Unraveling Representations for Face Recognition: from Handcrafted to Deep Learning

GAURAV GOSWAMI

IIIT-DELHI, INDIA

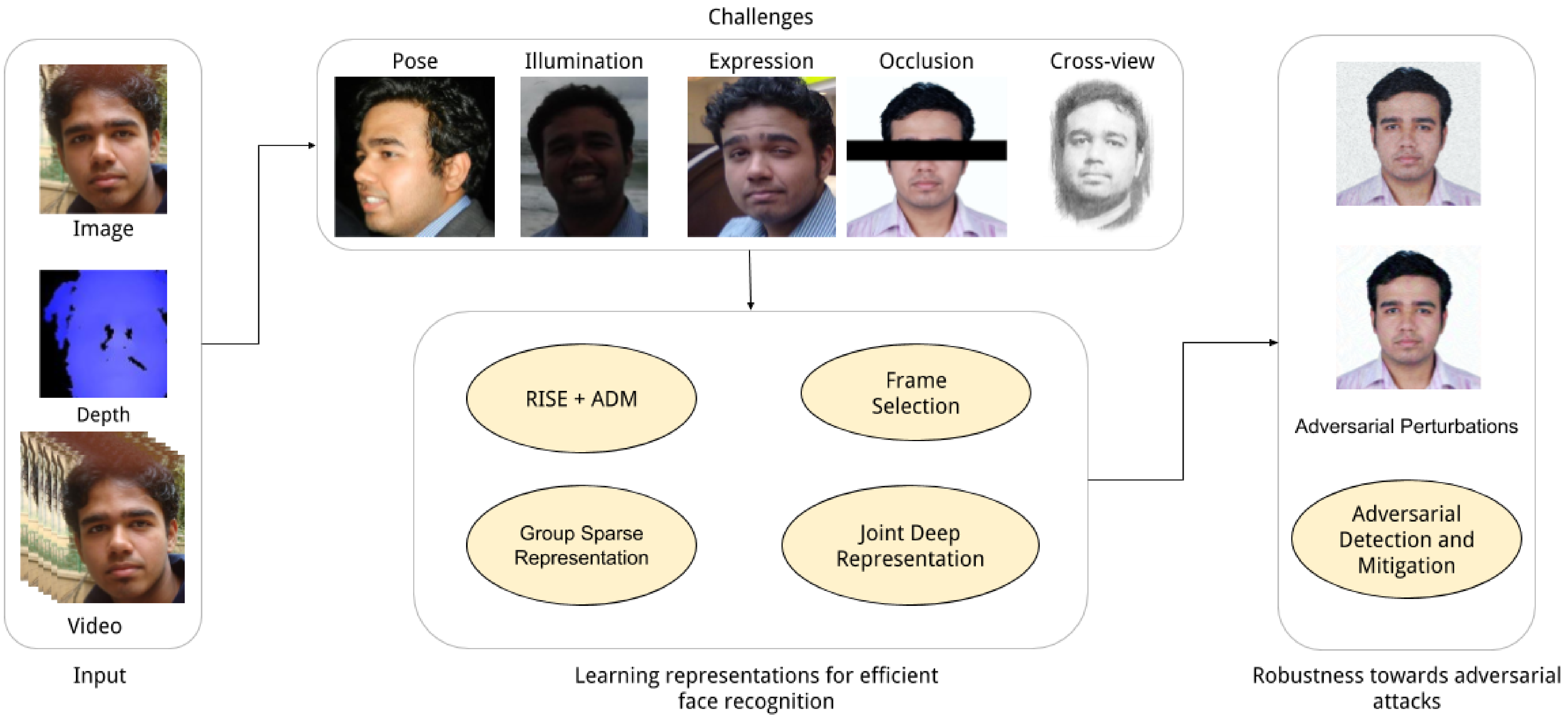
Face recognition

- Advantages:
  - Human perception and cognitive understanding
  - Does not require cooperation from the subject
  - Does not require specialized capture process and/or sensor
  - Forensic/law enforcement applications: sketch recognition, surveillance

(a) Pose (b) Occlusion (c) Expression (d) Illumination (e) Disguise (f) Resolution (g) Age (h) Spectrum

Images are taken from the AR face database (Martinez, 1998), the CMU Multi-PIE database (Gross et al, 2010), the SCFace database (Grgic, 2011), the Large Age Gap (LAG) database (Bianco, 2017), and the KaspAROV database (Chhokra, 2018)

