Video-based Face Recognition

Rama Chellappa

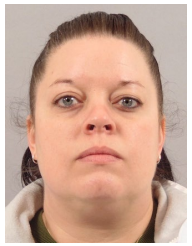
Johns Hopkins University

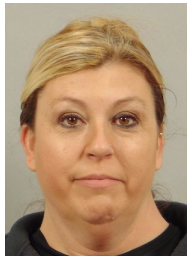
Baltimore, MD

Joint work with A. Bansal, N. Bodla, C. D. Castillo, C.H. Chen, J.C. Chen and Y.C. Chen, B. Lu, V. M. Patel, R. Ranjan, S. Sankaranarayanan and J. Zheng

The research in this presentation is based upon work supported by the Office of Naval Research under a MURI, the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), BEST program via a cooperative agreement Cooperative Agreement W911NF1020010 and the IARPA JANUS program via IARPA R&D Contracts No. 2019-022600002 and 2014-14071600012.

Introduction to Video-based Recognition





Gallery

Who Are They

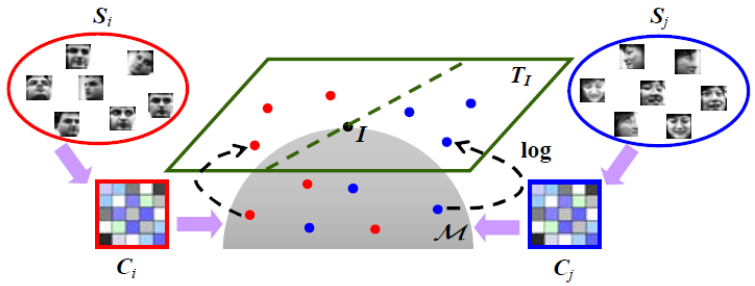
in the Video?

IJB-S Video Probes (Queries)

Pre-deep Learning Era: Representations

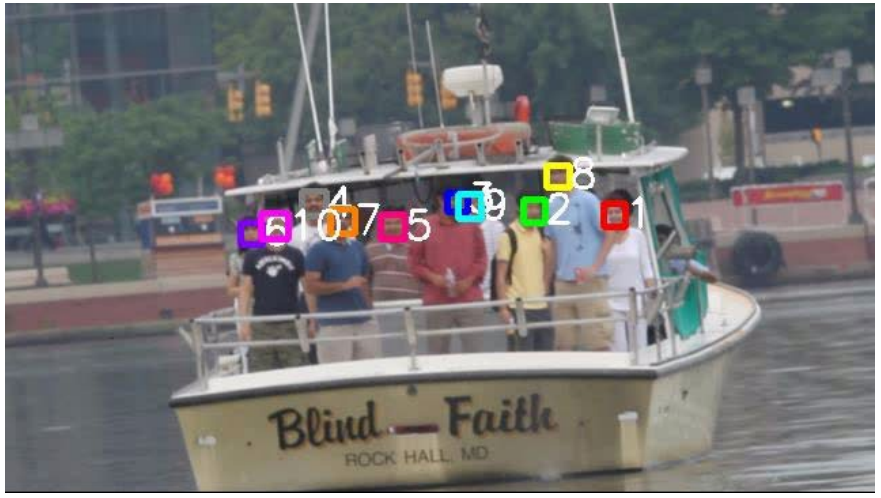
- Gabor jets and particle filtering: [Li and Chellappa, JOSA 2001]
- Online appearance models and particle filters: [Zhou & Chellappa, FG 2002, PAMI 2004]
- Temporal models: HMM [Liu & Chen CVPR 2003], ARMA [Aggarwal et al. ICPR 2004]
- 3D models: [Park & Jain ICB 2007]
- Manifolds: [Lee et al. CVIU 2005; Wang et al. CVPR 2008, CVPR 2009]
- Linear subspaces: Discriminative canonical correlation [Kim et al. PAMI 2007]
- Affine subspaces: Affine hull, convex hull [Cevikalp & Triggs CVPR 2010]
- Covariance matrices: [Wang et al. CVPR 2012]
- Dictionaries: [Chen et al. ECCV 2012]

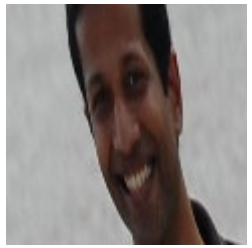
[Chen et al. ECCV 2012]

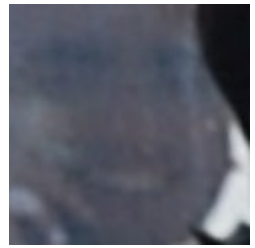


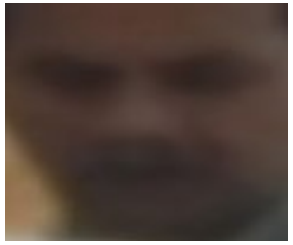
[Wang et al. CVPR 2012]

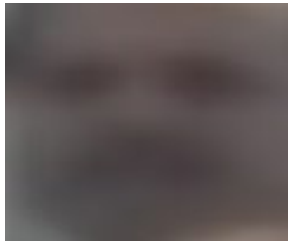
Video-based Face Recognition at > 50 meters 2008-2013









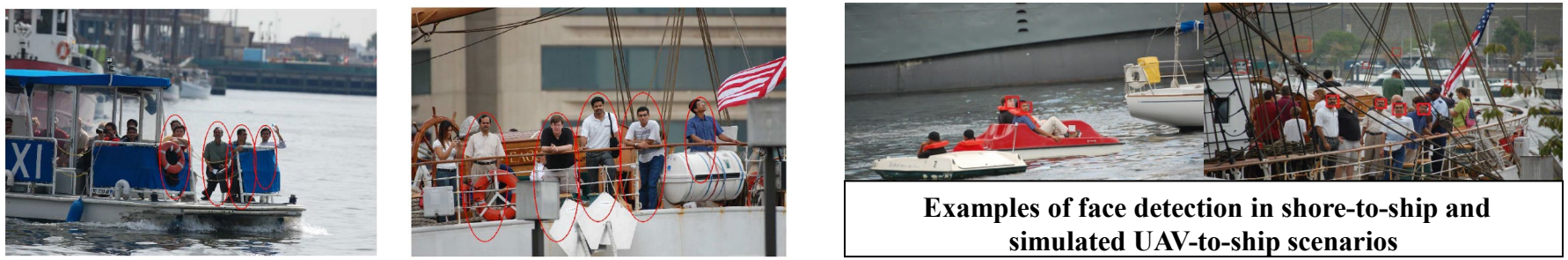


Challenges

- Atmospheric effects (fog, mist, rain, etc.)
- Blur
- Jitter due to ship motion
- Low-resolution
- Illumination, pose variations
- Occlusion
- Presence of others
- Collecting large data sets

Preprocessing – Face detection

- Before MURI
  - Viola Jones face detector
  - Video stabilization and face tracking using particle filters
- MURI
  - PLS method
  - Transitioned to DARPA VMR program
- Video stabilization and face tracking (UMD)
  - Association of frame-based detections using conditional random field models

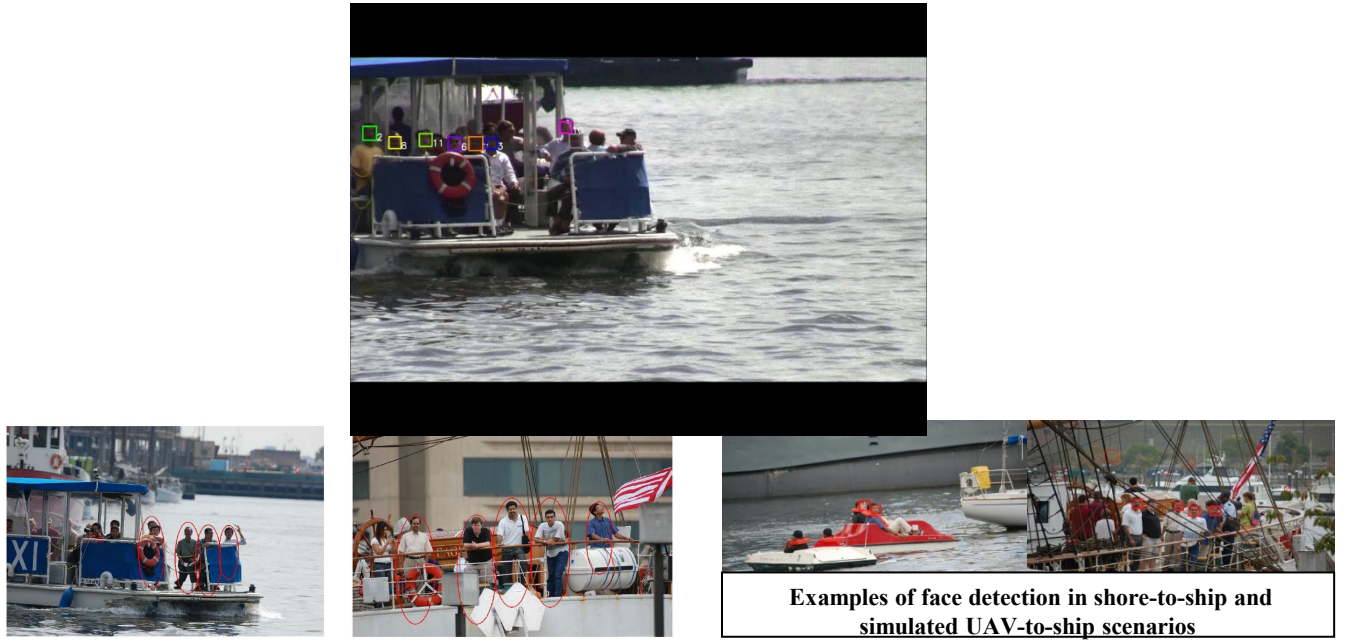


Examples of face detection in shore-to-ship and simulated UAV-to-ship scenarios

Acquisition of faces in motion

- We developed a Bayesian, scene-adaptive approach that is effective for scenarios involving sensor motion (ship to ship, ship to shore etc).
- Prior models tuned to scenes
- Online estimation of conditional random field models

