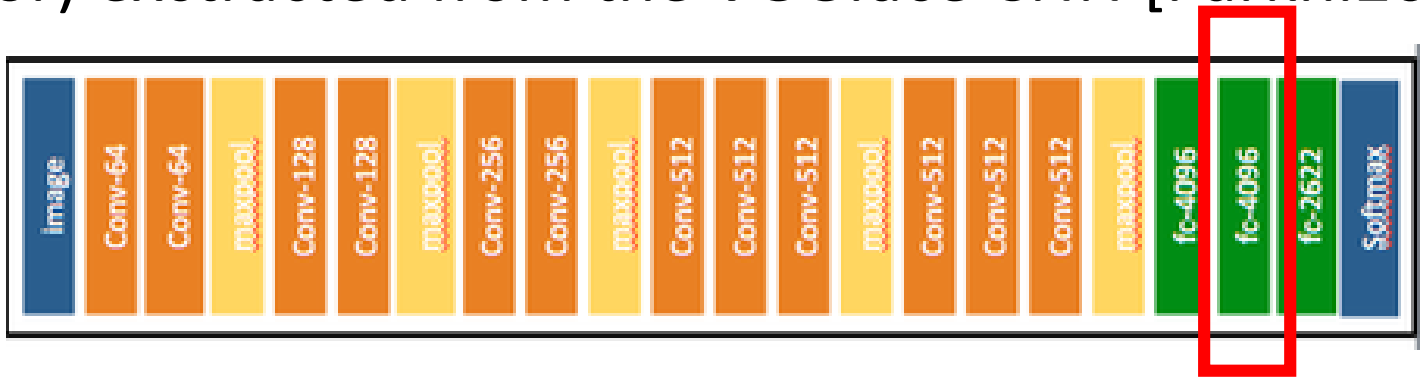


Face Dectection and Description

- Faces are **detected** using the Tiny Faces method [Peiyun2017]
  - The method uses a CNN with the ResNet101 architecure



- Detected faces are **represented** according CNN activations (the face descriptor) exctracted from the VGGface CNN [Parkhi2015]

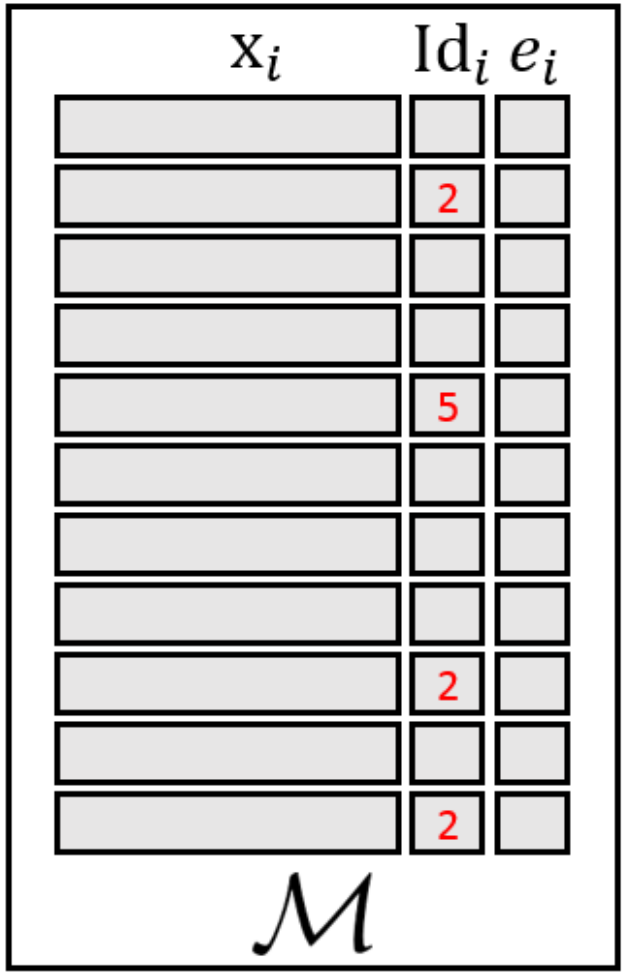


Main Idea: quick learning using Memory

- The memory module is used for fast learning and consists of the following triples:

$$\mathcal{M}(t) = \{(\mathbf{x}_i, \mathrm{Id}_i, e_i)\}_{i=1}^{N(t)}$$

- The **eligibility** e_i is a scalar quantity in [0,1] associated to each descriptor $\mathbf{x}_i$ (i.e. CNN activations)
  - It captures the **redundancy** of a descriptor with respect to the other descriptors in the memory.
  - Each descriptor has an associated identity Id_i.