Dissertation contributions overview



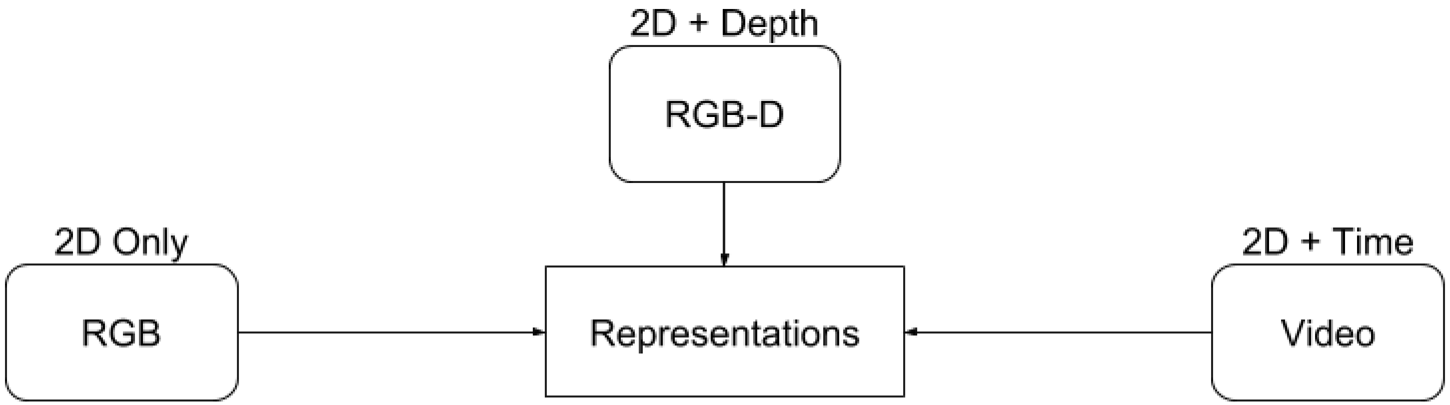
Progression of face recognition

2007-2010
- Introduction of the LFW database (unconstrained)
- Handcrafted feature ensembles with metric learning

2011-2013
- Introduction of the YTF database (unconstrained video)
- Handcrafted feature ensembles with metric learning

2014-Present
- Focus on deep learning (primarily CNNs) and hybrids
- Doubts on the robustness of deep learning