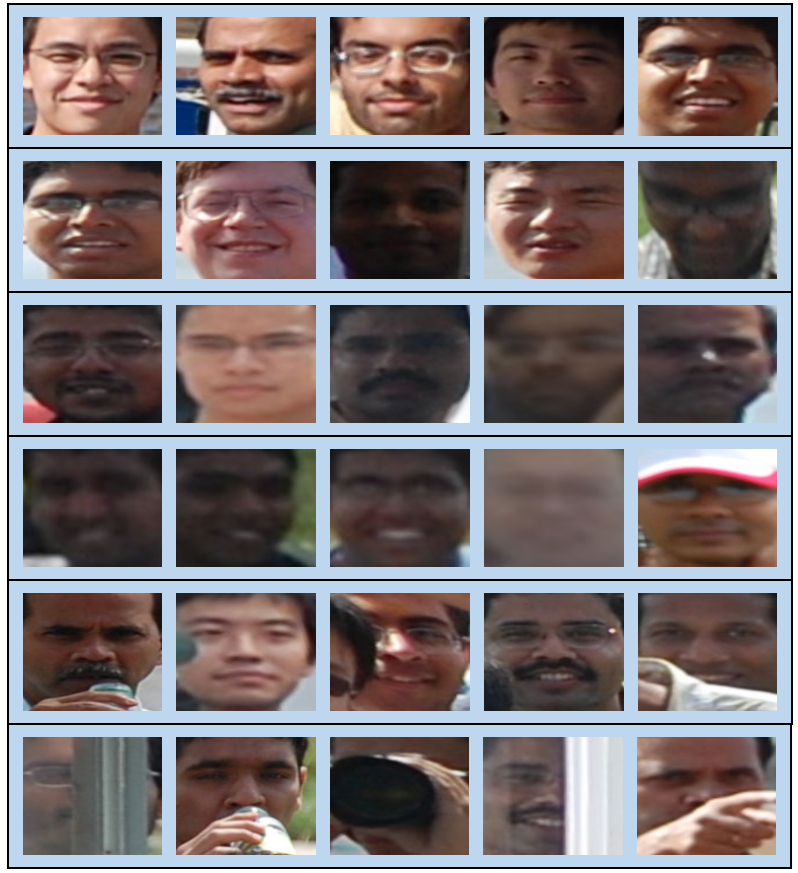
Examples of remote acquisition of faces

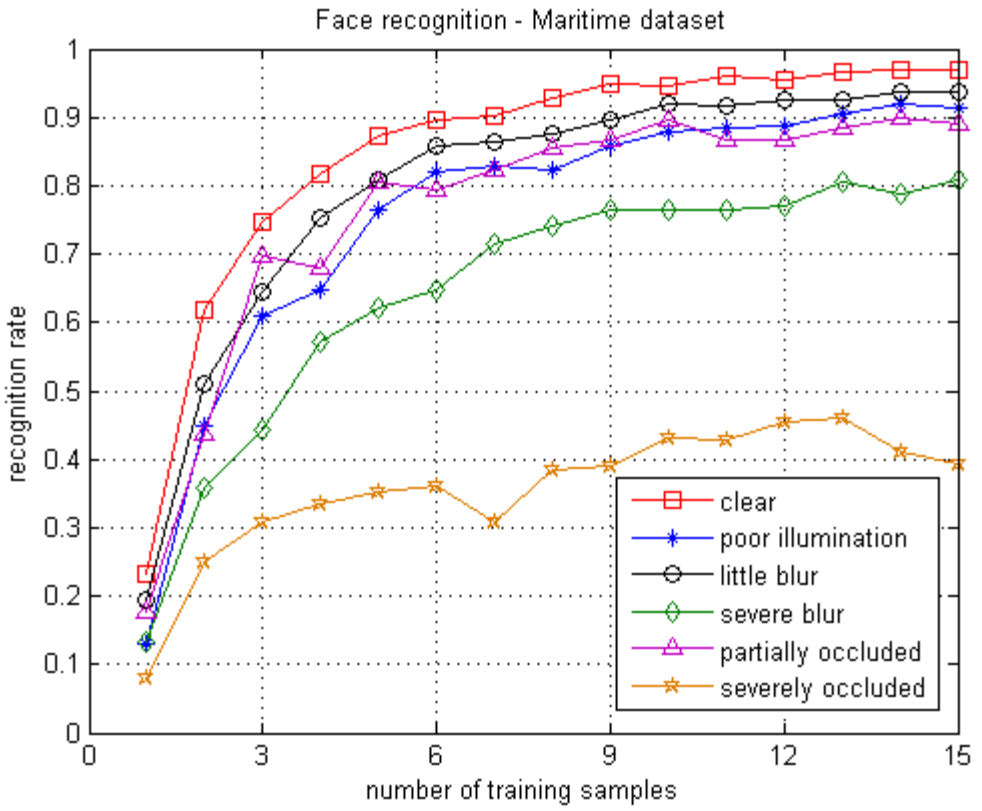


Examples of face detection in shore-to-ship and simulated UAV-to-ship scenarios

Face recognition on maritime data - partial least squares method

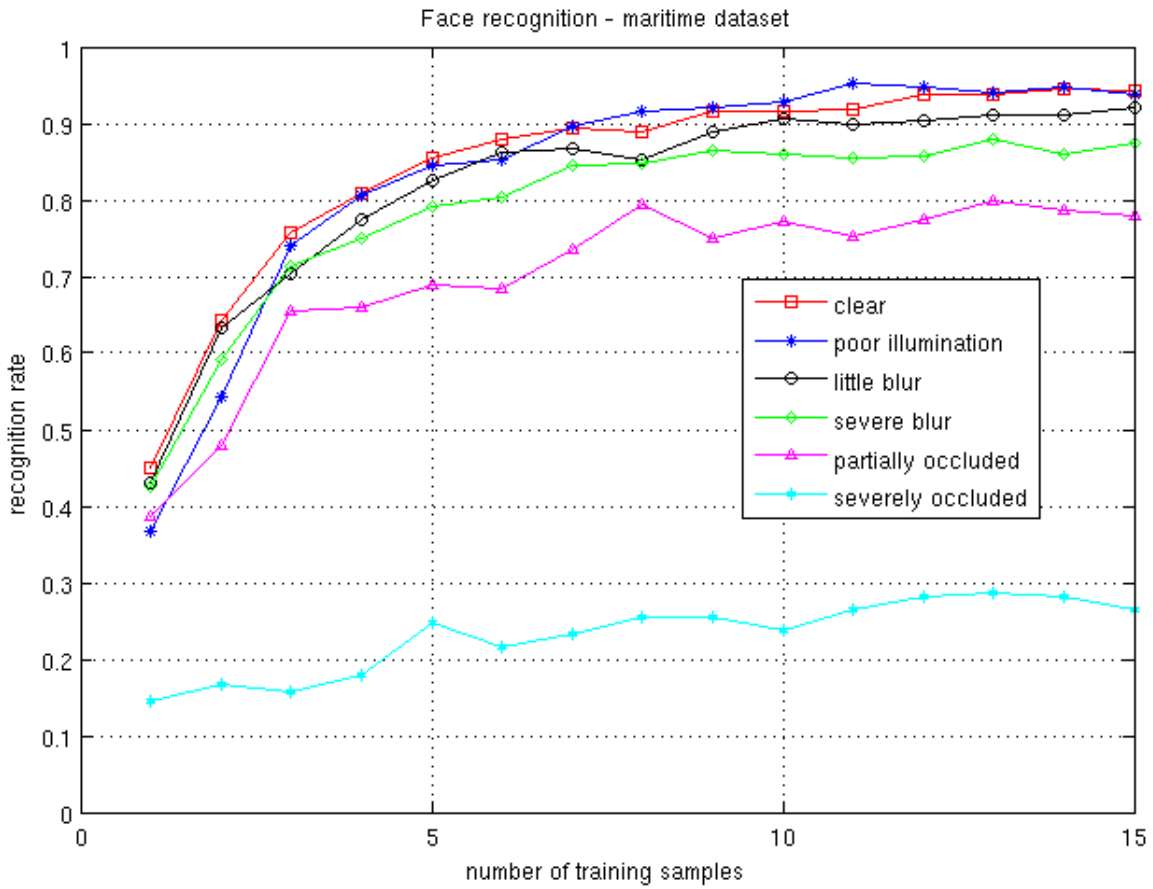
Feature dimensionality reduction 50,000 to 20.



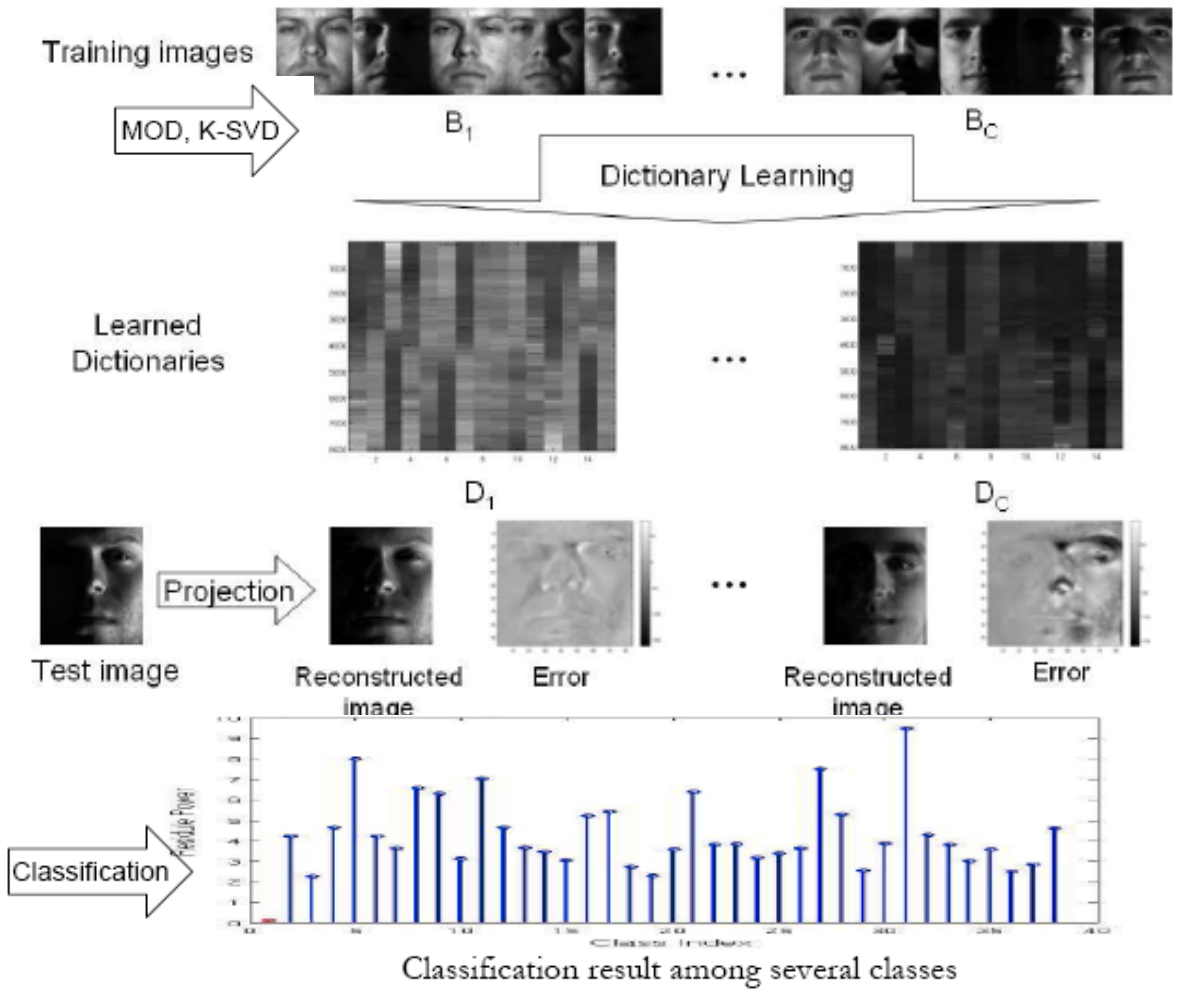


Face recognition on maritime intensity data - SVM

Kernel PCA
Regularized
LDA
SVM



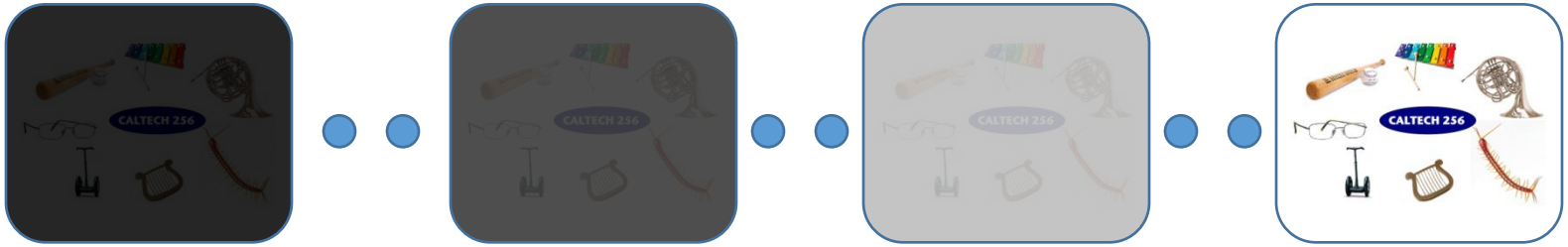
Dictionary-based face recognition



Classification result among several classes

How to learn Dictionaries?
K-SVD
M. Aharon, et al. 2006

Re-Identification of faces



Domain 1 (labeled Baltimore)

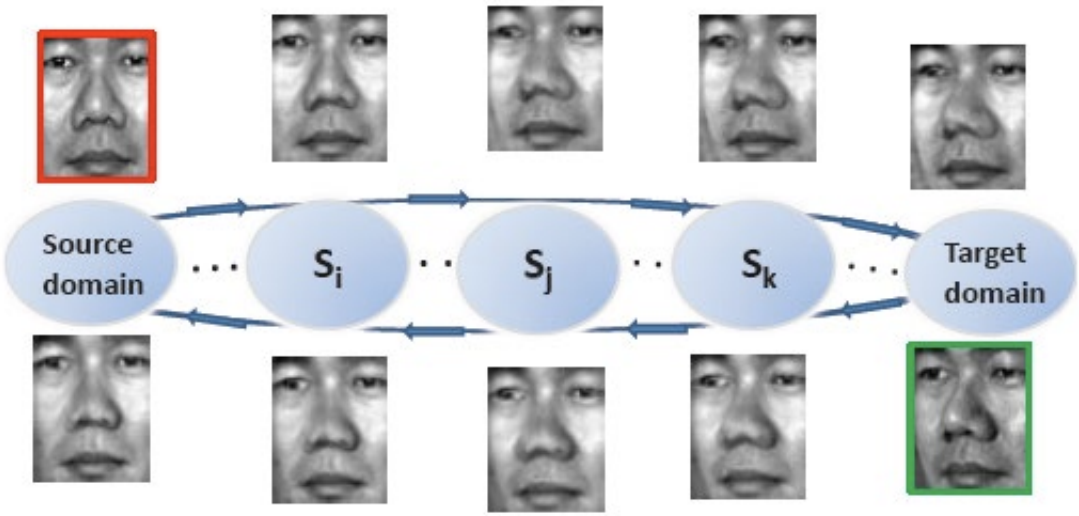
Intermediate domains (Incremental learning)

Domain 2 (unlabelled UMD Comcast center)

Questions

- How to obtain meaningful intermediate domains?
- How to characterize incremental domain shift information to perform recognition?
- Variations due to pose, illumination, background,..

