Self-Supervised Learning of Face Representations for Video Face Clustering

Vivek Sharma, Makarand Tapaswi, Saquib Sarfraz, and Rainer Stiefelhagen

https://github.com/vivoutlaw/SSIAM

Karlsruher Institut für Technologie

UNIVERSITY OF TORONTO

sharma.vivek@live.in

Motivation

- To learn discriminative face representation via self-supervision
  - Small intra-person-distance and large inter-person-distance.

- This will benefit potential applications in
  - Video understanding, video summarization, content-based indexing & retrieval
  - Automatic reasoning about multimedia content.

Introduction

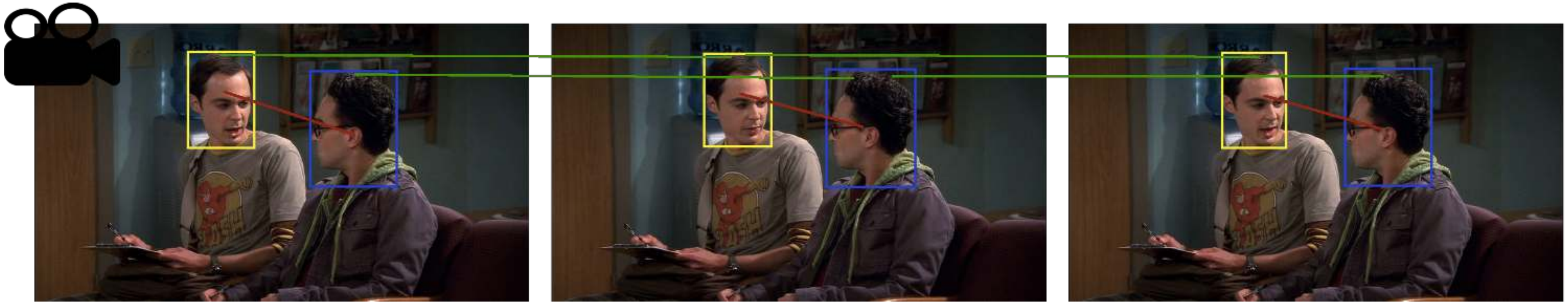
- Video face clustering is hard.
  - Discriminative features help.

Blurred Illumination Perspective

- Most prior works utilize: must-link and cannot-link information.

- Difficult to train from scratch (require lots of training data), typically handled by net surgery:
  - Fine-tuning
  - Use of additional embedding's on the features from the last layer
  - Both
- We propose two self-supervised discriminative methods.
  - Self-supervised Siamese network (SSiam)
  - Track-supervised Siamese network (TSiam)
- We evaluate on three video face clustering datasets.