Research contributions

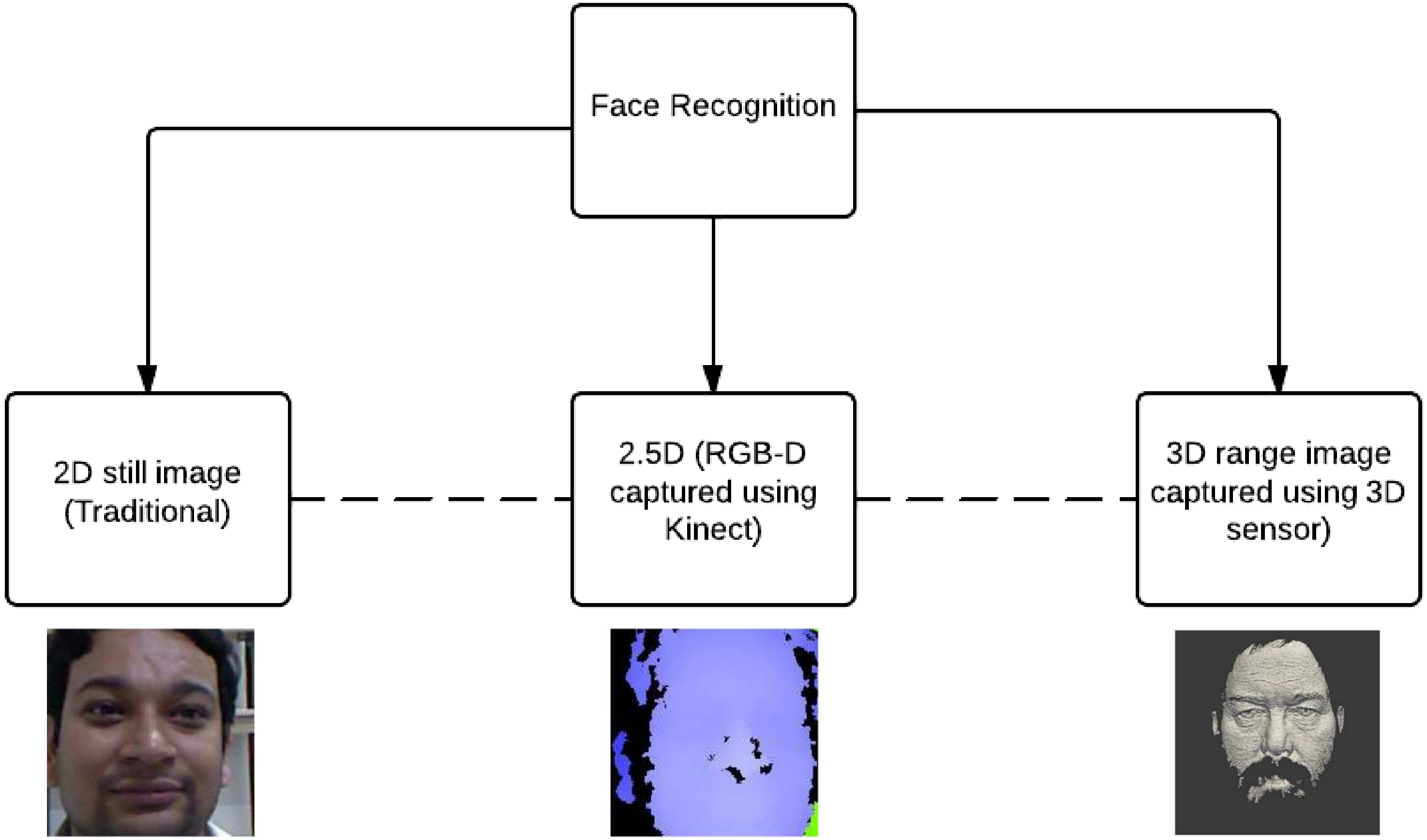


- Using depth data for improved face feature representations
- Combining multiple feature representations
- Using video data for improved feature representations
- Learning data-driven feature representations
- Evaluating and addressing the robustness of deep representations against adversaries

RGB-D face recognition

Objective: Use depth data from low cost sensors to obtain improved representations for face recognition

RGB-D face recognition



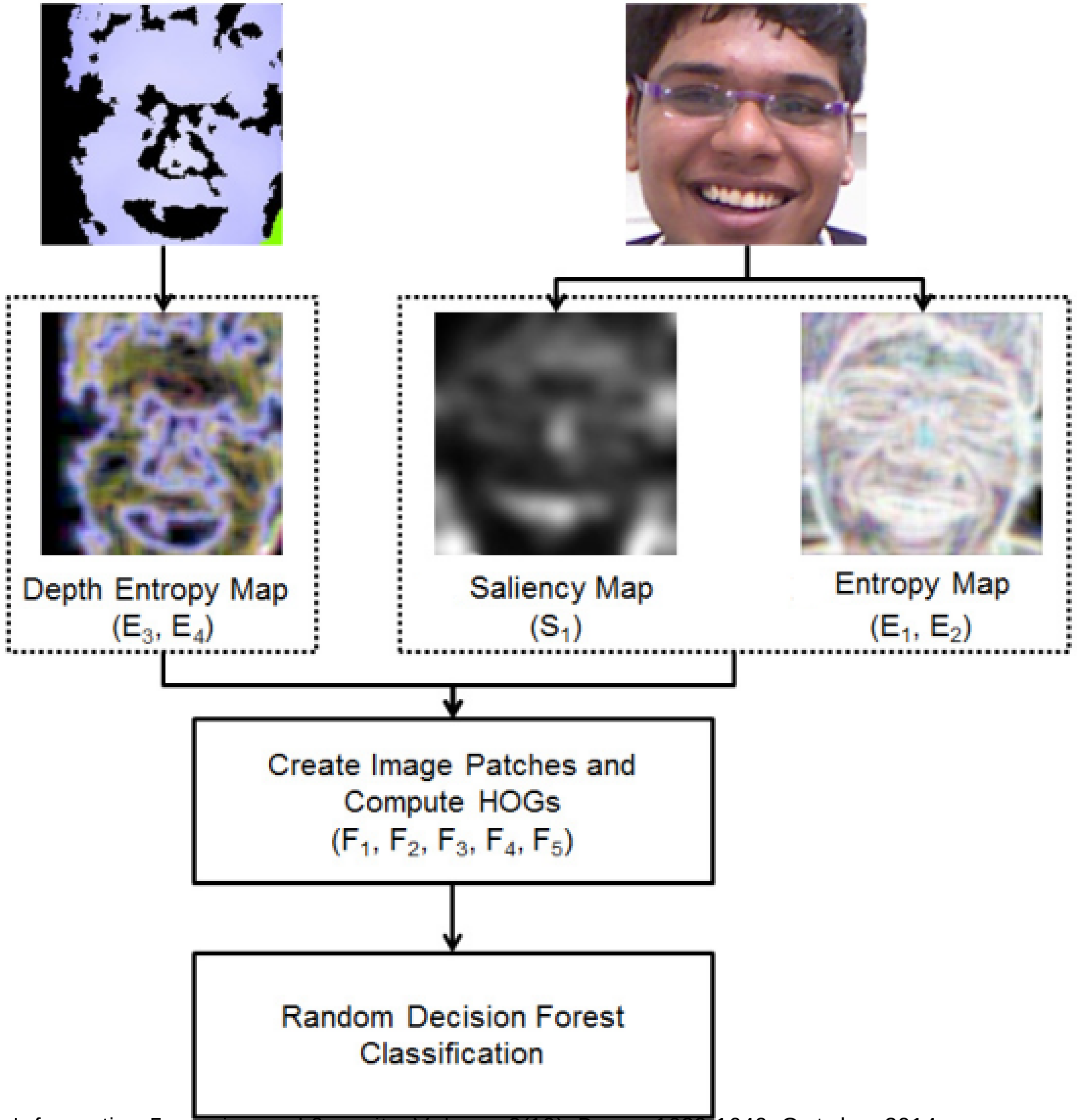
G. Goswami, M. Vatsa, and R. Singh, RGB-D Face Recognition with Texture and Attribute Features, IEEE Transactions on Information Forensics and Security, Volume 9(10), Pages 1629-1640, October 2014