Domain adaptation via dictionaries

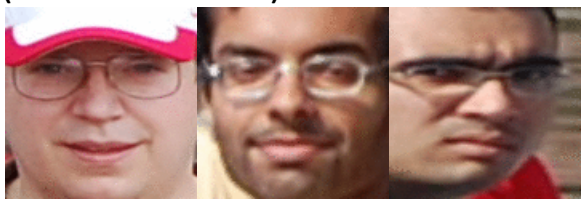


The top half of the figure shows some intermediate images synthesized from a given source image of frontal view (in red box). The bottom half shows the intermediate images generated from a given target image of side view (in green box).

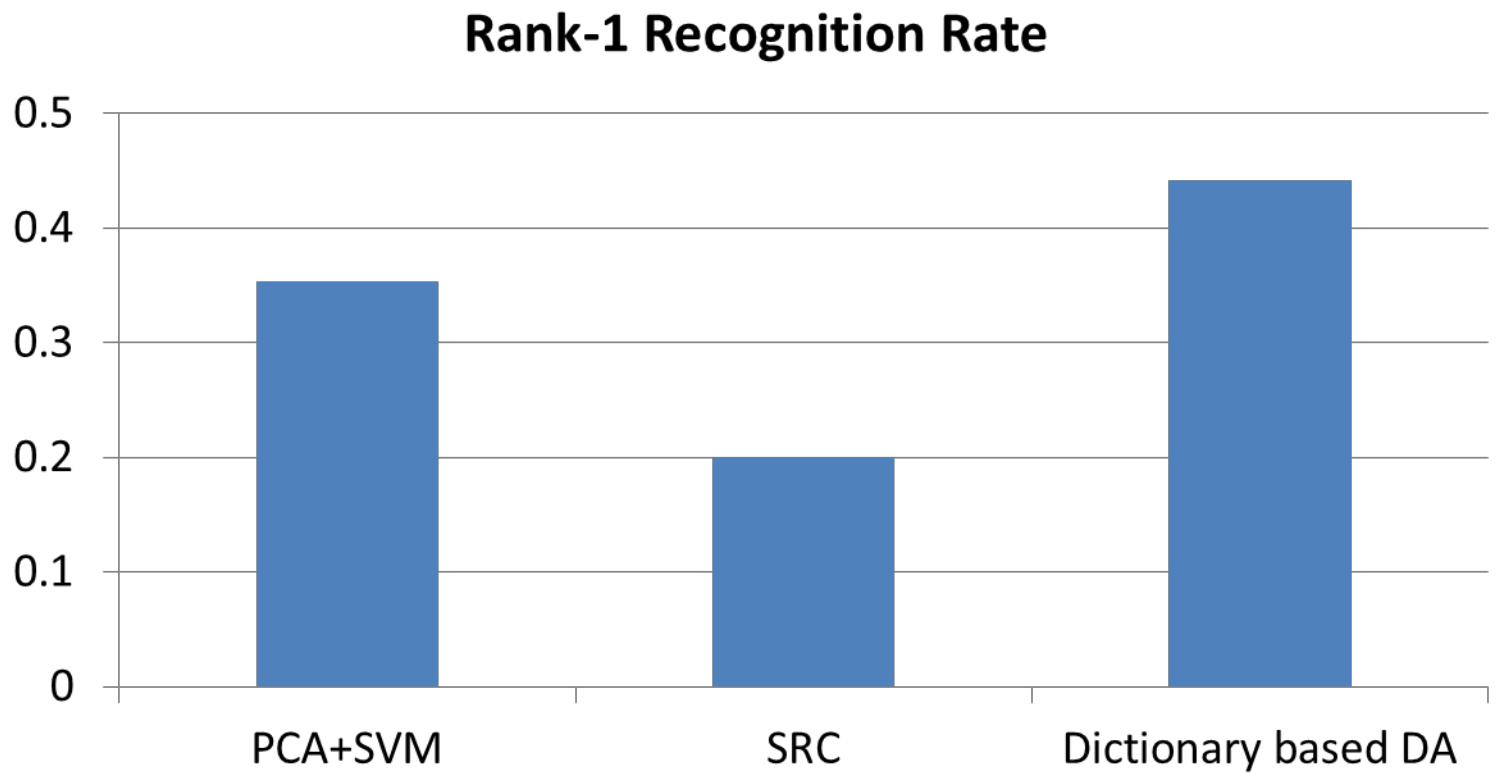
- Assume there exist K intermediate domains $\{S_k\}_{k=1}^{K}$ which smoothly bridge the information gap between the source and target domain. A domain dependent dictionary D_k is learned for each intermediate domain S_k.
- We learn the intermediate data to approximate the observations in the corresponding intermediate domains. The intermediate data is then utilized to build classifiers.

Results

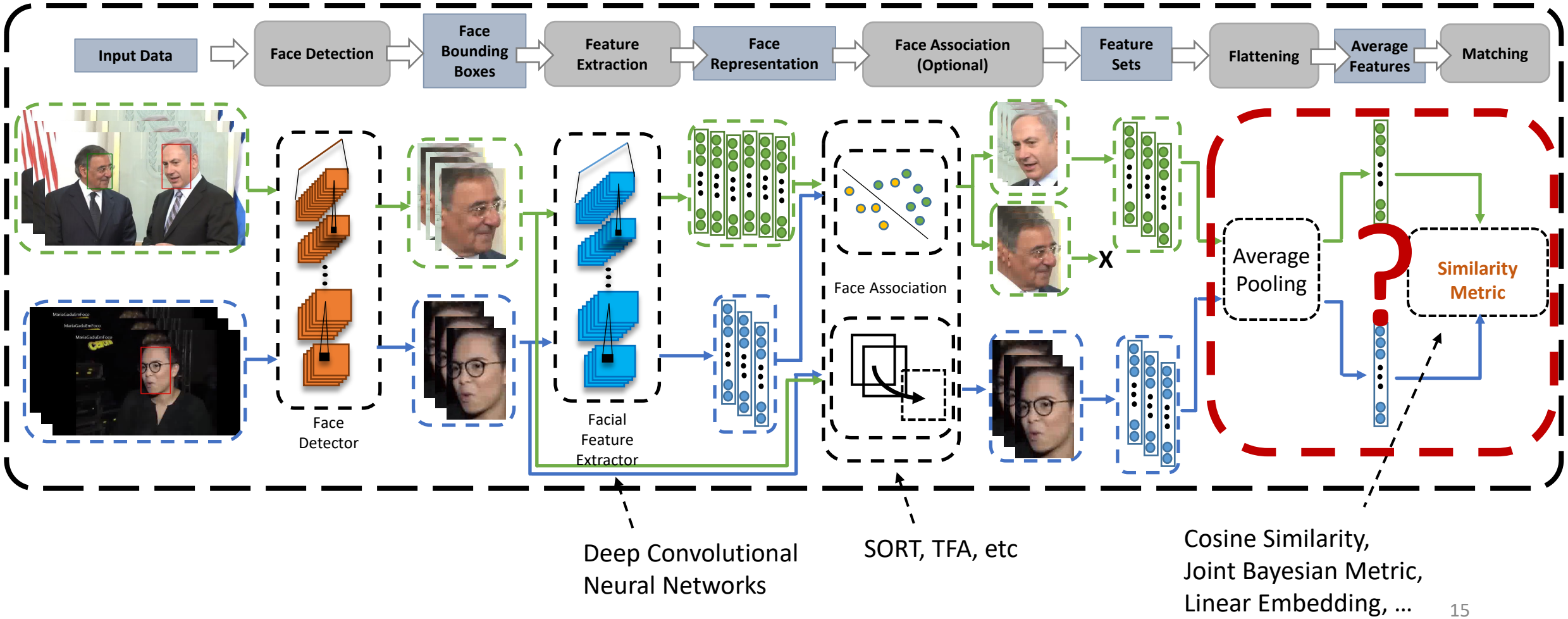
- 75 images from Baltimore dataset as gallery (source domain)



150 images from Comcast dataset as probe (target domain).



Deep Learning for Video-based Face Recognition

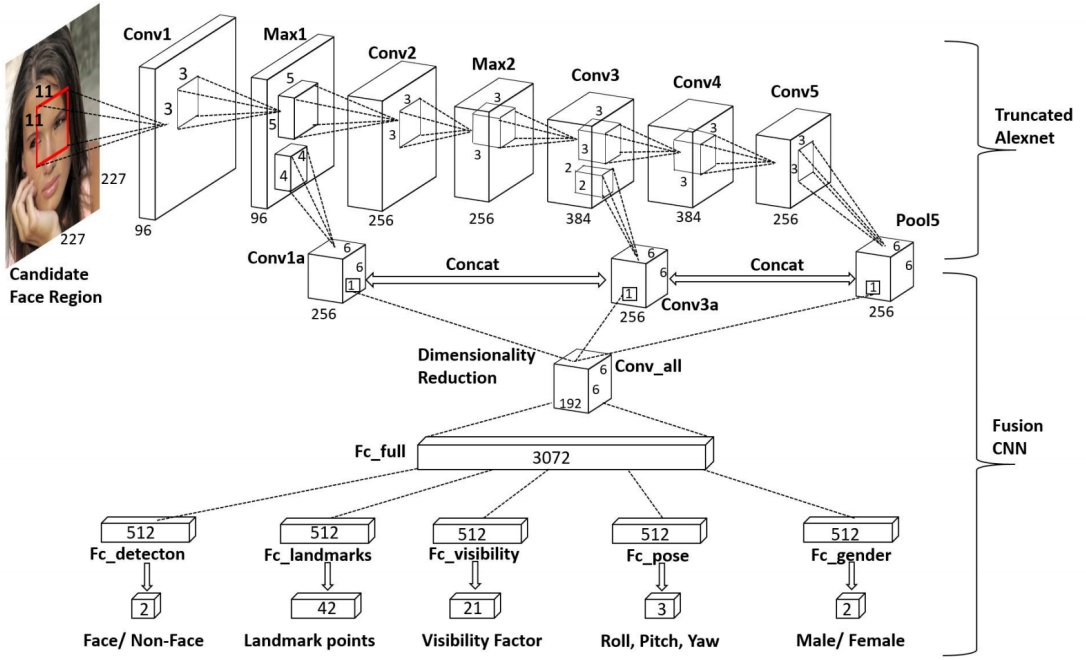


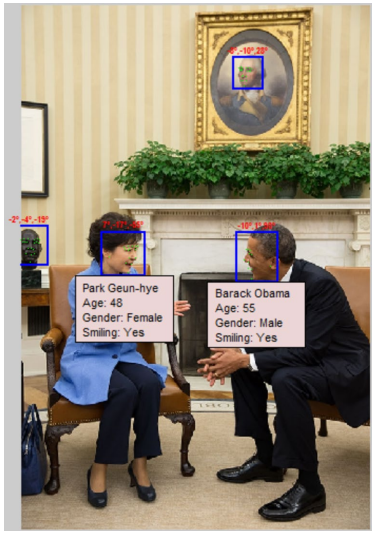
Methods

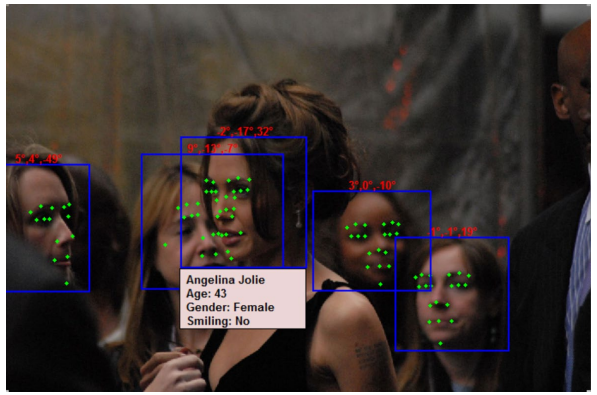
- Flattening frame-based deep features
- Image set-based method for unconstrained vide-based face recognition
- Incorporation of temporal information
- Face association using face and body