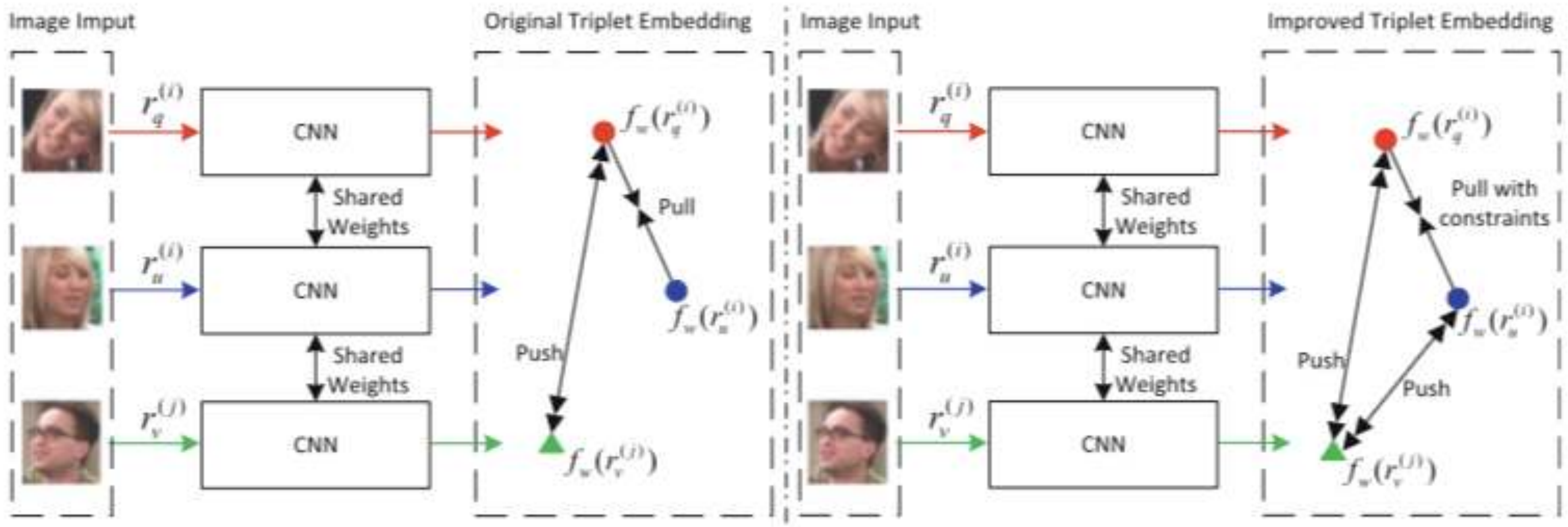
Temporal Constraints



- Video constraints: must-link and cannot-not link.

Everingham et al.: “Hello! My name is ... Buffy” Automatic Naming of Characters in TV Video. In: BMVC. (2006)
[**ULDML**] Cinbis et al.: Unsupervised Metric Learning for Face Identification in TV Video. In: ICCV. (2011)
Tapaswi et al.: “Knock! Knock! Who is it?” Probabilistic Person Identification in TV-Series. In: CVPR. (2012)

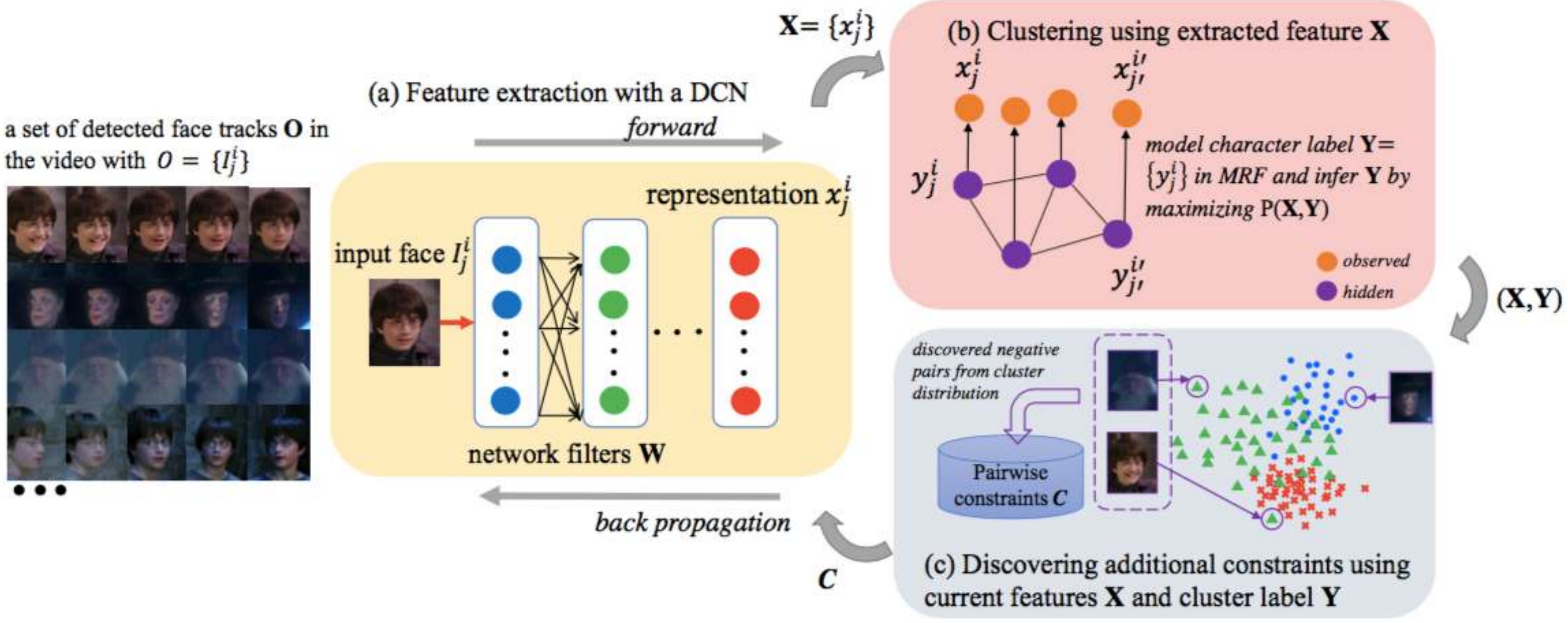
Related Work



- Link-constrained based improved triplet loss

[**Imp-Triplet**] Zhang et al.: "Deep metric learning with improved triplet loss for face clustering in videos." *Pacific Rim Conference on Multimedia*. Springer. (2016)