Comparison of 3D data acquisition devices

Device	Speed (sec)	Size ($inch^3$)	Price (USD)	Acc. (mm)
3dMD	0.002	N/A	>$50k	<0.2
Minolta	2.5	1408	>$50k	~0.1
Artec Eva	0.063	160.8	>$20k	~0.5
3D3 HDI R1	1.3	N/A	>$10k	>0.3
SwissRanger	0.02	17.53	>$5k	~10
DAVID SLS	2.4	N/A	>$2k	~0.5
Kinect	0.033	11 * 3 * 3	<$200	~1.5-50
Intel D415	0.01	3.9 * 0.79 * 0.9	<$150	~2.5-20

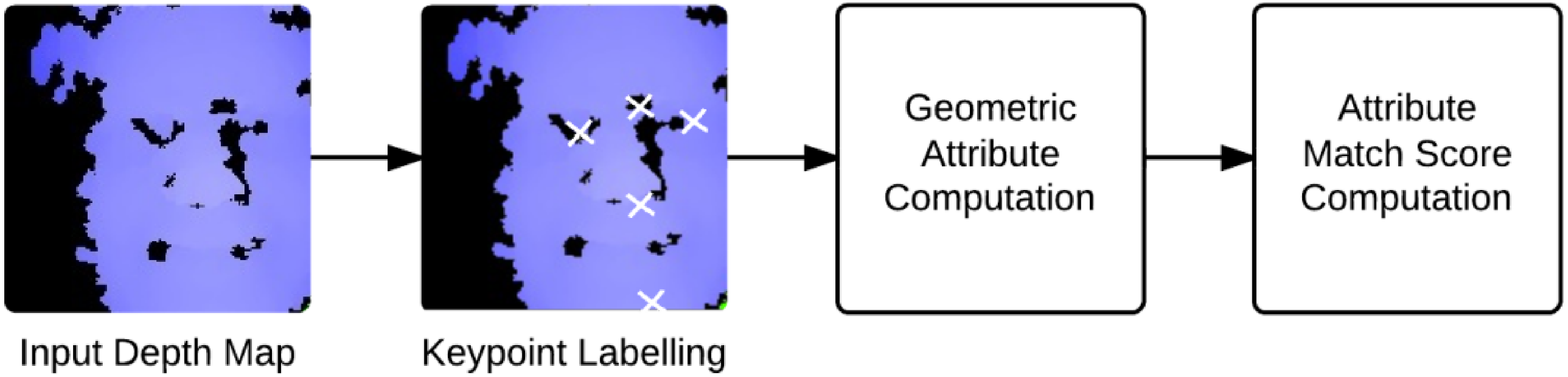
Face recognition using Kinect: RISE algorithm

- **R**GB-D **I**mage descriptor based on **S**aliency and **E**ntropy (RISE)
- Entropy is used to enhance the features of the face image and the depth map.
- Saliency provides additional features.
- Using various image patches helps to capture features at different granularities.
- HOG extracts robust and fixed length feature vector.
- Random Decision Forest classifier uses this vector in testing/training.