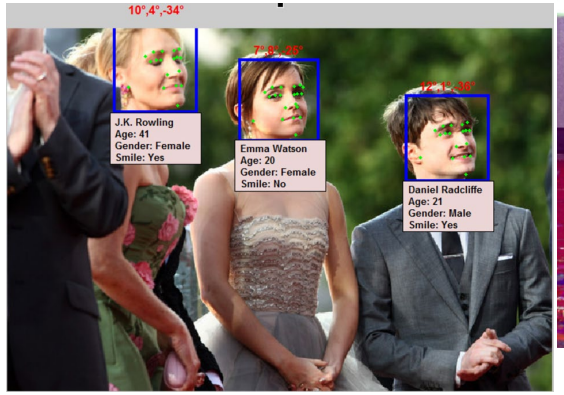
Method -1: Flattening Frame-based Deep Features

HyperFace-
PAMI 2018



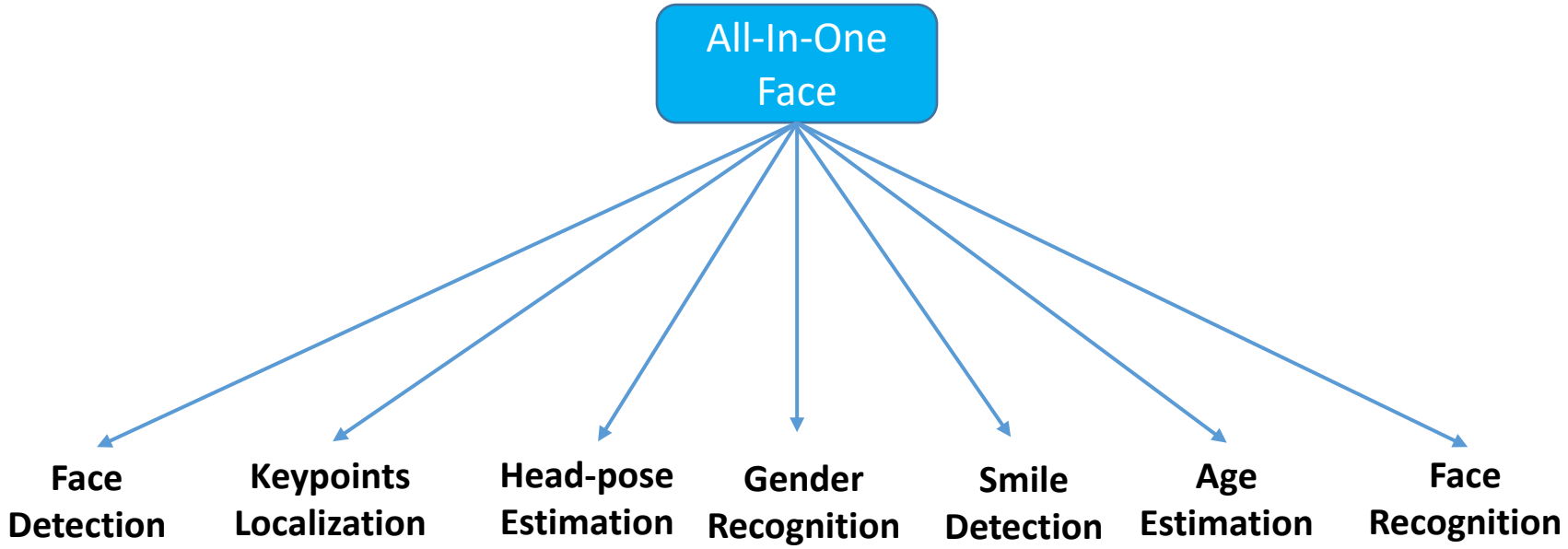




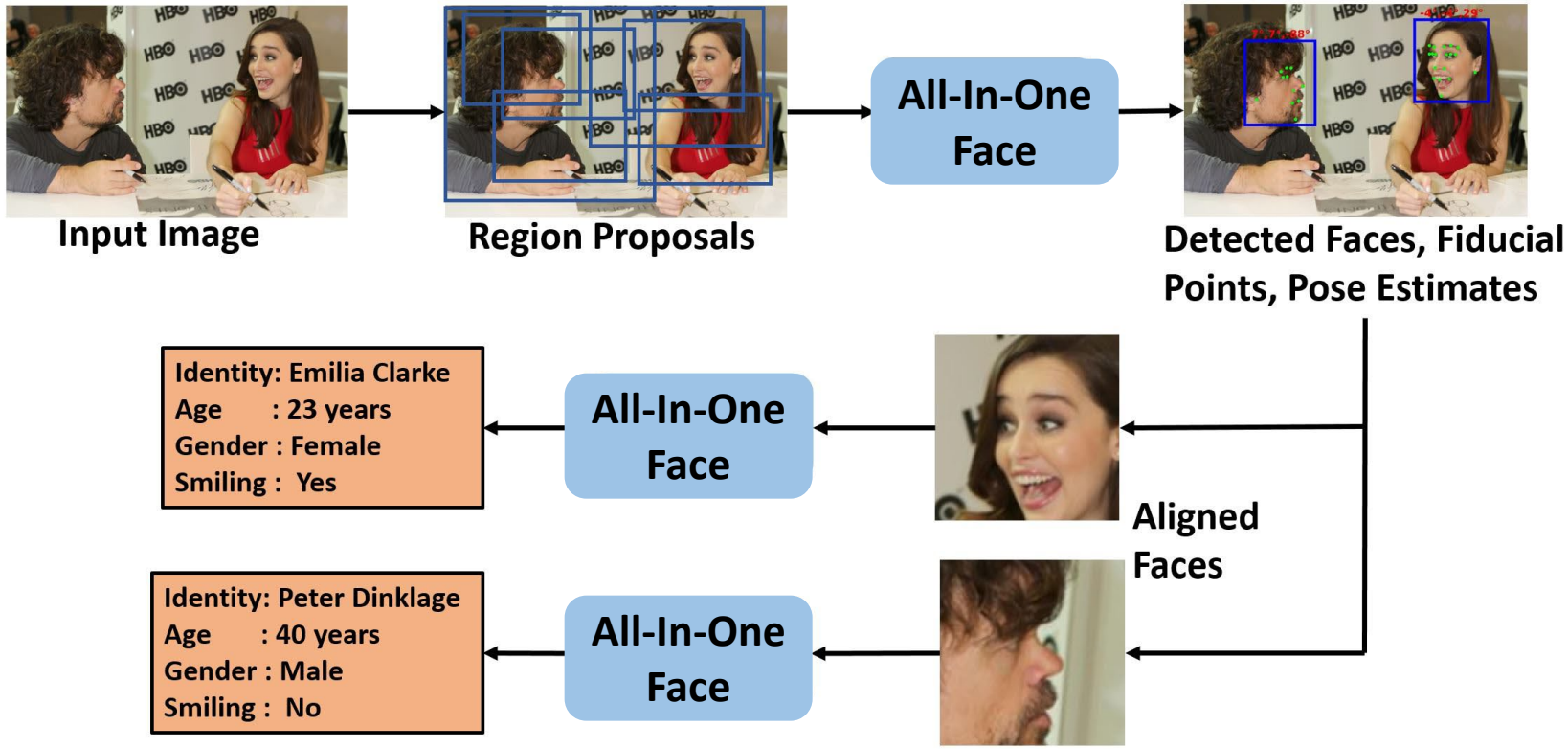


All-In-One Face (FG 2017)

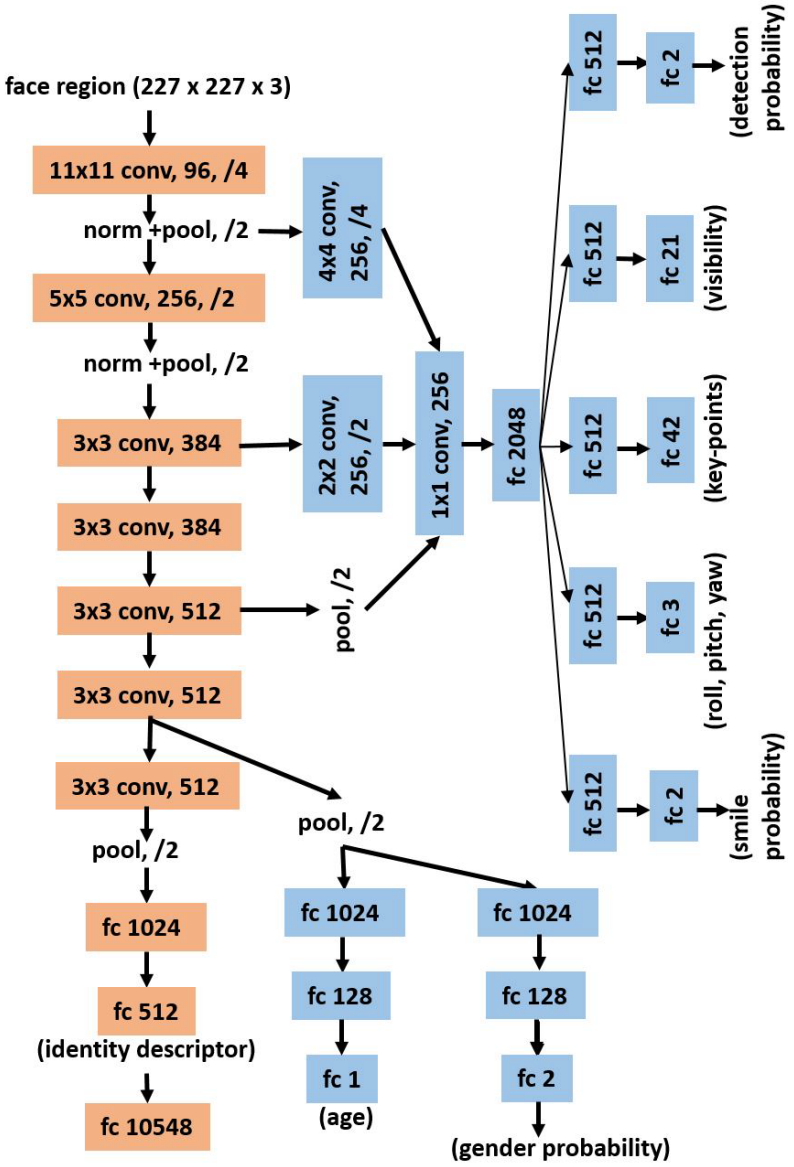
A single CNN model to perform multiple face analysis tasks.



All-In-One Face Pipeline



All-In-One Face Architecture



- Parameters initialized from face identification network[1]
- Subject-independent tasks share the lower layers of the network[2]
- Subject-specific features pooled from deeper layers of the network

1 (Sankarnarayanan et al., 2016).

2 (Ranjan et al., 2016).

Gang of Networks

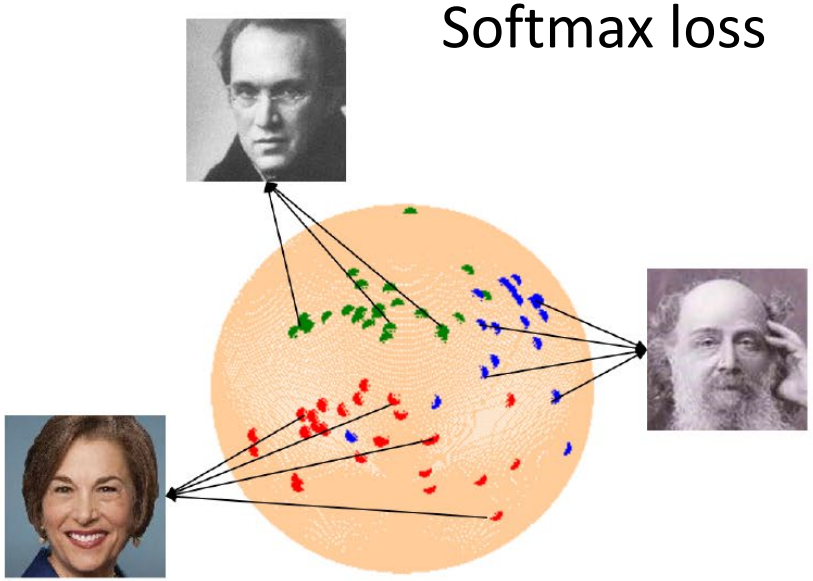
	Deep Networks			
	Rajeev-G1	Ankan-G1	Rajeev-G2	Ankan-G2
Training Set	MS1M-Curated + UMD Videos	MS1M-Curated + UMD Videos	MS1M-Curated - Megaface	MS1M-Curated - Megaface
Based Architecture	ResNet 101	Inception ResNet	ResNet 101	Inception ResNet
Loss Function	L2-Softmax (alpha=50)	L2-Softmax (alpha=50)	L2-Softmax (alpha=50)	L2-Softmax (alpha=50)
Embedding	TPE (UMDFaces stills)	TPE (UMDFaces stills)	TPE (UMDFaces stills)	TPE (UMDFaces stills)
Strengths	Just great	Good at high FAR	Good a high FAR	Good at high FAR
Alignment + Box size	Ultraface, 224x224	UltraFace, 299x299	Ultraface, 224x224	UltraFace, 299x299