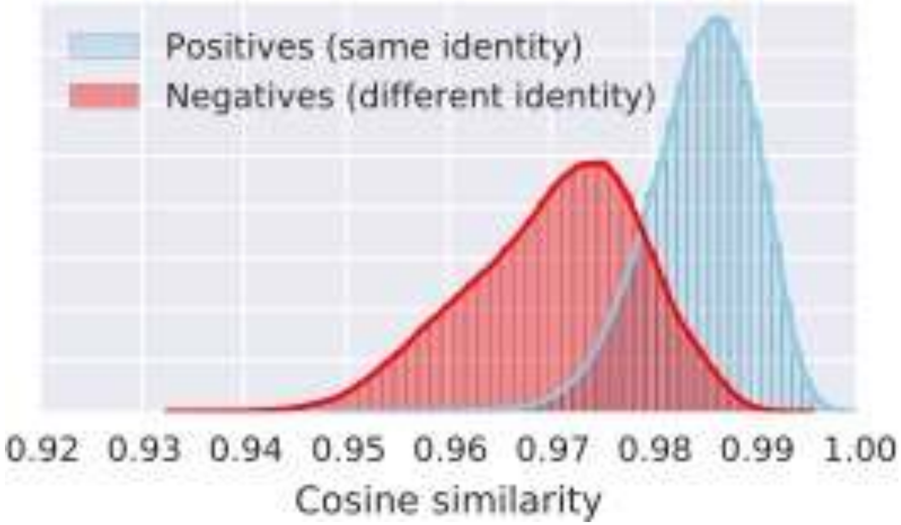
Related Work



- Based on loss function or MRF modeling.

[**JFAC**] Zhang et al.: "Joint face representation adaptation and clustering in videos." In *European conference on computer vision*, pp. 236-251. Springer. (2016)

Related Work: Pseudo-RF



- This is especially in light of CNN face representations that are very similar even across different identities.
- We see a large overlap between the cosine similarity distributions of positive (same id) and negative (across id) track pairs.

Pseudo-RF: Yan et al.: Negative pseudo-relevance feedback in content-based video retrieval. In *ACM MM*, 2003

Self-supervised Siamese network (**SSiam**)

- Does not need tracks or temporal information.
- Mechanism for mining positive and negative examples automatically.
- Compute a distance matrix (i.e. ranking) over random subset per iteration
  - Use the farthest positives and closest negatives pairs sets as labels.