G. Goswami, M. Vatsa, and R. Singh, RGB-D Face Recognition with Texture and Attribute Features, IEEE Transactions on Information Forensics and Security, Volume 9(10), Pages 1629-1640, October 2014

Face recognition using Kinect: ADM algorithm



- **A**ttributes based on **D**epth **M**ap (ADM)
- Rule template based on depth data and uniform structure of human face -> landmark points
- Various geometric attributes based on distance between landmark points

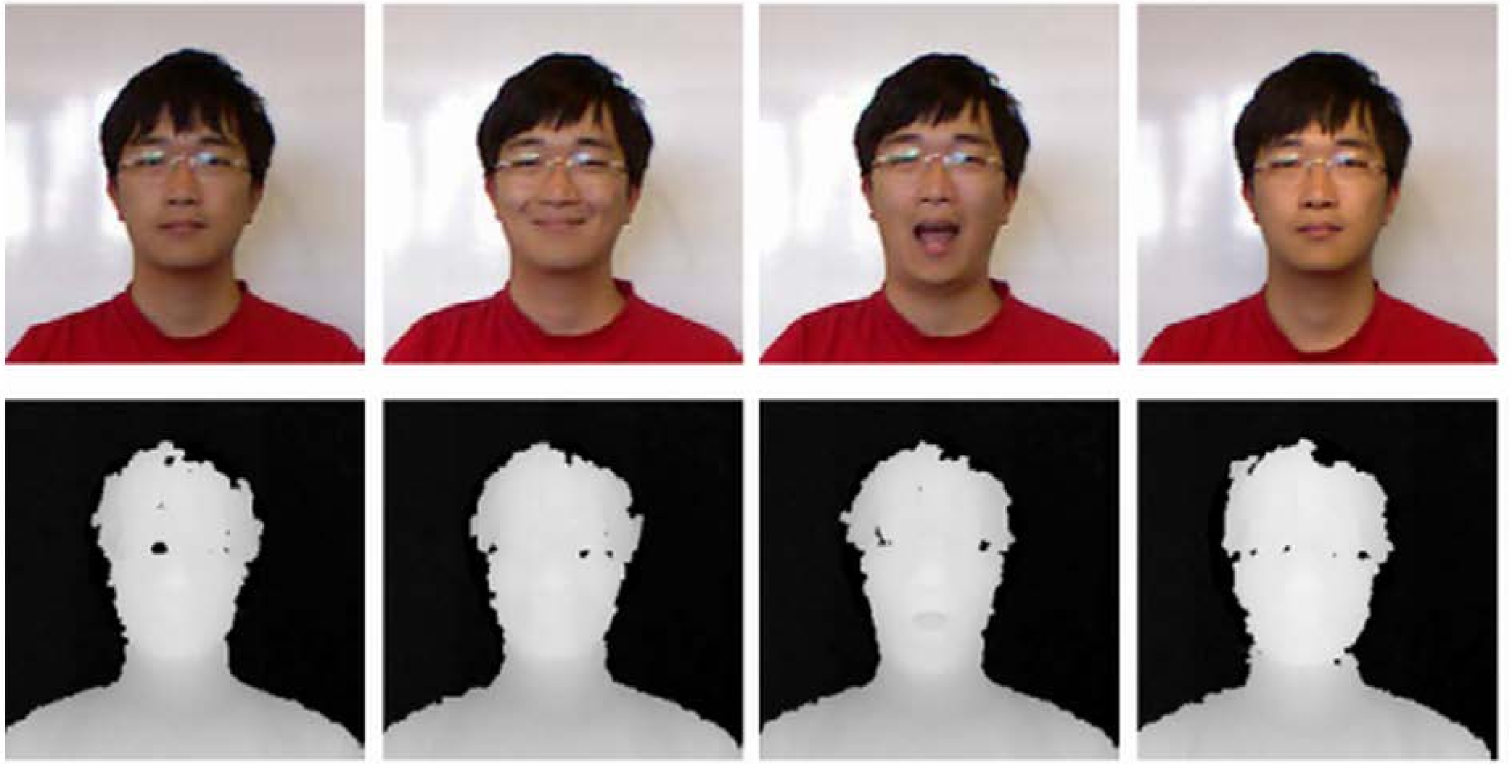
The IIIT-D Kinect RGB-D face database

- 106 subjects, over 4600 images pertaining to 2 sessions.

G. Goswami, M. Vatsa, and R. Singh, RGB-D Face Recognition with Texture and Attribute Features, IEEE Transactions on Information Forensics and Security, Volume 9(10), Pages 1629-1640, October 2014

EURECOM Kinect face database

- 936 images pertaining to 52 subjects and captured in two sessions.