Feature Distribution

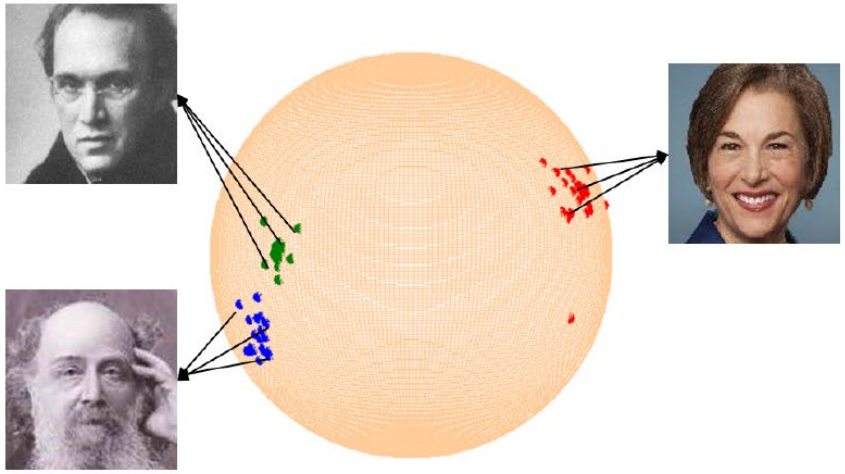
$$\text{minimize} \quad -\frac{1}{M}\sum_{i=1}^{M}\log\frac{e^{W_{y_i}^T f(\mathbf{x}_i)+b_{y_i}}}{\sum_{j=1}^{C} e^{W_j^T f(\mathbf{x}_i)+b_j}}$$

$$\text{subject to} \quad \|f(\mathbf{x}_i)\|_2 = \alpha, \quad \forall i = 1, 2, \ldots M,$$

Softmax loss



L2-Softmax loss



Hyperspheres are beautiful!

Other loss functions

- NormFace
- ArcFace
- CosFace
- RingLoss
- AM-Softmax
- SphereFace
- vMF-Loss
- Scalable Softmax
- GitLoss
- Decoupled Loss
- AAM-Loss
- SV-Softmax
- Marginal Loss
- Heated-Up-Softmax
- Copernican Loss
- KnotMagnify Loss

IARPA JANUS Benchmark B (IJB-B)

Example frames of a multiple-shot probe video in the IJBB dataset. The target annotation is in red box and face detection results from face detector are in green boxes

IJB-B Verification

The IJB-B dataset contains about 22,000 still images and 55, 000 video frames spread over 1, 845 subjects. Evaluation is done for 1:1 verification, and 1:N identification. The IJB-B verification protocol consists of 8,010,270 pairs between templates in the galleries (G1 and G2) and the probe templates. Out of these, 8 million are impostor pairs and the rest 10,270 are genuine comparisons.

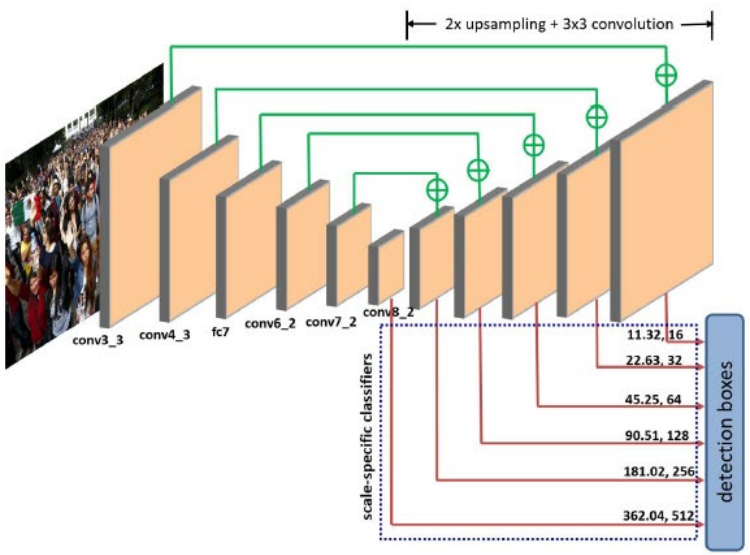
	True Accept Rate (%) @ False Accept Rate					
Method	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
GOTS	-	-	16.0	33.0	60.0	-
VGGFaces	-	-	55.0	72.0	86.0	-
FPN	-	-	83.2	91.6	96.5	-
Light CNN-29	-	-	87.7	92.0	95.3	-
VGGFace2	-	70.5	83.1	90.8	95.6	-
Center Loss	31.0	63.6	80.7	90.0	95.1	98.4
MN-vc	-	-	83.1	90.9	95.8	98.5
SENet50+DCN	-	-	84.9	93.7	**97.5**	**99.7**
ArcFace	37.5	**89.0**	**94.2**	**96.0**	97.5	98.4
Ours$_A$	27.7	61.6	89.1	94.3	97.0	98.7
Ours$_R$	**48.4**	80.4	89.8	94.4	97.2	98.9
Fusion (Ours)	45.6	77.8	90.3	94.6	97.3	98.9

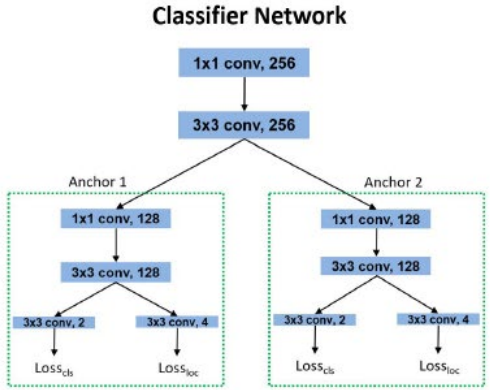
C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. C. Adams, T. Miller, N. D. Kalka, A. K. Jain, J. A. Duncan, K. Allen et al., "IARPA Janus Benchmark-B face dataset," in CVPR Workshops, 2017, pp. 592–600.

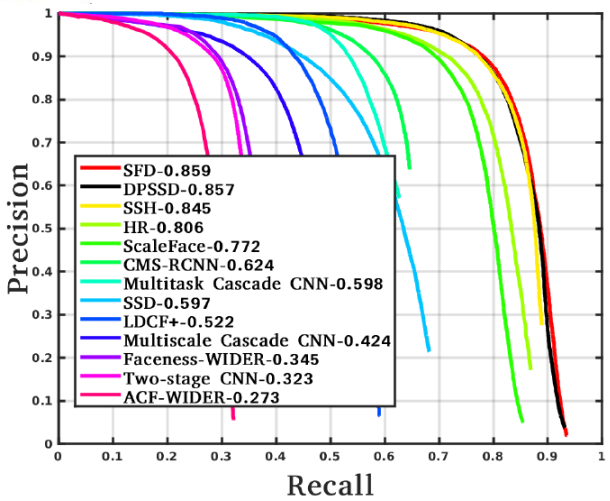
UMD-Janus: Results (IJB-B 1:N Identification)

	TPIR % @ FPIR (G1,G2)		Retrieval Rate (%) (G1,G2)		
	0.01	0.1	Rank=1	Rank=5	Rank=10
GOTS	-	-	42.0		62.0
VGGFace	-	-	78.0		89.0
FPN	-	-	91.1		96.5
Light CNN-29	-	-	91.9	94.8	-
VGGFace2	74.3	86.3	90.2	94.6	95.9
Center Loss	75.5, 67.7	87.5, 82.8	92.2, 86.0	95.4, 92.5	96.2, 94.4
UMD_A	83.1, 75.5	93.6, 89.3	95.5, 90.8	97.5, 94.2	98.0, 95.8
UMD_R	86.9, 78.6	94.0, 89.1	95.6, 91.5	**97.7, 95.4**	98.0, **96.5**
$UMD_{(Fused)}$	**88.2, 79.4**	**94.3, 89.7**	**95.8, 91.8**	97.7, 95.2	**98.1**, 96.4

Deep Pyramid Single Shot Face Detector (DPSSD)







WiderFace Hard

- Ranjan, Rajeev, Ankan Bansal, Jingxiao Zheng, Hongyu Xu, Joshua Gleason, Boyu Lu, Anirudh Nanduri, Jun-Cheng Chen, Carlos D. Castillo, and Rama Chellappa. "A Fast and Accurate System for Face Detection, Identification, and Verification." *arXiv preprint arXiv:1809.07586* (2018).

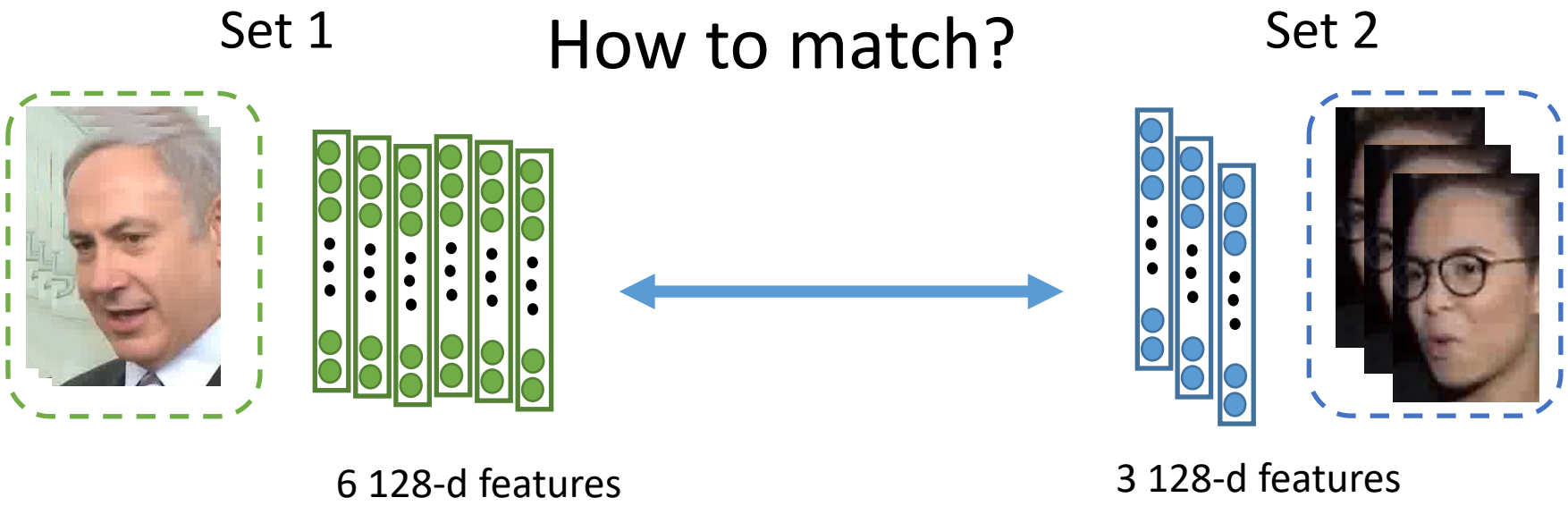
Method -2: Leverage the Set Information

- Faces in videos are sometimes associated into sets (non-sequential).
- Correlation exists between faces in the same set.

J. Zheng, R. Ranjan, C.-H. Chen, J.-C. Chen, C. D. Castillo and R. Chellappa. "An Automatic System for Unconstrained Video-Based Face Recognition.T-BIOM, 2020.

Motivation

Leverage the Set Information



- Aggregate face features into a unified fixed-size representation for efficient face recognition.

Motivation

Leverage the Set Information

- To exploit the correlation information in the sets and generate unified representations:
  - RNN-based methods:

    Need large scale labeled training data. Very expensive. Only applicable to sequential data.
  - Subspace-based methods:

    Encodes the correlation between samples. No external training data needed.