Distance Matrix

	1	2	3	4
1	0	0.1	0.5	0.7
2	0.1	0	0.9	0.4
3	0.5	0.9	0	0.3
4	0.7	0.4	0.3	0

Pairs

	1	2	3	4
1	1-1	1-2	1-3	1-4
2	2-1	2-2	2-3	2-4
3	3-1	3-2	3-3	3-4
4	4-1	4-2	4-3	4-4

Sort distance row-wise

1	0	0.1	0.5	0.7
2	0	0.1	0.4	0.9
3	0	0.3	0.5	0.9
4	0	0.3	0.4	0.7

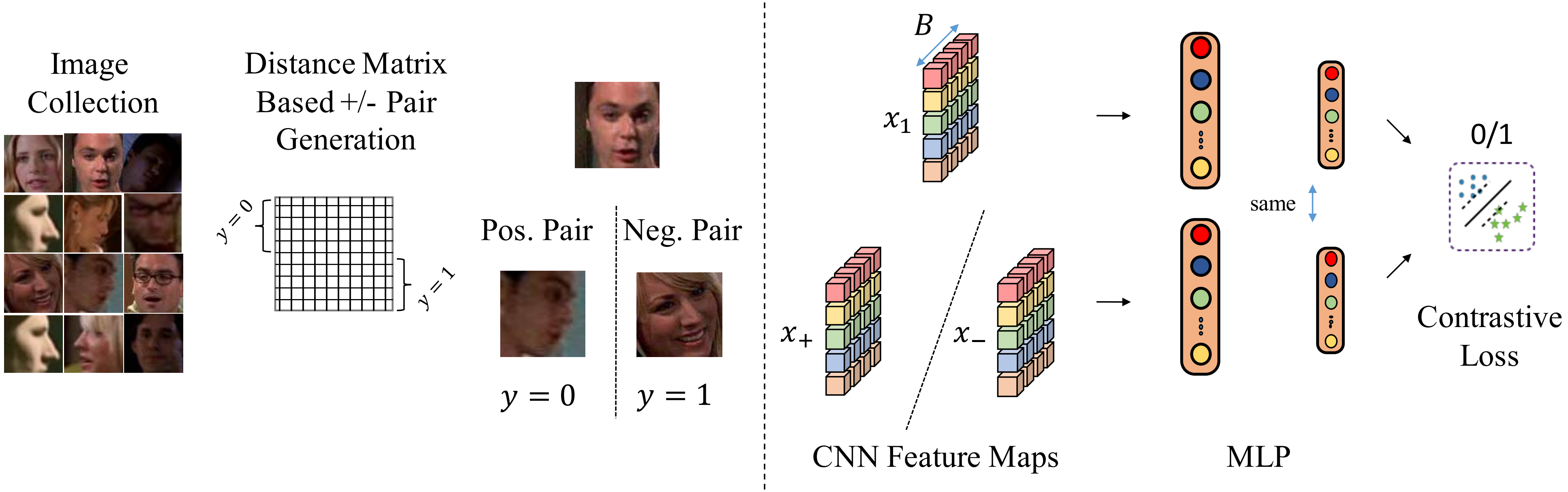
1	1-1	1-2	1-3	1-4
2	2-2	2-1	2-4	2-3
3	3-3	3-4	3-1	3-2
4	4-4	4-3	4-2	4-1

- Choose positive pairs from the second column with the largest distance, and negative pairs from the last column with the smallest distance.
- These pairs are semi-hard.
- Example
  - 1 positive (3-4) and 1 negative pair (1-4)