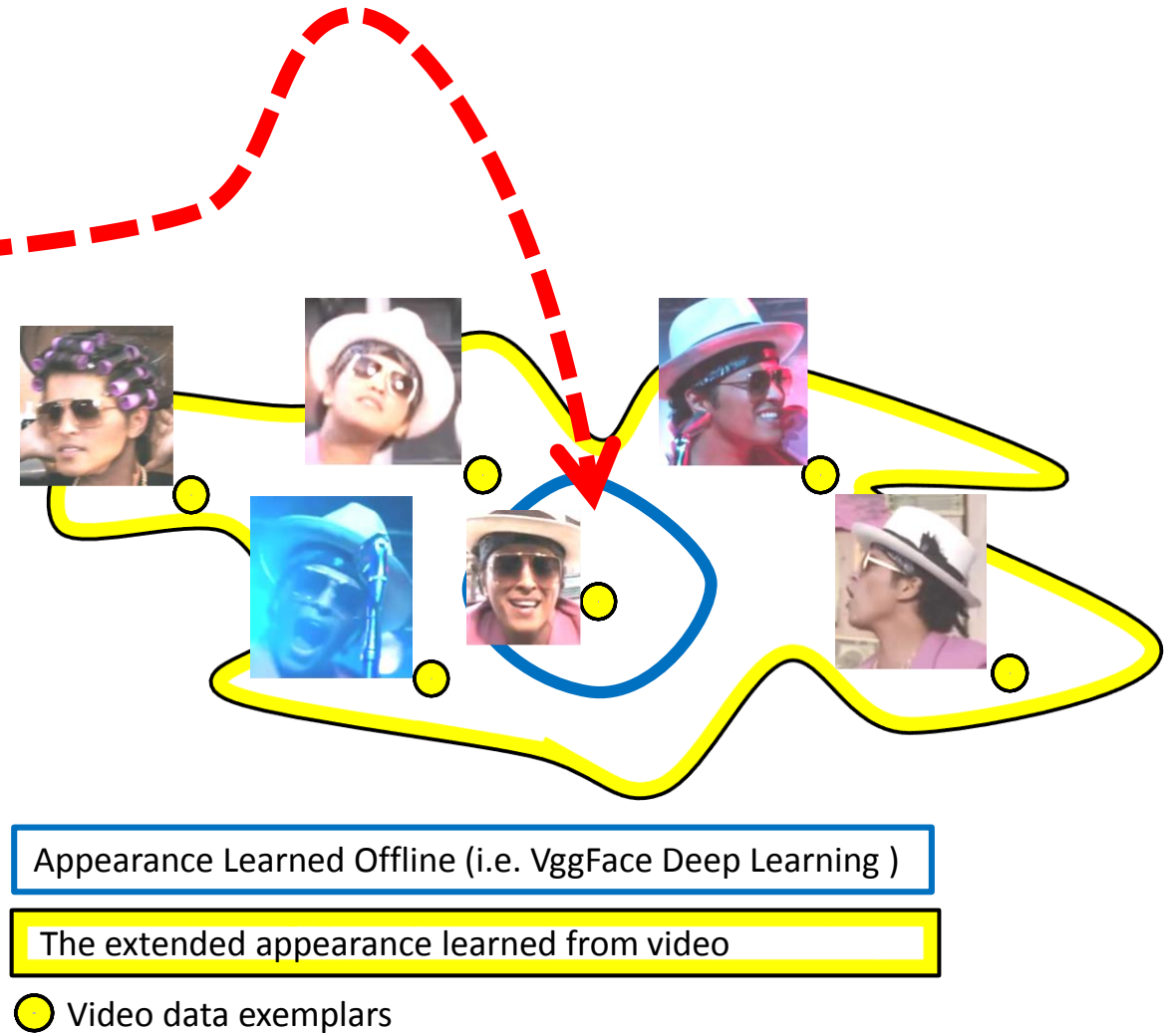
Intuition: Memory and Eligibilities

- Faces appearance model is extended using the video **exemplars** collected while tracking.
- To control **redundancy** the eligibilities e_i of **matching** descriptors are time updated according to:

$$e_i(t+1) = \eta_i \, e_i(t)$$

where η_i take into account descriptor distance (i.e. spatial redundancy).

- Descriptors are **removed** when their corresponding eligibilities e_i drops below a given threshold.
- The eligibility is:
  - Low for **ordinary** «events»
  - High for **rare** «events»
- **Unmatched** descriptors are **added** to the memory with a novel Id and *e*=1.

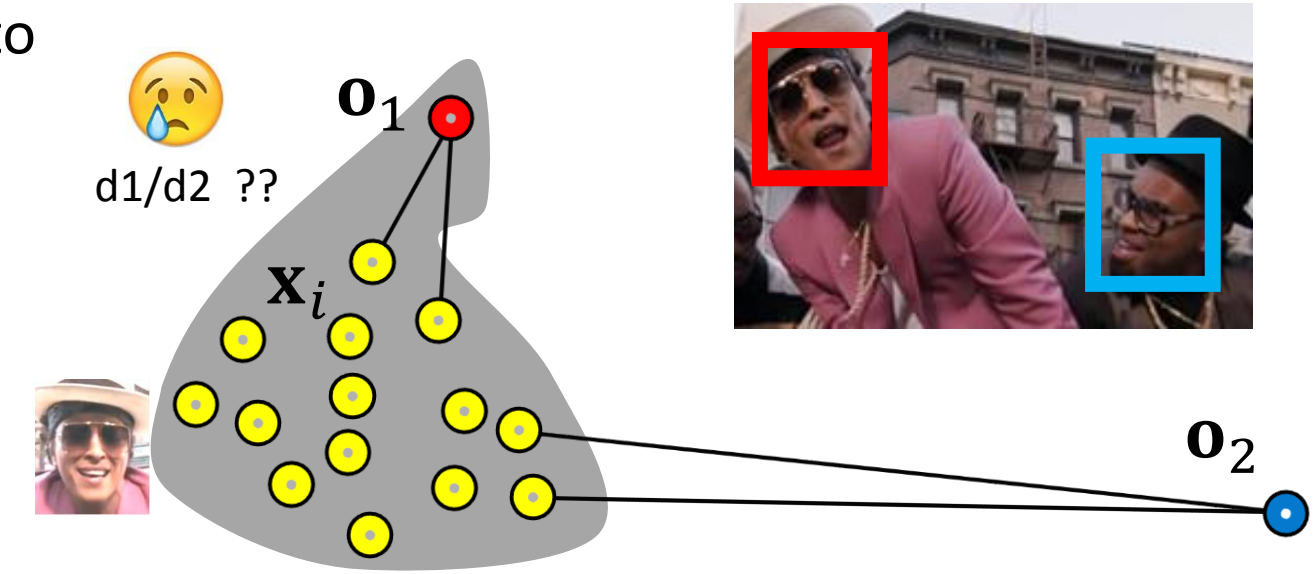


Discriminative Matching

- Video temporal coherence:
  - Faces in consecutive frames have little differences.
  - Similar descriptors will be stored in the memory (Repeated Temporal Structure).