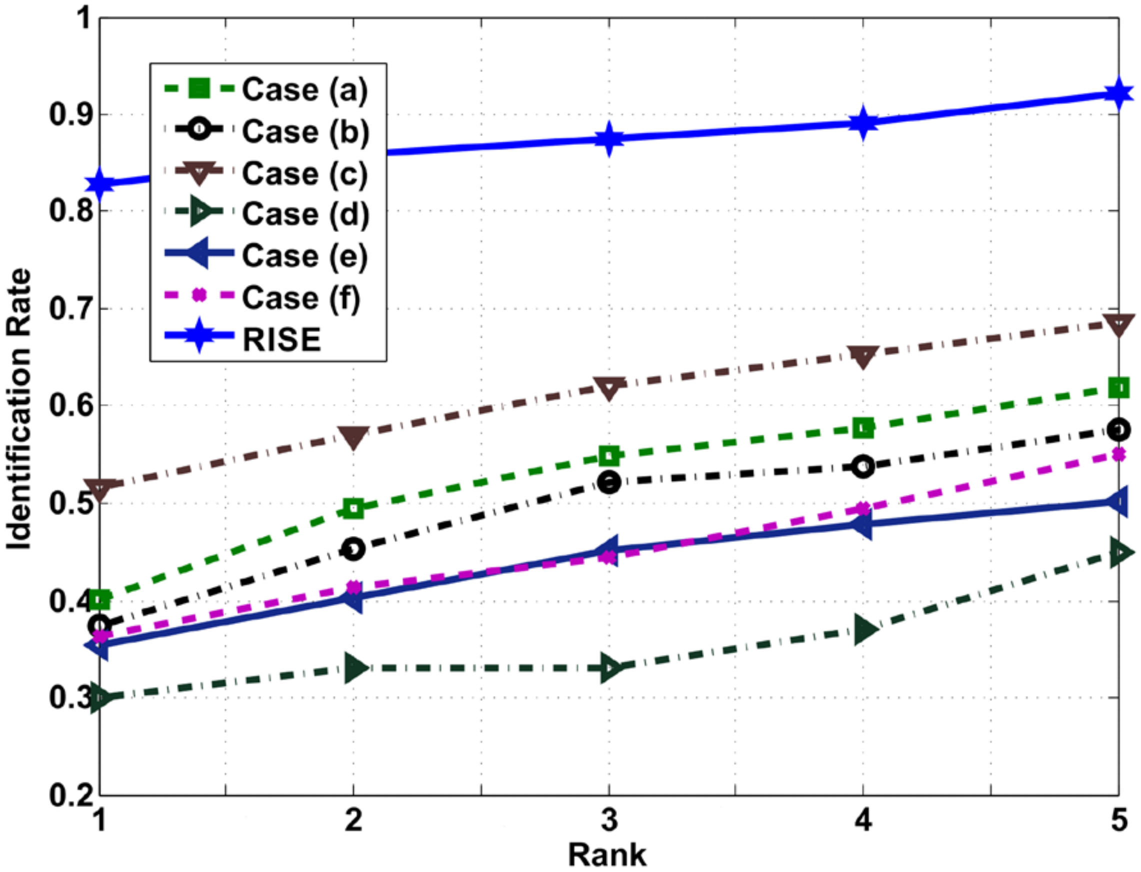
T. Huynh, R. Min, and J. L. Dugelay. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In Asian Conference on Computer Vision, 2012.

Experimental protocol

Experiment	Database	No. of Images	No. of Subjects	
			Training	Testing
Experiment 1	IIIT-D RGB-D	4605	42	64
Experiment 2 (Extended database)	IIIT-D RGB-D + VAP + EURECOM	5694	75	114

G. Goswami, M. Vatsa, and R. Singh, RGB-D Face Recognition with Texture and Attribute Features, IEEE Transactions on Information Forensics and Security, Volume 9(10), Pages 1629-1640, October 2014
R. I. Hg, P. Jasek, C. Rofidal, K. Nasrollahi, T. B. Moeslund, and G. Tranchet. An RGB-D database using Microsoft's kinect for windows for face detection. In International Conference on Signal Image Technology and Internet Based Systems, pages 42–46, 2012.
T. Huynh, R. Min, and J. L. Dugelay. An efficient LBP-based descriptor for facial depth images applied to gender recognition using RGB-D face data. In Asian Conference on Computer Vision, 2012.

Contribution of individual components



- Case (a) RGB-D and saliency without entropy
- Case (b) RGB only
- Case (c) RGB-D only
- Case (d) RGB and saliency without entropy
- Case (e) RGB-D only without entropy
- Case (f) RGB only

G. Goswami, M. Vatsa, and R. Singh, RGB-D Face Recognition with Texture and Attribute Features, IEEE Transactions on Information Forensics and Security, Volume 9(10), Pages 1629-1640, October 2014