Most dissimilar

Most similar

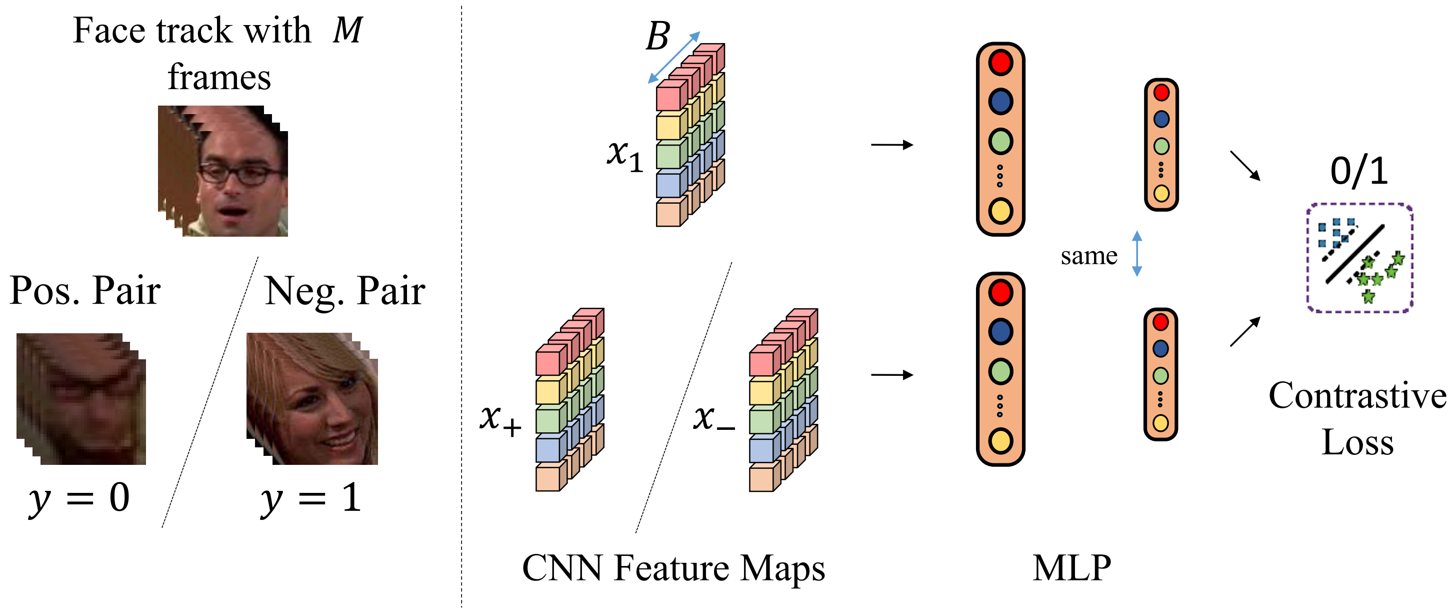
SSiam



SSiam selects hard pairs: farthest positives and closest negatives using a ranked list based on distance matrix. *B* corresponds to batch.

Track-supervised Siamese networks (**TSiam**)

- Use temporal information (must-link/cannot-link).
- Also include negative pairs for singleton tracks
  - based on track-level distances (computed on base features)
  - randomly sample frames from the farthest F = 25 tracks.



Evaluation

- We present our evaluation on three challenging datasets.
  - Buffy the Vampire Slayer (BF) (season 5, episodes 1 to 6)
  - Big Bang Theory (BBT) (season 1, episodes 1 to 6)
  - Harry Potter 1 Movie (ACCIO)

		This work		Previous work
Datasets	#Cast	#TR (#FR)	LC/SC (%)	#TR (#FR)
BBT0101	5	644 (41220)	37.2 / 4.1	182 (11525)
BF0502	6	568 (39263)	36.2 / 5.0	229 (17337)
ACCIO	36	3243 (166885)	30.93/0.05	3243 (166885)

- Metrics
  - Clustering acc. for *BBT*, *BF*
  - BCubed, P, R, F1 for *ACCIO*

Implementation details

- We extract VGGFace2 features. The features are of 2048 Dimensions.
- Siamese network. Fully-connected neural network (2048 $\rightarrow$ 512 $\rightarrow$ 2). We extract the feature representations of 512D for clustering.

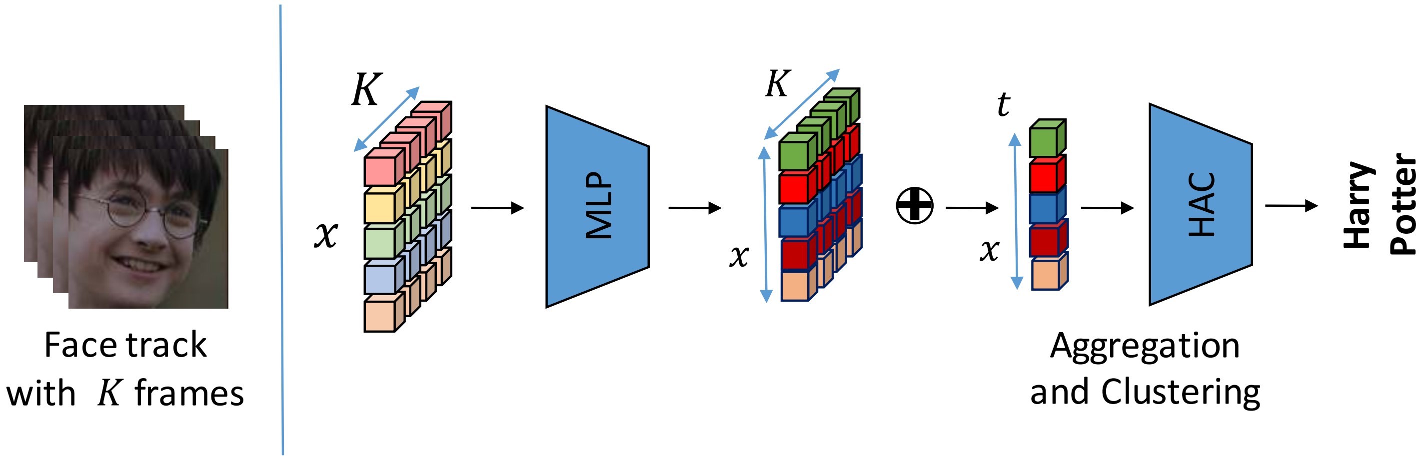
Cao, et al..: VGGFace2: A Dataset for Recognising Faces across Pose and Age. In: FG. (2018)

SSiam and TSiam labels mining

- For SSiam,
  - We use a random subset of size B = 3000
  - Choose 2K: positive and negative pairs, K = 64.
  - Higher values of B did not improve.
- For TSiam, we mine 2 positive and 4 negative pairs for each frame.