- **Distance Ratio test**: compares the distance to the closest neighbor with the distance to the second closest neighbor.
  - If they are far apart (d1/d2<thresh): OK.

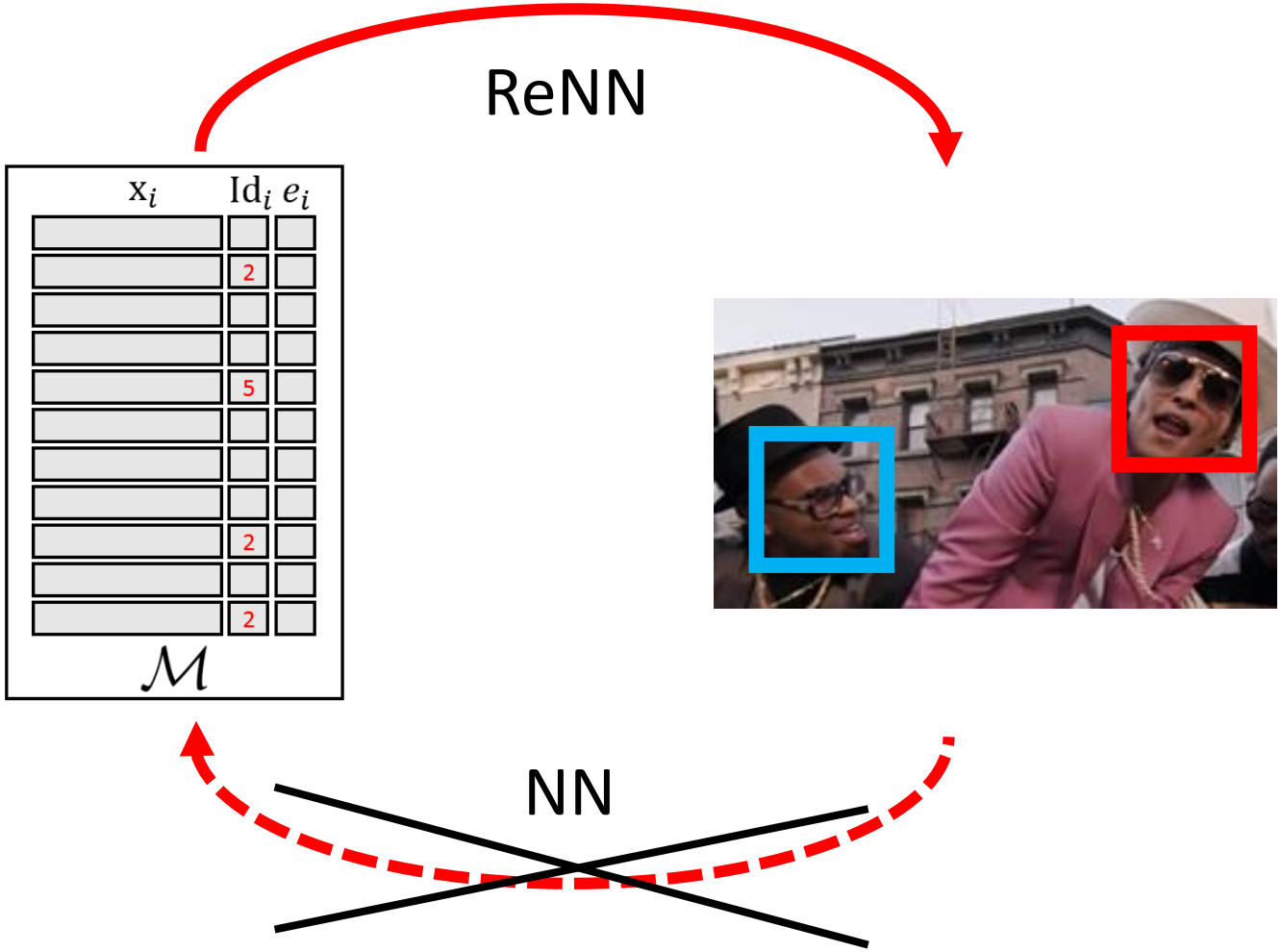
- If repeated structure distances are comparable, the discriminative match **cannot** be assessed.
  - This limit is solved using **Reverse Nearest Neighbor**(ReNN)



Repeated Temporal Structure (Memory)

Reverse Nearest Neighbour (ReNN)

- In ReNN **Roles** are **exchanged**
  - Each entry of the database is a query.
  - Faces in the current frame are the database.