Subspace representation with subspace-based similarity

Subspace-based Representations

- Given deep features $\mathbf{Y}$, we learn the subspace representation $\mathbf{P}$ by

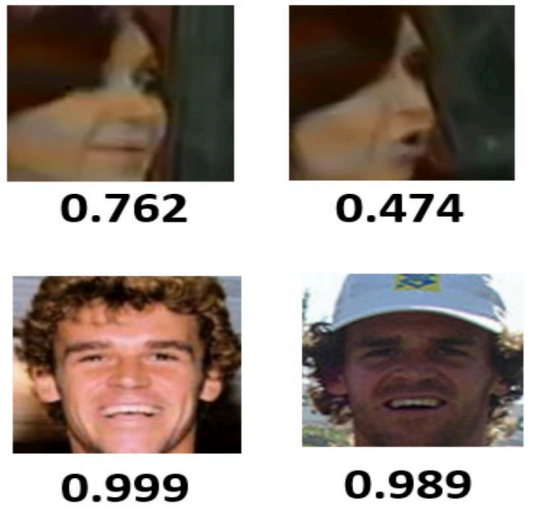
1. Subspace Learning (Sub):

$$\underset{\mathbf{P},\mathbf{X}}{\text{minimize}} \ \|\mathbf{Y} - \mathbf{P}\mathbf{X}\|_F^2 \quad s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I}$$

2. Quality-aware Subspace Learning (QSub):

$$\underset{\mathbf{P},\mathbf{X}}{\text{minimize}} \ \sum_{i=1}^{N} \tilde{d}_i \|\mathbf{y}_i - \mathbf{P}\mathbf{x}_i\|_2^2 \quad s.t.\ \mathbf{P}^T\mathbf{P} = \mathbf{I}$$

Normalized Detection Confidence as face quality indicator



Examples of faces with detection probability.

Global Representations

1. Average Pooling:

$$\mathbf{e} = \frac{1}{L}\sum_{i=1}^{L}\mathbf{y}_i$$

2. Quality-aware Average Pooling:

$$\mathbf{e}_D = \frac{1}{L}\sum_{i=1}^{L}\tilde{d}_i\mathbf{y}_i$$

Normalized Detection Confidence as face quality indicator

Similarity Metrics

- The similarity metrics between two sets of deep representations $\mathbf{Y}_1$ and $\mathbf{Y}_2$:
  1. Projection Metric (PM)

Principle angles between bases

$$s_{PM}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\frac{1}{r}\sum_{k=1}^{r}\cos^2\theta_k} = \sqrt{\frac{1}{r}\|\mathbf{P}_1^T\mathbf{P}_2\|_F^2}$$

  2. Variance-aware Projection Metric (VPM)

$$S_{VPM}(\mathbf{P}_1, \mathbf{P}_2) = \sqrt{\frac{1}{r}\sum_{k=1}^{r}\alpha^2(\lambda_{1k})\alpha^2(\lambda_{2k})\cos^2\theta_k} = \sqrt{\frac{1}{r}\|\tilde{\mathbf{P}}_1^T\tilde{\mathbf{P}}_2\|_F^2}$$

$$\text{where } \tilde{\mathbf{P}}_i = \mathbf{P}_i diag\{\alpha(\lambda_{ik})\}$$

Eigenvalues in PCA

Similarity Metrics

3. Cosine Similarity (Cos):

$$s_{cos}(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{\mathbf{e}_1^T \mathbf{e}_2}{\|\mathbf{e}_1\|_2 \|\mathbf{e}_2\|_2}$$

4. Quality-aware Cosine Similarity (QCos):

$$s_{Qcos}(\mathbf{Y}_1, \mathbf{Y}_2) = \frac{\mathbf{e}_{D1}^T \mathbf{e}_{D2}}{\|\mathbf{e}_{D1}\|_2 \|\mathbf{e}_{D2}\|_2}$$

5. Combining the quality-aware subspace learning, quality-aware average pooling and variance-aware projection metric, the overall similarity is

$$s(\mathbf{Y}_1, \mathbf{Y}_2) = s_{Qcos}(\mathbf{Y}_1, \mathbf{Y}_2) + \lambda s_{VPM}(\mathbf{P}_{D1}, \mathbf{P}_{D2})$$

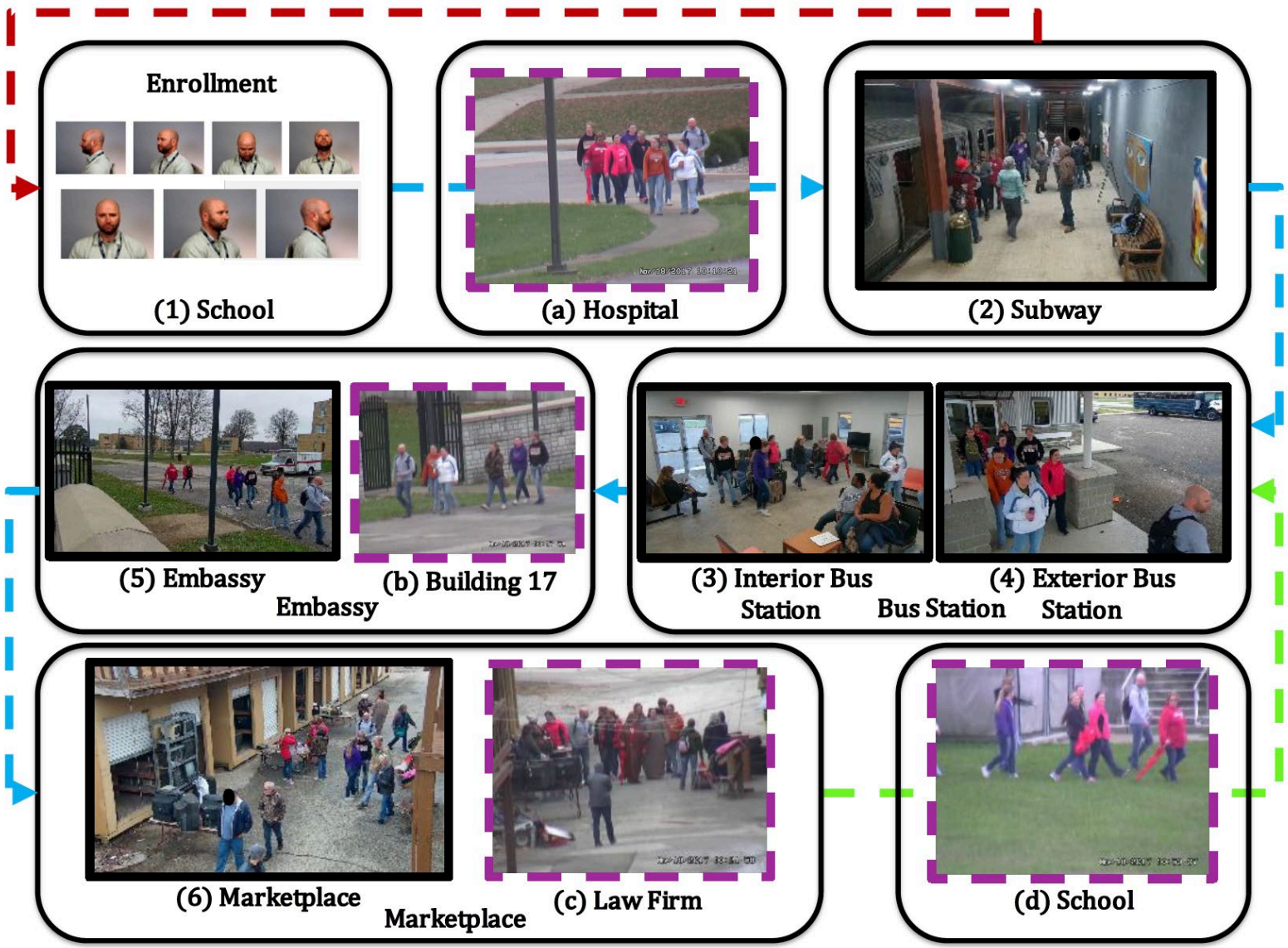
System Details

- Face Detection:
  - Multi-task SSD (Chen *et al.* 2018) for high quality faces,
  - DPSSD (Ranjan *et al.* 2019) for tiny faces.
- Facial Landmark Estimation:
  - All-in-One Face (Ranjan *et al.* 2017)
- Face Association:
  - SORT tracking for single-shot videos,
  - TFA (Chen *et al.* 2017) association for multi-shot videos.

A. Bewley, Z. Ge, L. Ott, F. Ramos, B. Upcroft. "Simple Online and Realtime Tracking". *ICIP*, 2016.

Face Association

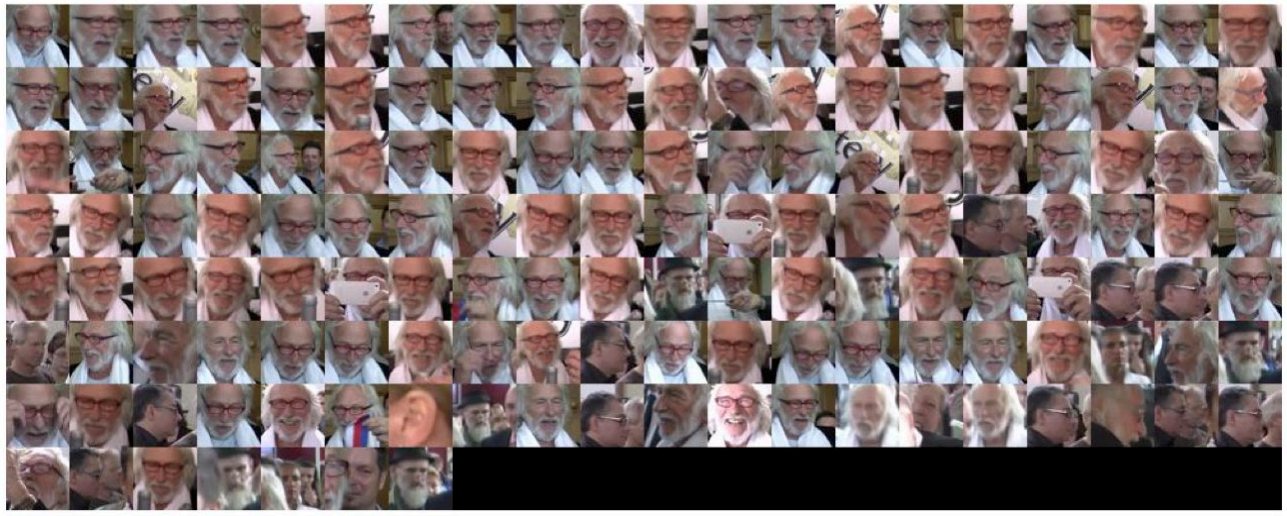
- Face Association for Videos Using Conditional Random Fields and Max-Margin Markov Networks, (Du and Chellappa, PAMI 2016)
- SORT
  - A real-time online tracking algorithm which approximates the dynamics with linear Gaussian state space models and associates detection bounding boxes in every frame using Kalman Filters.
- Target face association (Chen, at al., FG 2017)
  - Retrieves a set of representative face images in a given video that are likely to have the same identity as the target face.
- We have used a method that combines face and body features for face association using CRFs (ICCV 2019)



N. Kalka, B. Maze, J. Duncan, K. O'Connor, S. Elliott, K. Hebert, J. Bryan, A. Jain. "IJB–S: IARPA Janus Surveillance Video Benchmark." *BTAS* 2018.

Face Association

Face tracking using SORT on IJB-S



Face association using TFA on IJB-B

System Details

Deep Face Representation

- ResNet-101 and Inception-ResNet-v2, both trained on the union of MSCeleb-1M, UMDFaces, and UMDFaces Video datasets with the crystal loss.
- Features are further reduced to 128-dimensional by a Triplet Probabilistic Embedding (TPE).

R. Ranjan, A. Bansal, **J. Zheng**, H. Xu, J. Gleason, B. Lu, A. Nanduri, J.-C. Chen, C. Castillo, R. Chellappa. "A Fast and Accurate System for Face Detection, Identification, and Verification". *TBIOM,* 2019.

S. Sankaranarayanan, A. Alavi, C. Castillo, R. Chellappa. "Triplet Probabilistic Embedding for Face Verification and Clustering". *BTAS,* 2016.

Experiments and Compared Baselines

- We evaluated the proposed system on four datasets:
  1. Multiple Biometric Grand Challenge (MBGC),
  2. Face and Ocular Challenge Series (FOCS),
  3. IARPA Janus Benchmark B (IJB-B),
  4. IARPA JANUS Surveillance Video Benchmark (IJB-S).
- Metrics:
  1. Cosine Similarity (Cos)
  2. Quality-aware Cosine Similarity (QCos)
  3. Cos + Subspace with Projection Metric (Sub-PM)
  4. QCos + Sub-PM
  5. QCos + Quality-aware Subspace with Projection Metric (QSub-PM)
  6. QCos + Quality-aware Subspace with Variance-aware Projection Metric (QSub-VPM)

Compared Baselines

- Methods:
  1. Dictionary-Based Face Recognition from Video (DFRV)
- Representations:
  1. ArcFace (Arc-)

J. Deng, J. Guo, X. Niannan, and S. Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition." In CVPR, 2019.

Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. "Dictionary-based face recognition from video." ECCV, October 2012.