Comparison on extended dataset

Modality	Descriptor	Rank 1	Rank 5
2D	SIFT	55.3 ± 1.7	72.8 ± 2.1
	HOG	58.8 ± 1.4	76.3 ± 1.8
	PHOG	60.5 ± 1.6	78.1 ± 1.1
	FPLBP	64.0 ± 1.1	80.7 ± 2.0
	Sparse	65.8 ± 0.6	84.2 ± 0.8
3D	3D-PCA	67.5 ± 1.2	82.5 ± 1.9
	RISE+ADM (W.B.C.)	76.3 ± 1.0	90.3 ± 1.1
	RISE+ADM (W.S.)	$\mathbf{78.9} \pm \mathbf{1.7}$	$\mathbf{92.9} \pm \mathbf{1.3}$

Bowyer et al. 2004, Dalal and Triggs 2005, Wolf et al. 2008, Bai et al. 2009,Wright et al. 2009, Goswami et al. 2014

RGB-D face recognition: outcomes

Journal Article

- G. Goswami, M. Vatsa, and R. Singh, RGB-D Face Recognition with Texture and Attribute Features, IEEE Transactions on Information Forensics and Security, Volume 9(10), Pages 1629-1640, October 2014.

Conference Article

- G. Goswami, S. Bharadwaj, M. Vatsa, and R. Singh, On RGB-D Face Recognition using Kinect, 6th IEEE International Conference on Biometrics: Theory, Applications and Systems, 2013 **(Received the Best Poster Award)**.

Book Chapter

- G. Goswami, M. Vatsa, and R. Singh, Face Recognition with RGB-D images using Kinect, in Face Recognition across the Imaging Spectrum, Springer International Publishing, 2016, pp. 281-303.

Group Sparse Classifier

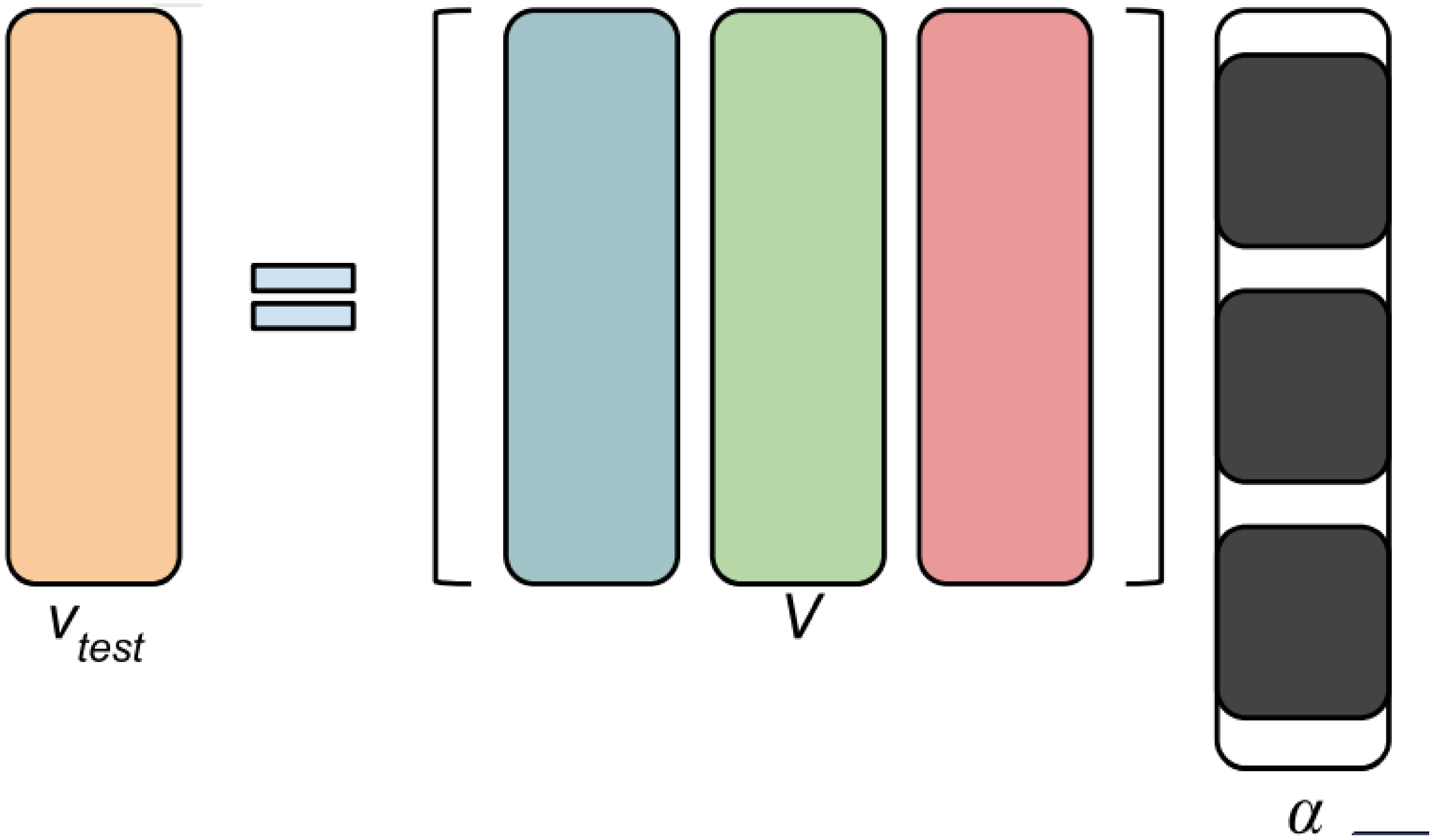
Objective: Leverage multiple feature representations obtained using different feature extractors, modalities, and input types