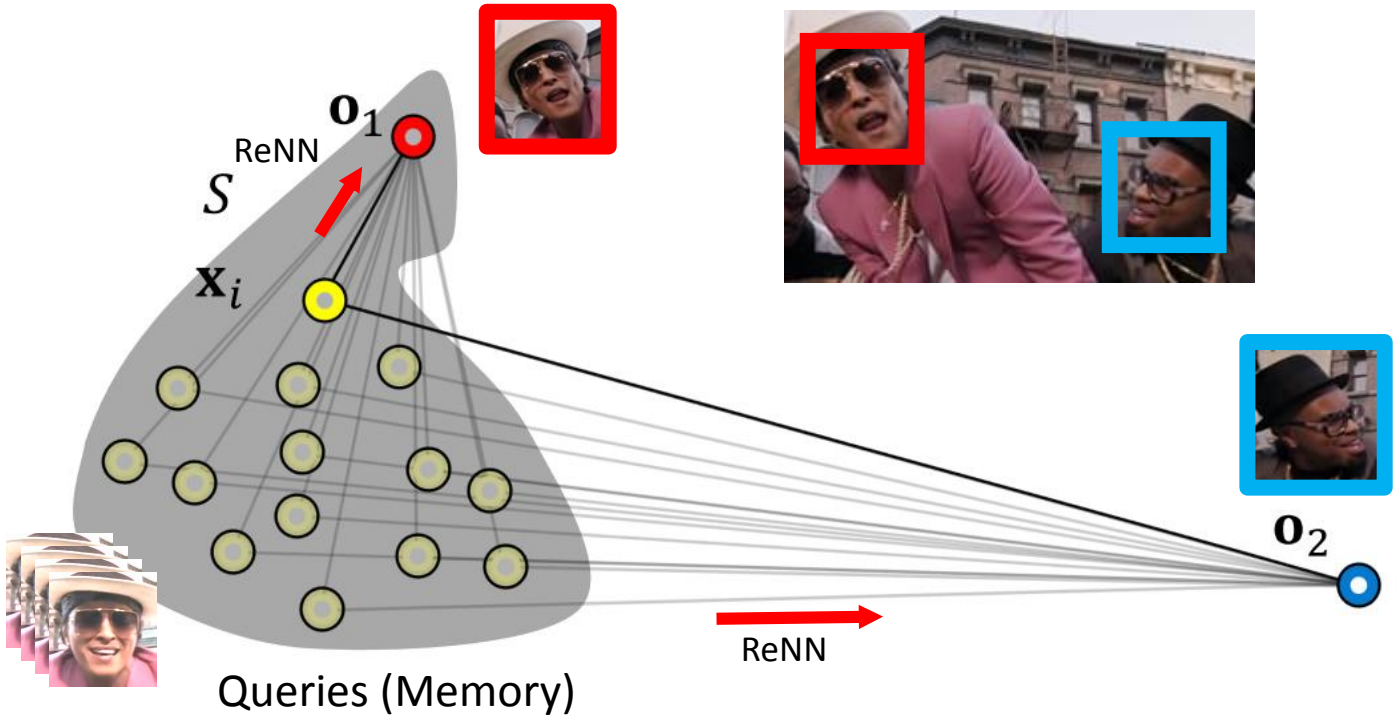
ReNN and distance ratio

- This strategy exploits discriminatively the **uniqueness of face** in the current frame.

- The other important advantage ReNN is that all the descriptors $\mathbf{x}_i$ of the repeated structure match with $\mathbf{o}_1$:

$$\{\mathbf{o}_1\} \leftrightarrow \{\mathbf{x}_i\}$$

- This allows the **automatic selection** of the descriptors that need to be **condensed** into a more compact representation.

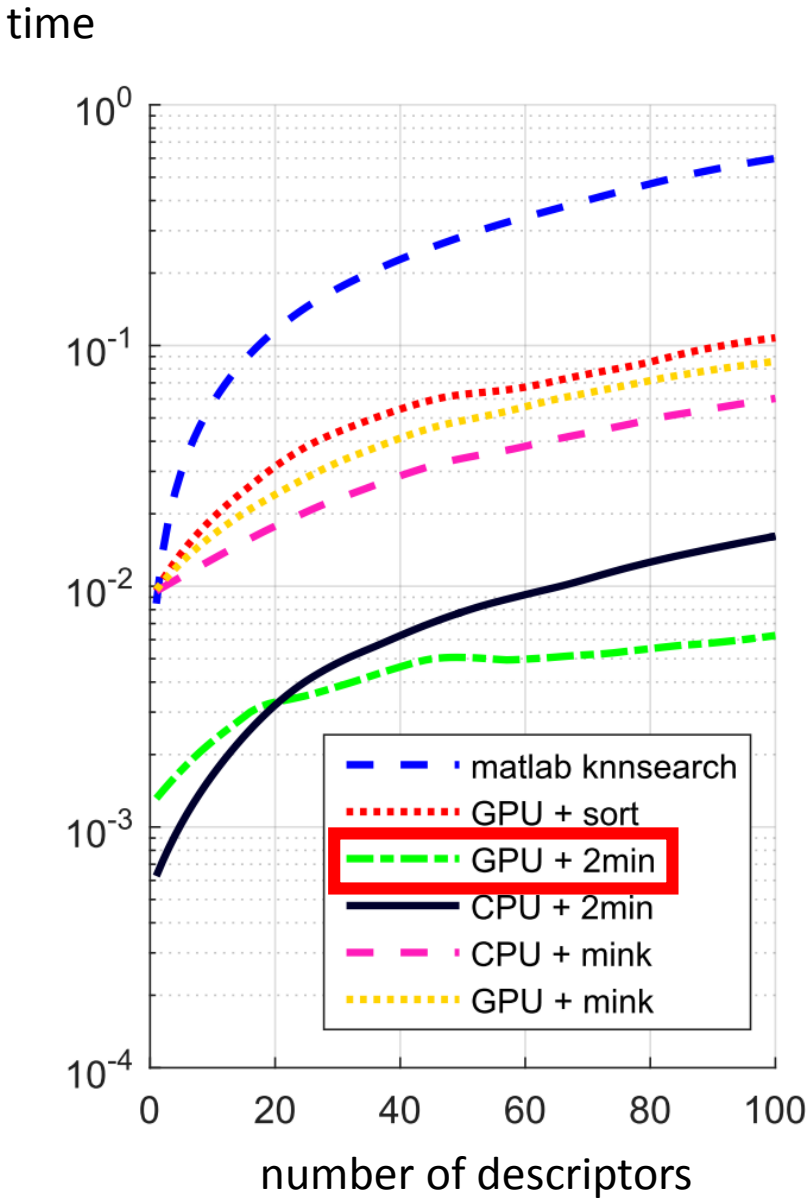


GPU based ReNN

- Reverse Nearest Neighbor under the distance ratio criterion can be **effectively accelerated** on the GPU.