Y. C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. "Dictionary-based face and person recognition from unconstrained video." IEEE Access, 2015.

IARPA JANUS Benchmark B (IJB-B)

- A template-based unconstrained face recognition dataset.
- Contains 1845 subjects with 11,754 images, 55,025 frames and 7,011 multiple-shot videos.
- We focus on the video identification protocol.
- It is an open set 1:N identification protocol where each given probe is collected from a video and is searched among all gallery faces. Gallery candidates are ranked according to their similarity scores to the probes. Top-K rank accuracy and True Positive Identification Rate (TPIR) over False Positive Identification Rate(FPIR) are used to evaluate the performance.

C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, and K. Allen. IARPA Janus Benchmark-B face dataset. CVPRW, 2017.

Identification Results on IJB-B

Methods	Rank=1	Rank=2	Rank=5	Rank=10	Rank=20	Rank=50
Chen *et al.* with Iteration 0	55.94%	-	68.40%	72.89%	-	83.71%
Chen *et al.* with Iteration 3	61.01%	-	73.39%	77.90%	-	87.62%
Chen *et al.* with Iteration 5	61.00%	-	73.46%	77.94%	-	87.69%
Cos	78.37%	81.35%	84.39%	86.29%	88.30%	90.82%
QCos	78.43%	81.41%	84.40%	86.33%	88.34%	90.88%
Cos+Sub-PM	77.99%	81.45%	84.68%	86.75%	88.96%	91.91%
QCos+Sub-PM	78.02%	81.46%	84.76%	86.72%	88.97%	91.91%
QCos+QSub-PM	78.04%	81.47%	84.73%	86.72%	88.97%	91.93%
QCos+QSub-VPM	**78.93%**	**81.99%**	**84.96%**	**87.03%**	**89.24%**	**92.02%**

IARPA JANUS Surveillance Video Benchmark (IJB-S)

- An unconstrained video-based face recognition dataset.
- Galleries: high-resolution still images. Probes: low quality, remotely captured surveillance videos.
- 202 subjects from 1421 images and 398 single-shot surveillance videos.
- We focus on surveillance-to-single , surveillance-to-booking and surveillance-to-surveillance identification protocols.

N. Kalka, B. Maze, J. Duncan, K. O'Connor, S. Elliott, K. Hebert, J. Bryan, A. Jain. "IJB–S: IARPA Janus Surveillance Video Benchmark." *BTAS* 2018.

Identification Results on IJB-S

Surveillance-to-Single

Methods	Top-K Average Accuracy **with Filtering**						EERR metric **without Filtering**					
	R=1	R=2	R=5	R=10	R=20	R=50	R=1	R=2	R=5	R=10	R=20	R=50
Arc-Cos (Deng *et al.*)	52.03%	56.83%	63.16%	69.05%	76.13%	88.95%	24.45%	26.54%	29.35%	32.33%	36.38%	44.81%
Arc-QCos+QSub-PM	60.92%	65.06%	70.45%	75.19%	80.69%	90.29%	28.73%	30.44%	32.98%	35.40%	38.70%	45.46%
Cos	64.86%	70.87%	77.09%	81.53%	86.11%	93.24%	29.62%	32.34%	35.60%	38.36%	41.53%	46.78%
QCos	65.42%	71.34%	77.37%	81.78%	86.25%	93.29%	29.94%	32.60%	35.85%	38.52%	41.70%	46.78%
Cos+Sub-PM	69.52%	75.15%	80.41%	84.14%	87.83%	94.27%	32.22%	34.70%	37.66%	39.91%	42.65%	47.54%
QCos+Sub-PM	69.65%	75.26%	80.43%	84.22%	87.81%	94.25%	32.27%	34.73%	37.66%	39.91%	42.67%	47.54%
QCos+QSub-PM	**69.82%**	**75.38%**	**80.54%**	**84.36%**	**87.91%**	**94.34%**	**32.43%**	**34.89%**	**37.74%**	**40.01%**	**42.77%**	**47.60%**
QCos+QSub-VPM	69.43%	75.24%	80.34%	84.14%	87.86%	94.28%	32.19%	34.75%	37.68%	39.88%	42.56%	47.50%

Identification Results on IJB-S

Surveillance-to-Booking

Methods	Top-K Average Accuracy **with Filtering**						EERR metric **without Filtering**					
	R=1	R=2	R=5	R=10	R=20	R=50	R=1	R=2	R=5	R=10	R=20	R=50
Arc-Cos (Deng *et al.*)	54.59%	59.12%	65.43%	71.05%	77.84%	89.16%	25.38%	27.58%	30.59%	33.42%	37.60%	45.05%
Arc-QCos+QSub-VPM	60.86%	65.36%	71.30%	76.15%	81.63%	90.70%	28.66%	30.64%	33.43%	36.11%	39.57%	45.70%
Cos	66.48%	71.98%	77.80%	82.25%	86.56%	93.41%	30.38%	32.91%	36.15%	38.77%	41.86%	46.79%
QCos	66.94%	72.41%	78.04%	82.37%	86.63%	93.43%	30.66%	33.17%	36.28%	38.84%	41.88%	46.84%
Cos+Sub-PM	69.39%	74.55%	80.06%	83.91%	87.87%	**94.34%**	32.02%	34.42%	37.59%	39.97%	42.64%	**47.58%**
QCos+Sub-PM	69.57%	74.78%	80.06%	83.89%	87.94%	94.33%	32.16%	34.61%	37.62%	39.99%	42.71%	47.57%
QCos+QSub-PM	69.67%	74.85%	80.25%	84.10%	88.04%	94.22%	32.28%	34.77%	37.76%	40.11%	**42.76%**	47.57%
QCos+QSub-VPM	**69.86%**	**75.07%**	**80.36%**	**84.32%**	**88.07%**	94.33%	**32.44%**	**34.93%**	**37.80%**	**40.14%**	42.72%	**47.58%**

Identification Results on IJB-S

Surveillance-to-Surveillance

Methods	Top-K Average Accuracy **with Filtering**						EERR metric **without Filtering**					
	R=1	R=2	R=5	R=10	R=20	R=50	R=1	R=2	R=5	R=10	R=20	R=50
Arc-Cos (Deng *et al.*)	8.68%	12.58%	18.79%	26.66%	39.22%	68.19%	4.98%	7.17%	10.86%	15.42%	22.34%	37.68%
Arc-QCos+QSub-PM	8.64%	12.57%	18.84%	26.86%	39.78%	**68.21%**	**5.26%**	**7.44%**	**11.31%**	15.90%	**22.68%**	**37.83%**
Cos	8.54%	11.99%	19.60%	28.00%	37.71%	59.44%	4.42%	6.15%	10.84%	15.73%	21.14%	33.21%
QCos	8.62%	12.11%	19.62%	28.14%	37.78%	59.21%	4.46%	6.20%	10.80%	15.81%	21.06%	33.17%
Cos+Sub-PM	8.19%	11.79%	19.56%	28.62%	39.77%	63.15%	4.26%	6.25%	10.79%	16.18%	22.48%	34.82%
QCos+Sub-PM	8.24%	11.82%	19.68%	28.68%	39.68%	62.96%	4.27%	6.25%	10.92%	16.18%	22.39%	34.69%
QCos+QSub-PM	8.33%	11.88%	19.82%	28.65%	39.78%	62.79%	4.33%	6.21%	10.96%	16.19%	22.48%	34.69%
QCos+QSub-VPM	8.66%	12.27%	**19.91%**	**29.03%**	**40.20%**	63.20%	4.30%	6.30%	10.99%	**16.23%**	22.50%	34.76%

Method 3: Leverage the Temporal Information

- To exploit the temporal information in a video for video-based face recognition:
  - RNN-based methods:
    Need large scale labeled training data. Very expensive.
  - Dictionary learning:
    Robust to noise. No external training data needed.
  - Linear Dynamical Systems (LDSs):
    Represent sequential data.
- Given face feature sequence $\mathbf{Y}$, we learn two kinds of dictionaries:
  1. Structural Dictionaries $\mathbf{D}_s$:
     Represent non-sequential correlation.
  2. Dynamical Dictionaries $\mathbf{D}_d$:
     Represent sequential correlation.

Structural Dictionary Learning

- Learn dictionary $\mathbf{D}_s$ from face feature sequence $\mathbf{Y}$, ignoring the order.

$$\begin{aligned} \min_{\mathbf{D}_s, \mathbf{X}} \quad & \|\mathbf{Y} - \mathbf{D}_s \mathbf{X}\|_F^2 \\ \text{s.t} \quad & \|\mathbf{x}_i\|_0 \leq T \\ & \mathbf{d}_i^T \mathbf{d}_i = 1 \quad \forall i \end{aligned}$$

Dynamical Dictionary Learning

Basic Problem

- Learn dictionary $\mathbf{D}_d$ and motion $\mathbf{A}$ from face feature sequence $\mathbf{Y}$ jointly.

Dictionary Learning:

$$\min_{\mathbf{D}_d, \mathbf{X}} \|\mathbf{Y} - \mathbf{D}_d \mathbf{X}\|_F^2$$

$$\text{s.t} \quad \|\mathbf{x}_i\|_0 \leq T,$$

$$\mathbf{d}_i^T \mathbf{d}_i = 1 \quad \forall i$$

Combine

LDS Model:

$$\mathbf{y}_t = \mathbf{D}_d \mathbf{x}_t + \mathbf{w}_t$$

$$\mathbf{x}_{t+1} = \mathbf{A}\mathbf{x}_t + \mathbf{v}_t$$

Dynamical Dictionary Learning

Basic Problem

- Learn dictionary D_d and motion A from video features Y jointly.

Reconstruction Error

Motion Error

Motion smoothness

$$\min_{\mathbf{D}_d,\mathbf{X},\mathbf{A}} \|\mathbf{Y}-\mathbf{D}_d\mathbf{X}\|_F^2+\eta\|\mathbf{X}_1-\mathbf{A}\mathbf{X}_0\|_F^2+\gamma\|\mathbf{A}\|_F^2$$

$$\text{s.t.} \quad \|\mathbf{x}_i\|_0 \le T, \quad \mathbf{d}_i{}^T\mathbf{d}_i = 1 \quad \forall i$$

$$\text{where } \mathbf{X}_0 = \left[\mathbf{x}_1,\ldots,\mathbf{x}_{L-1}\right],\ \mathbf{X}_1 = \left[\mathbf{x}_2,\ldots,\mathbf{x}_L\right].$$

Dynamical Dictionary Learning

Extended Problem

- For a subject with N tracklets:

$$\min_{\mathbf{D}_d,\mathbf{X},\mathbf{A}} \sum_{n=1}^{N} \|\mathbf{Y}^n - \mathbf{D}_d\mathbf{X}^n\|_F^2 + \eta \sum_{n=1}^{N} \|\mathbf{X}_1^n - \mathbf{A}^n\mathbf{X}_0^n\|_F^2 + \gamma \sum_{n=1}^{N} \|\mathbf{A}^n\|_F^2$$
$$\text{s.t.} \quad \|\mathbf{x}_i^n\|_0 \leq T, \quad \mathbf{d}_i{}^T\mathbf{d}_i = 1 \quad \forall i, n$$

Difficult Term

Dynamical Dictionary Learning

Auxiliary Variable W

$$\min_{\mathbf{D}_d,\mathbf{X},\mathbf{A},\mathbf{W}} \sum_{n=1}^{N} \|\mathbf{Y}^n - \mathbf{D}_d\mathbf{X}^n\|_F^2$$
$$+\beta \sum_{n=1}^{N} \|\mathbf{X}^n - \mathbf{W}^n\|_F^2$$
$$+\eta \sum_{n=1}^{N} \|\mathbf{W}_1^n - \mathbf{A}^n\mathbf{W}_0^n\|_F^2$$
$$+\gamma \sum_{n=1}^{N} \|\mathbf{A}^n\|_F^2$$
$$\text{s.t.} \quad \|\mathbf{x}_i^n\|_0 \leq T \quad , \mathbf{d}_i^T\mathbf{d}_i = 1 \quad \forall i, n$$

Sparse coding term and motion term are separated

LDDL Algorithm

Algorithm 1: The proposed LDDL algorithm.