Testing Setup

- Extract features from base network and trained MLP: SSiam or TSiam.
- Perform clustering via HAC



TSiam, impact of singleton tracks

	TSiam		# Tracks		
Dataset	w/o Single (FG_Best)	Ours	Total	Single	Co-oc
BBT-0101	0.936	**0.964**	644	331	313
BF-0502	0.849	**0.893**	568	395	173

- Ignoring singleton tracks leads to significant performance drop.
- Approx. 50-70% tracks are singleton and ignoring them lowers accuracy by 4%.

FG_Best: Datta et al.: Unsupervised learning of face representations. In *FG*. IEEE, 2018

SSiam, comparison to pseudo-RF

- In Pseudo-RF, all samples are treated independent of each other.
- A pair of samples closest in distance are chosen as positive, and farthest as negative.
- SSiam that involves sorting a batch of queries is much more efficient over pseudo-RF

Method	BBT-0101	BF-0502
Pseudo-RF	0.930	0.814
SSiam	**0.962**	**0.909**

Pseudo-RF: Yan et al. Negative pseudo-relevance feedback in content-based video retrieval. In *ACM MM*, 2003

Performance on training videos.

Train/Test	Base	TSiam	SSiam
BBT-0101	0.932	**0.964**	0.962
BF-0502	0.836	0.893	**0.909**

	#cluster=36		
Methods	P	R	F
JFAC (ECCV '16)	0.690	0.350	0.460
Ours (with HAC)			
TSiam	**0.749**	**0.382**	**0.506**
SSiam	**0.766**	**0.386**	**0.514**

- Training is done at frame-level information.
- Testing is done at track-level i.e. mean representation.

Comparison with the SOTA at Frame-Level

Method	BBT-0101	BF-0502
ULDML (ICCV '11)	57.00	41.62
HMRF (CVPR '13)	59.61	50.30
HMRF2 (ICCV '13)	66.77	–
WBSLRR (ECCV '14)	72.00	62.76
VDF (CVPR '17)	89.62	87.46
Imp-Triplet (PacRim '16)	96.00	–
JFAC (ECCV '16)	–	92.13
Ours (with HAC)		
TSiam	**98.58**	**92.46**
SSiam	**99.04**	90.87

- Training the SSiam for about 15 epochs on BBT-0101 requires less than 25 minutes.

[**HMRF**] Wu et al.: Constrained Clustering and its Application to Face Clustering in Videos. In: CVPR. (2013)
[**HMRF2**] Wu et al.: Simultaneous Clustering and Tracklet Linking for Multi-face Tracking in Videos. In: ICCV. (2013)
[**WBSLRR**] Xiao et al.: Weighted Block-sparse Low Rank Representation for Face Clustering in Videos. In: ECCV. (2014)
[**McAFC**] Zhou et al..: Multi-cue augmented face clustering. In: ACM'MM. (2015)
[**CMVFC**] Cao et al.: Constrained Multi-view Video Face Clustering. IEEE TIP (2015)
[**VDF**] Sharma et al.. A simple and effective technique for face clustering in tv series. In *CVPR: Workshops* (2017)

Comparison with SOTA on ACCIO

	# clusters=40		
Methods	P	R	F
K-means-DeepID2$^+$ (ECCV '16)	0.543	0.201	0.293
DIFFRAC-DeepID2$^+$ (ICCV '11)	0.557	0.213	0.301
WBSLRR-DeepID2$^+$ (ECCV '14)	0.502	0.206	0.292
HMRF-DeepID2$^+$ (CVPR '13)	0.599	0.23.0	0.332
DeepID2$^+$·C0·Intra (ECCV '16)	0.657	0.312	0.423
JFAC (ECCV '16)	0.711	0.352	0.471
Ours (with HAC)			
TSiam	**0.763**	**0.362**	**0.491**
SSiam	**0.777**	**0.371**	**0.502**

Conclusion

- Presented two variants of discriminative methods to learn strong face representations
  - Self-supervised Siamese network (SSiam)
  - Track-supervised Siamese network (TSiam)
- State-of-the-art representation learning approach on BBT, BF and ACCIO.

Thank you!

https://vivoutlaw.github.io/

sharma.vivek@live.in