- This is achieved using the *min* function twice in a GPUarray (Matlab, PyCuda).
  - Cuda Parallel Reduction is exploited.

- Complexity is almost constant as the number of descriptors in the memory increases (Nvidia Titan X Maxwell).



...

Asymptotic Stability

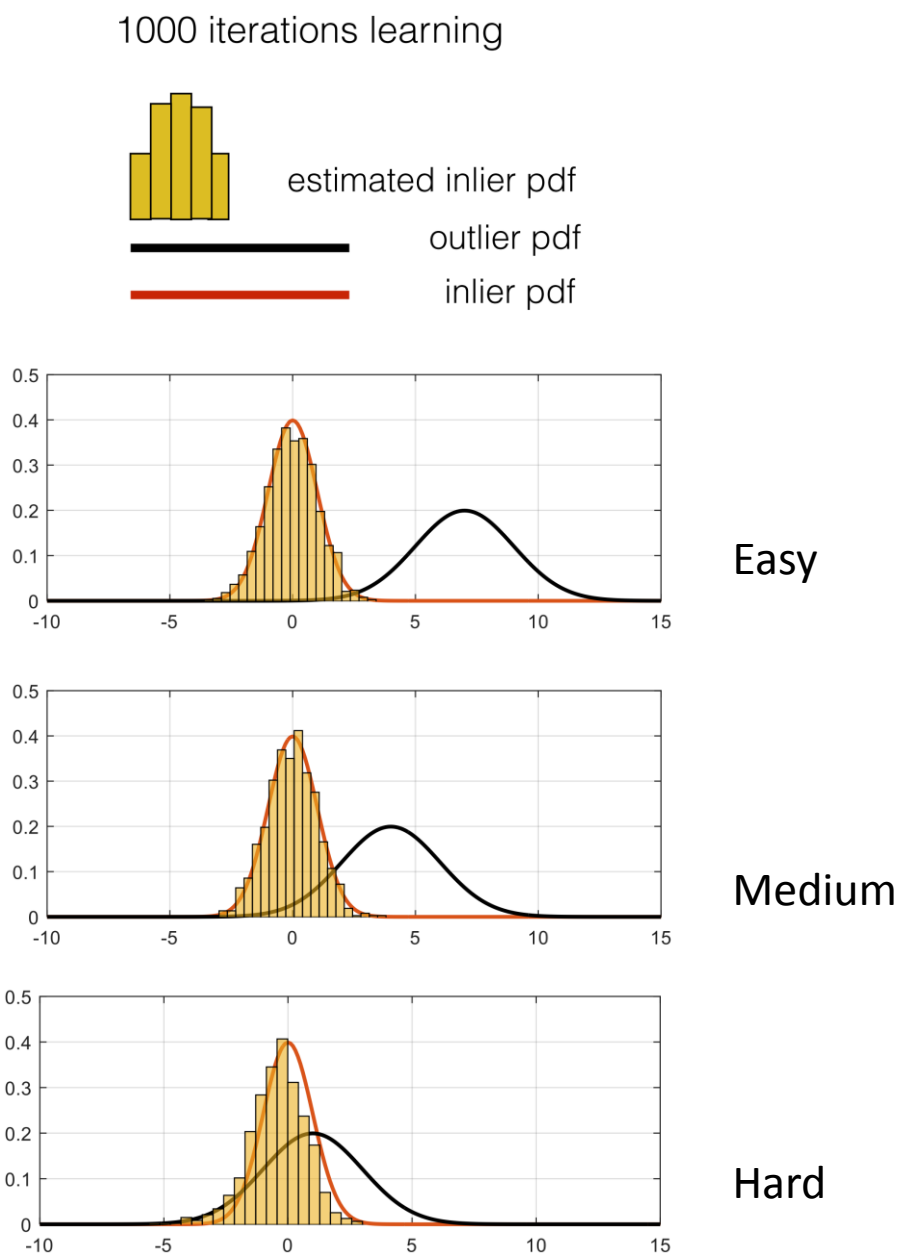
- Eligibility updating **stabilizes** around the pdf of each individual subject face.

- The eligibility updating rule:

$$e_i(t+1) = \eta_i \, e_i(t)$$

  is a **contraction** (i.e. η_i<1), it converges to its unique fixed point.

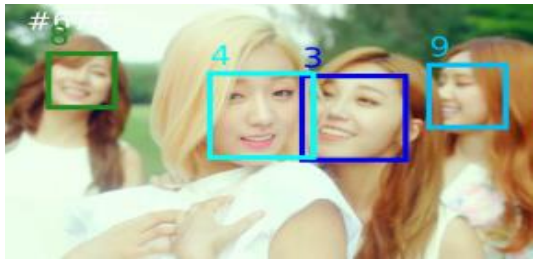
- Toy problem with increasing difficulty…



Experimental Results

- We used the **Music-dataset** [Zhang2016].
- **8 music videos** downloaded from YouTube with annotations of 3,845 face tracks

- **Big Ban Theory** 1° season (Ep1,2,...,6).
- 6 videos, about 23 minutes each.