Data: Training data:
$\mathbf{Y} = [\mathbf{Y}^1, \cdots, \mathbf{Y}^N]$.
Initialize $\mathbf{D}$ using K-SVD;
for $n = 1 : N$ **do**
    Initialize $\mathbf{X}^n$ using OMP;
    Initialize $\mathbf{W}^n$ by zeros;
    Initialize $\mathbf{A}^n$ based on initialized $\mathbf{X}^n$;
end
while $Iter < Loop$ **do**
    for $n = 1 : N$ **do**
        Update $\mathbf{X}^n$;
        Update $\mathbf{W}^n$;
        Update $\mathbf{A}^n$;
    end
    Update $\mathbf{D}$;
    $Iter = Iter + 1$;
end
Result: Dictionary $\mathbf{D}$, transition matrices $\{\mathbf{A}^n\}$, sparse codes $\{\mathbf{X}^n\}$.

Similarity Metric between Videos

- Given video face representations $\mathbf{Y}_i$
  1. Dictionaries $\mathbf{D}_{si}, \mathbf{D}_{di}$ are learned using the proposed method.
  2. Orthonormal bases $\mathbf{P}_{si}, \mathbf{P}_{di}$ are computed using QR decomposition.
  3. Global representations are computed by average pooling

$$\mathbf{e}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{y}_{ij}$$

Similarity Metric between Videos

- The video-to-video similarity is computed as:

Principle angles between bases

$$s(\mathbf{Y}_1, \mathbf{Y}_2) = \cos\theta_0 + \lambda_s \sqrt{\frac{1}{r}\sum_{k=1}^{r} \cos^2\theta_{sk}} + \lambda_d \sqrt{\frac{1}{r}\sum_{k=1}^{r} \cos^2\theta_{dk}}$$

$$= \frac{\mathbf{e}_1^T \mathbf{e}_2}{\|\mathbf{e}_1\|_2 \|\mathbf{e}_2\|_2} + \lambda_s \sqrt{\frac{1}{r}\|\mathbf{P}_{s1}^T \mathbf{P}_{s2}\|_F^2} + \lambda_d \sqrt{\frac{1}{r}\|\mathbf{P}_{d1}^T \mathbf{P}_{d2}\|_F^2}$$

Projection Metric

Deep Face Representations

- 15-layer VGG-like DCNN model.
- Trained with cross-entropy loss, using the CASIA-WebFace dataset.
- We evaluated the proposed method on three datasets:
  1. Multiple Biometric Grand Challenge (MBGC),
  2. Face and Ocular Challenge Series (FOCS),
  3. IARPA Janus Benchmark A (IJB-A).

J.-C. Chen, V. M. Patel, R. Chellappa, “Unconstrained face verification using deep CNN features.” WACV 2016

Compared Baselines

- Methods:
  1. Dictionary-Based Face Recognition from Video (DFRV)
  2. Adaptive Video Dictionary Learning (AVDL)
  3. Structural Dictionary only (SDL)
  4. Linear Dynamical Dictionary only (LDDL)
  5. SDL+LDDL fusion (Hybrid)

Y.-C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. "Dictionary-based face recognition from video." ECCV, October 2012.

Y. C. Chen, V. M. Patel, P. J. Phillips, and R. Chellappa. "Dictionary-based face and person recognition from unconstrained video." IEEE Access, 2015.

X. Wei, H. Shen, and M. Kleinsteuber. "An adaptive dictionary learning approach for modeling dynamical textures." ICASSP 2014.

Compared Baselines

- Metrics:
  1. Cosine similarity on the global representations (Cosine)
  2. Reconstruction Error (RE)
- Representations:
  1. ArcFace (Arc-)
  2. Resnet-101 with Crystal Loss (CL-)

J. Deng, J. Guo, X. Niannan, and S. Zafeiriou. "Arcface: Additive angular margin loss for deep face recognition." In CVPR, 2019.
R. Ranjan, A. Bansal, **J. Zheng**, H. Xu, J. Gleason, B. Lu, A. Nanduri, J.-C. Chen, C. Castillo, and R. Chellappa. "A fast and accurate system for face detection, identication, and verication." TBIOM, 2019.

Verification Results on IJB-A

Methods	TAR@FAR		
	0.1%	1%	10%
Abd-Almageed *et al.*	-	78.70%	91.10%
Masi *et al.*	72.50%	88.60%	-
Sankaranarayanan *et al.*	81.30%	90.00%	96.40%
Cao *et al.*	92.10%	96.80%	**99.00%**
Xie *et al.*	92.00%	96.20%	***98.90%***
Shi *et al.*	**95.25%**	**97.50%**	-
Cosine	76.95%	88.73%	96.04%
DFRV_{deep} (Chen *et al.*)	58.55%	83.31%	93.83%
RE	64.63%	85.37%	94.35%
AVDL (Wei *et al.*)	34.86%	81.44%	94.83%
SDL	78.00%	89.60%	96.32%
LDDL	78.58%	89.67%	96.51%
Hybrid	78.30%	89.65%	96.45%
CL-Cosine (Ranjan *et al.*)	94.73%	97.01%	98.46%
CL-Hybrid	***95.04%***	***97.18%***	98.56%

IJB-MDF dataset

- The IARPA JANUS Benchmark Multidomain Face (IJB-MDF) dataset consists of images and videos of 251 subjects captured using a variety of cameras corresponding to
  - visible,
  - short-, mid-, and long-wave infrared,
  - long range surveillance domains
- There are 1,757 visible enrollment images, 40,597 short-wave infrared (SWIR) enrollment images and over 800 videos spanning 161 hours

Figure 1. Example images and frames from the different domains present within the IJB-MDF dataset.

Kalka, Nathan D., et al. "IARPA Janus benchmark multi-domain face." *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2019.

IJB-MDF dataset

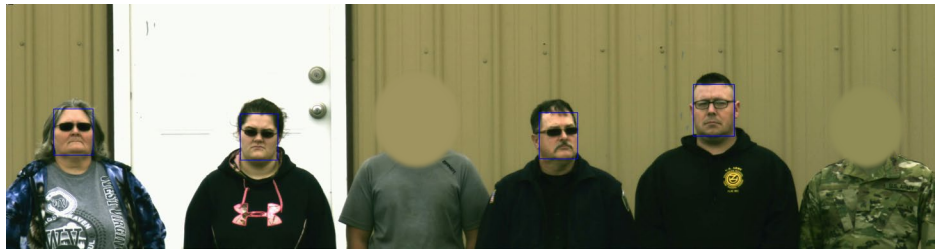
300m

400m

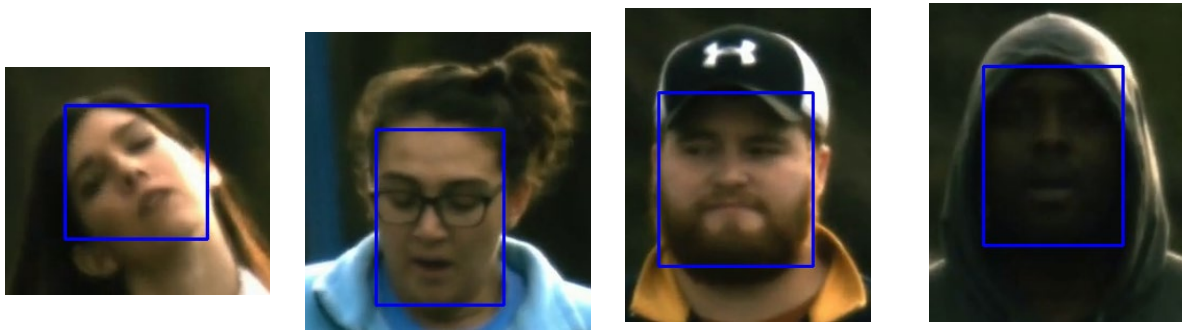
500m

IJB-MDF dataset

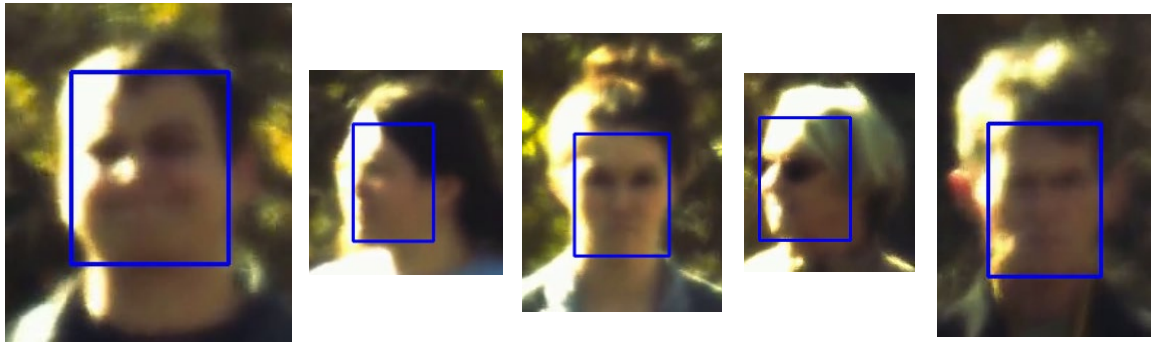
300m



400m



500m



Face Verification Evaluation Protocol

- Gallery – visible enrollment images
- Probe videos are from the following domains:
  - visible 500m
  - visible 400m
  - visible 300m
  - visible 500m 400m walking
- Every frame from the probe videos is compared against the gallery enrollment images separately

Implementation details

- Face detection
  - **SCRFD** algorithm used to detect faces in the video frames
  - Achieves a recall of about 95%
- Keypoint detection and alignment:
  - **AdaptiveWingLoss** algorithm is applied on the cropped faces to detect the face key-points
- Feature extraction:
  - We use a Resnet-101 model trained with **ArcFace** loss for feature extraction

1. Guo, Jia, et al. "Sample and computation redistribution for efficient face detection." *arXiv preprint arXiv:2105.04714* (2021).
2. Wang, Xinyao, Liefeng Bo, and Li Fuxin. "Adaptive wing loss for robust face alignment via heatmap regression." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.
3. Deng, Jiankang, et al. "Arcface: Additive angular margin loss for deep face recognition." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

Results

Domain	Rank 1	Rank 2	Rank 5	Rank 10
Visible 500m	20.8%	25.5%	33.2%	41.8%
Visible 400m	95.0%	97.1%	98.6%	99.1%
Visible 300m	**98.5%**	99.3%	99.7%	99.9%
Visible 500m 400m walking	57.8%	64.4%	71.6%	77.5%
All domains together	76.9%	79.5%	82.5%	85.1%

- The ArcFace model trained on visible images adapts well to the surveillance videos at 300m and 400m, but there is a huge drop in performance at 500m.

Related Works

1. Remote face recognition under atmospheric turbulence (FG 2020, T-BIOM2022)
2. Expressivity of covariates in deep features (FG 2020)
3. Adversarial learning for gender bias mitigation (ICCV2021).

Summary

1. Frame-level deep features with proper flattening are effective for video-based face recognition
2. Video face recognition can also be solved as an image set matching problem
3. Introducing temporal information in deep learning for vide-based face recognition is not fully developed.
    1. Pros and cons of incorporating motion?

Select List of Publications

- **R. Ranjan**, C. D. Castillo and R. Chellappa, “L2-constrained Softmax Loss for Discriminative Face Verification,” arXiv preprint arXiv:1703.09507
- **R. Ranjan**, S. Sankaranarayanan, C. D. Castillo and R. Chellappa, “An all-in-one convolutional neural network for face analysis,” in *Automatic Face & Gesture Recognition **(FG)**, 2017*. (**Oral**)
- **R. Ranjan**, V. M. Patel and R. Chellappa, “HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (**PAMI**), 2017*.
- **R. Ranjan,** J-C Chen, S. Sankaranarayanan, A. Kumar, C-H Chen, V. M. Patel, C. D. Castillo and R. Chellappa, “An End-to-End System for Unconstrained Face Verification with Deep Convolutional Neural Networks, ” in *International Journal of Computer Vision **(IJCV)**, 2017.*
- **Jingxiao Zheng**, Ruichi Yu, Jun-Cheng Chen, Boyu Lu, Carlos D. Castillo and Rama Chellappa. "Uncertainty Modeling of Contextual-Connections between Tracklets for Unconstrained Video-based Face Recognition." **ICCV 2019**.
- **Jingxiao Zheng**, Jun-Cheng Chen, Vishal M. Patel, Carlos D. Castillo and Rama Chellappa. "Hybrid Dictionary Learning and Matching for Video-based Face Verification." BTAS 2019.
- **Jingxiao Zheng**, Rajeev Ranjan, Ching-Hui Chen, Jun-Cheng Chen, Carlos D. Castillo and Rama Chellappa. "An Automatic System for Unconstrained Video-Based Face Recognition." IEEE Transactions on Biometrics, Behavior, and Identity Science, 2020.