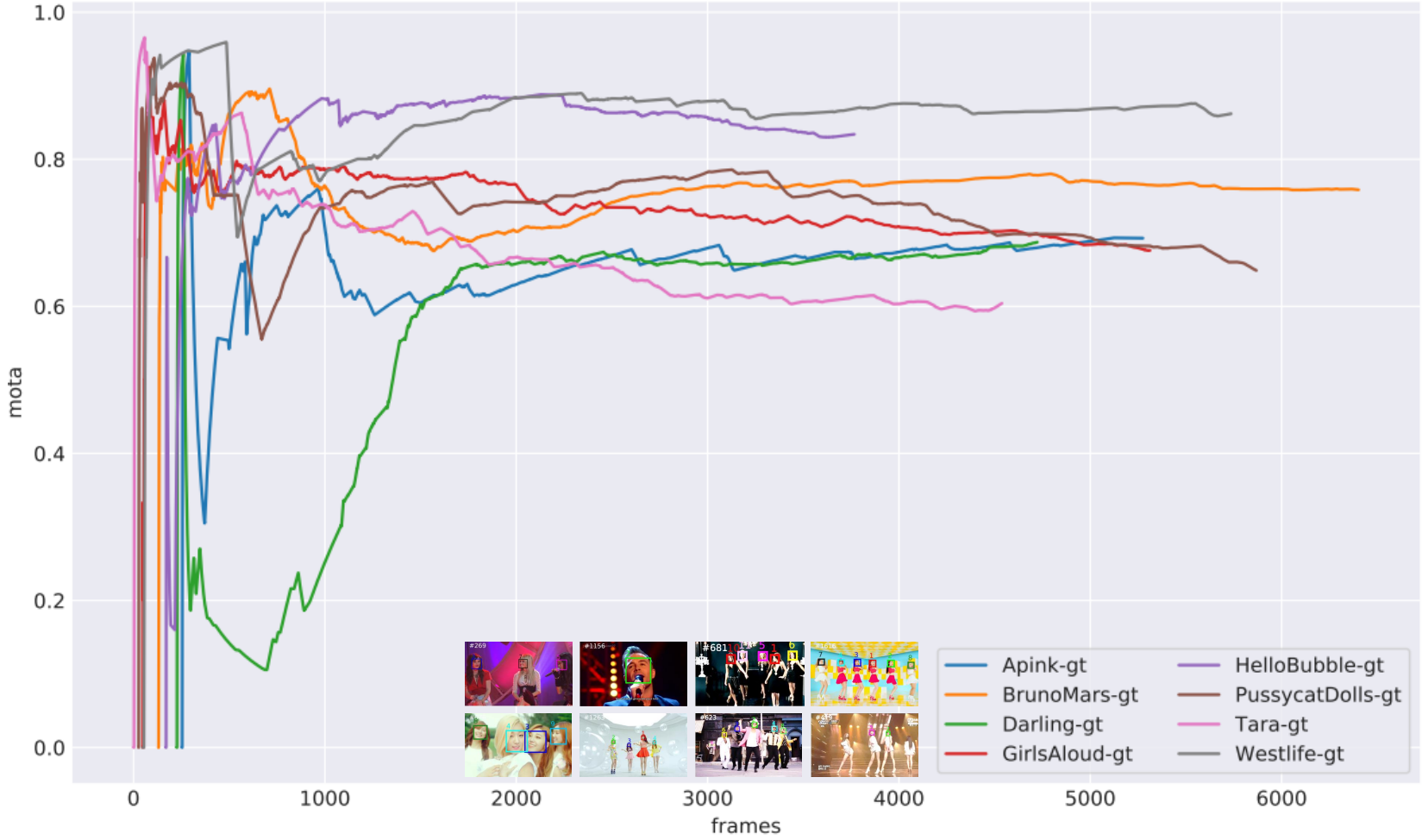


Experimental Results: drifting analisys

- Ground Truth as detections
- Accuracy:

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \cancel{\mathrm{FP}_t} + \mathrm{IDS}_t)}{\sum_t \mathrm{GT}_t}$$

- Fluctuations: no information at the beginning.
- Stability is common to all the videos.

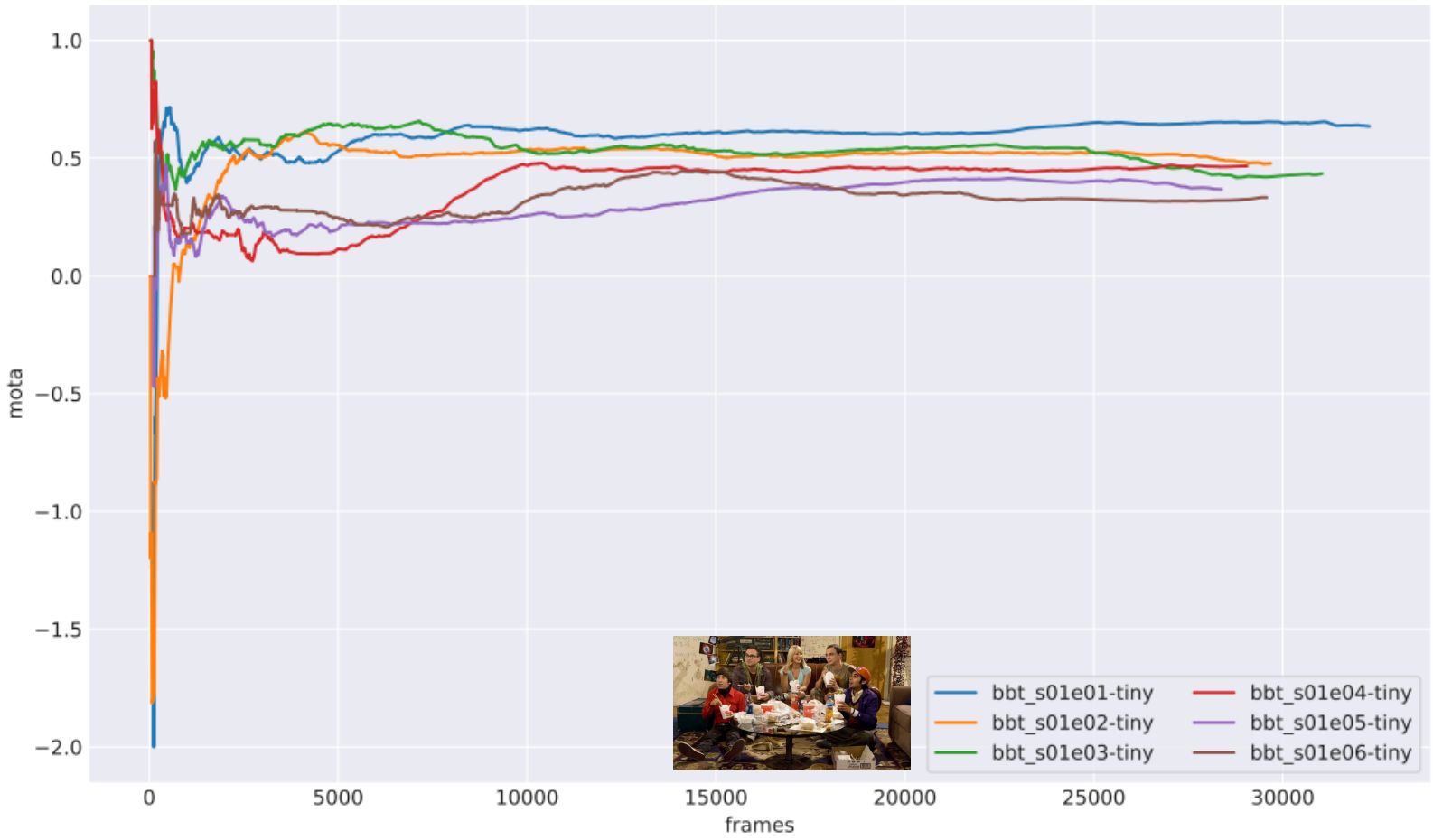


Experimental Results: drifting analisys

- Ground Truth as detections
- Accuracy:

$$\mathrm{MOTA} = 1 - \frac{\sum_t (\mathrm{FN}_t + \mathrm{FP}_t + \mathrm{IDS}_t)}{\sum_t \mathrm{GT}_t}$$

- Fluctuations: no information at the beginning.
- Stability is common to all the videos.



Comparison with Offline Methods

MUSIC DATASET								
Videos	Apink	Bruno Mars	Darling	Girls Aloud	Hello Bubble	Pussycat Dolls	T-ara	Westlife
HOG	0.20	0.36	0.19	0.29	0.35	0.28	0.22	0.27
AlexNet	0.22	0.36	0.18	0.30	0.31	0.31	0.25	0.37
Pre-trained	0.29	0.50	0.24	0.33	0.34	0.31	0.31	0.37
VGG-Face	0.24	0.44	0.20	0.31	0.29	0.46	0.23	0.27
Siamese	0.48	0.88	0.46	0.67	0.54	0.77	0.69	0.54
Triplet	0.60	0.83	0.49	0.67	0.60	0.77	0.68	0.52
SymTriplet	**0.72**	0.90	0.70	0.69	**0.64**	0.78	0.69	0.56
MuFTiR-tiny	0.51	**0.96**	**0.73**	**0.89**	0.59	**0.97**	**0.72**	**0.98**

Scores are based on **Purity.** Purity is a measure of the extent to which clusters contain a single class.

BIG BANG THEORY						
Episodes	BBT01	BBT02	BBT03	BBT04	BBT05	BBT06
HOG	0.37	0.32	0.38	0.35	0.29	0.26
AlexNet	0.47	0.32	0.45	0.35	0.29	0.26
Pre-trained	0.62	0.72	0.73	0.57	0.52	0.52
VGG-Face	0.91	0.85	0.83	0.54	0.65	0.46
Siamese	0.94	0.95	0.87	0.74	0.70	0.70
Triplet	0.94	0.95	0.92	0.74	0.68	0.70
SymTriplet	0.94	0.95	0.92	0.78	0.85	0.75
MuFTiR-tiny	**0.98**	**0.98**	**0.98**	**0.85**	**0.98**	**0.94**

Comparison with Offline Methods

BBT_S01E01				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD [1]	Offline	1	-16.3	74.8
ADMM [2]	Offline	323	42.5	64.0
IHTLS [3]	Offline	312	45.7	64.0
Pre-Trained [4]	Offline	171	41.9	73.3
mTLD2 [1]	Offline	223	58.4	73.8
Siamese [4]	Offline	144	69.0	73.7
Triplet [4]	Offline	164	69.3	73.6
SymTriplet [4]	Offline	156	72.2	73.7
MuFTiR-tiny	**Online**	36	51.6	70.1
MuFTiR-tiny	Offline	34	51.6	70.1

BRUNOMARS				
Method	Mode	IDS ↓	MOTA ↑	MOTP ↑
mTLD [1]	Offline	35	-8.7	65.3
ADMM [2]	Offline	428	50.6	85.7
IHTLS [3]	Offline	375	52.7	85.8
Pre-Trained [4]	Offline	151	48.3	88.0
mTLD2 [1]	Offline	278	52.6	87.9
Siamese [4]	Offline	126	56.7	87.8
Triplet [4]	Offline	126	56.6	87.8
SymTriplet [4]	Offline	105	56.8	87.8
MuFTiR-dpm	**Online**	78	4.5	61
MuFTiR-tiny	**Online**	420	48.8	65.5

Not too far, but online

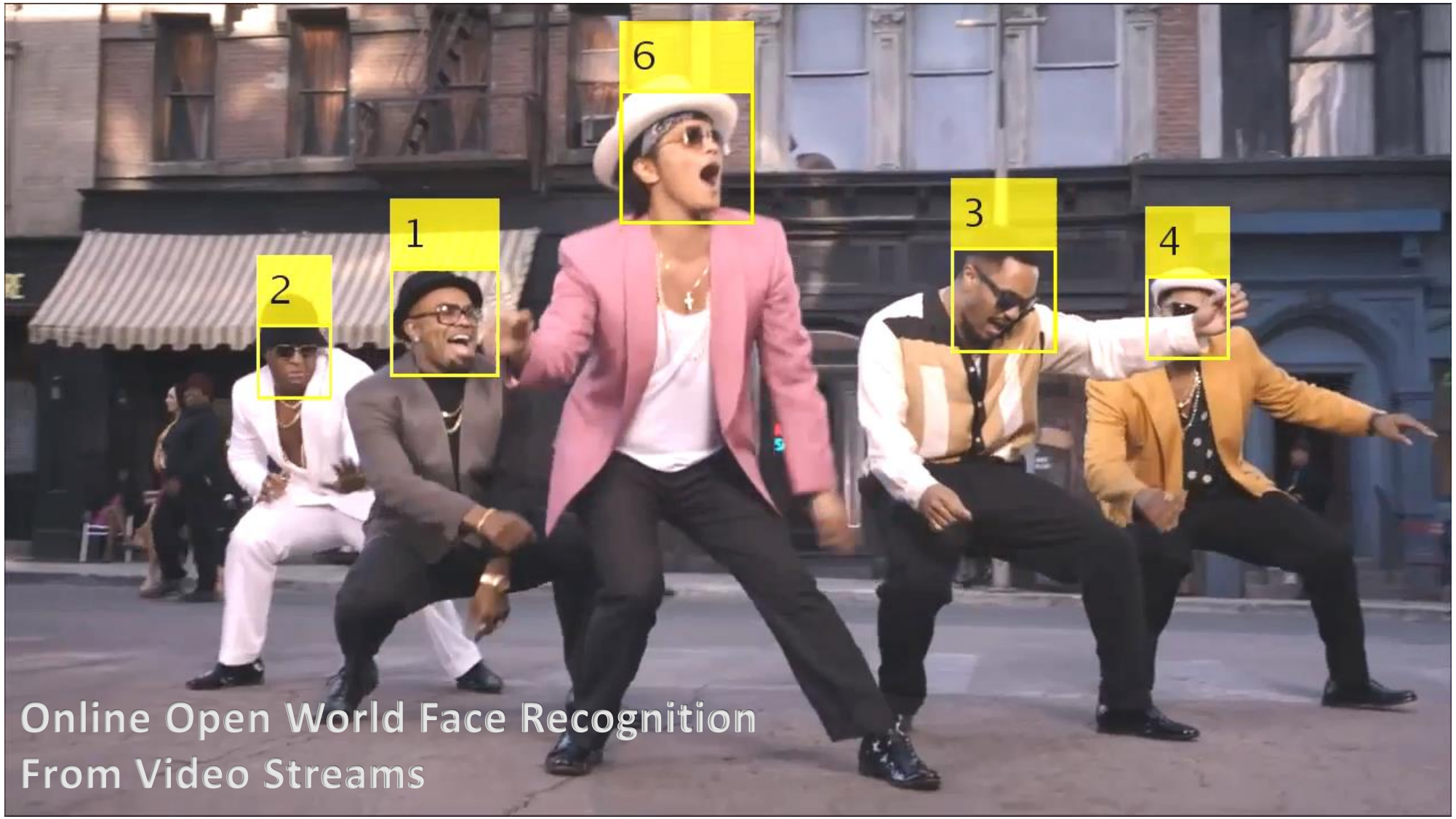
Bruno Mars

[1] Z. Kalal, K. Mikolajczyk, J. Matas, *Tracking-learning-detection*, IEEE TPAMI 2012.

[2] C. Dicle, O. I Camps, M. Sznaier, *The way they move: Tracking multiple targets with similar appearance*, *IEEE ICCV*,2013

[3] M. Ayazoglu, M. Sznaier, O.I Camps, *Fast algorithms for structured robust principal component analysis*, *CVPR 2012*

[4] S. Zhang,Y. Gong, ,J-B. Huang, ,J-J Wang, N. Ahuja, M-H Yang, *Tracking persons-of-interest via adaptive discriminative features*, *ECCV* 2016



Link: https://youtu.be/6S7D6Dgmt3Y

Qualitative results

Conclusion

- Online Open World Face Recognition From Video Streams
  - Fully implemented on a GPU
  - Wide applicability: Enables face recognition with auto enrollment of subjects

- Applicability in other contexts:
  - Person Detector – Person Descriptor
  - Car detector – Car Descriptor
  - Traffic Signal Detector – Traffic Signal Descriptor
  - …

- Future developments:
  - Exploit the data diversity in the memory to train online a Deep CNN.