Feature-level fusion

- Advantages:
  - Less prone to noise as compared to sensor-level
  - Preserves more information than score/rank/decision level
- Challenges:
  - Relationships between features are unknown
  - Variable/fixed length of features
  - Feature compatibility

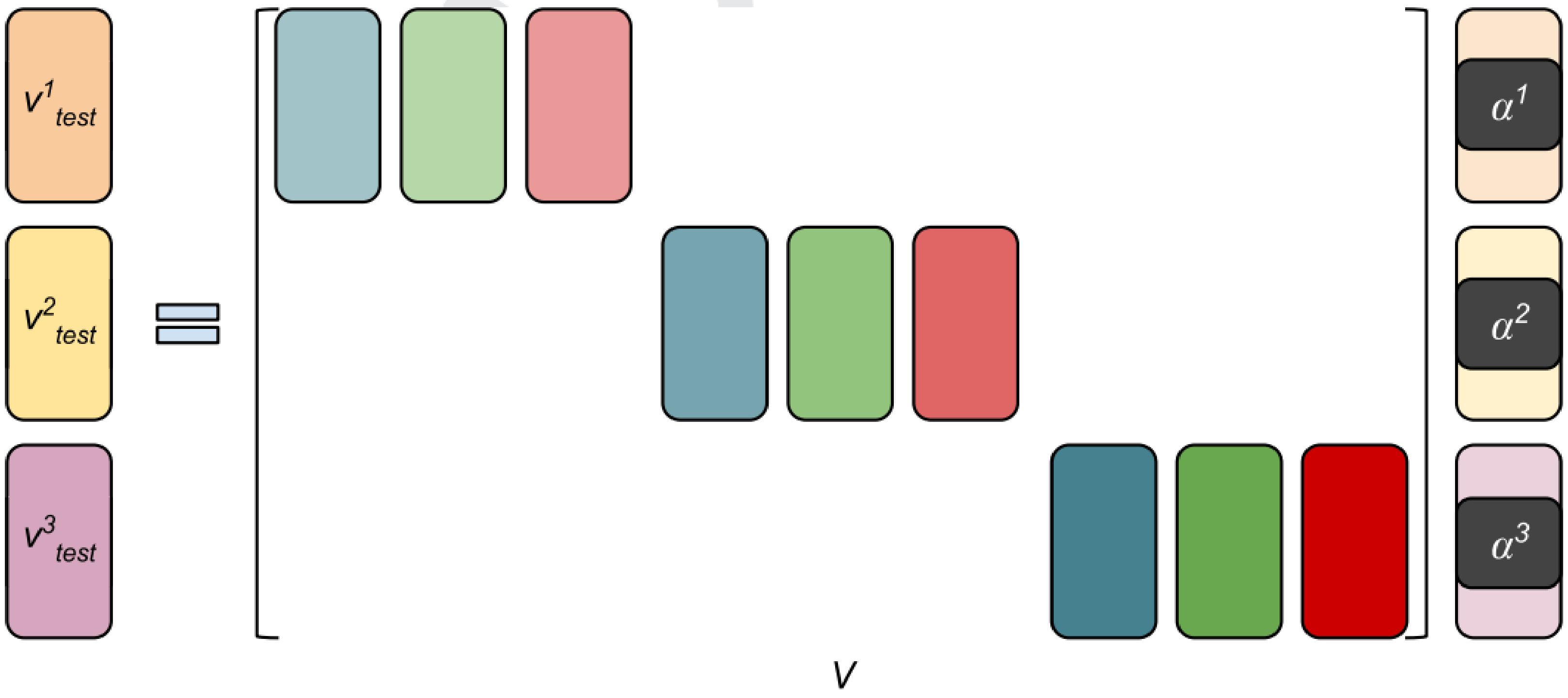
Naïve approach: Concatenation followed by feature selection/reduction

Sparse Representation based Classification (SRC)



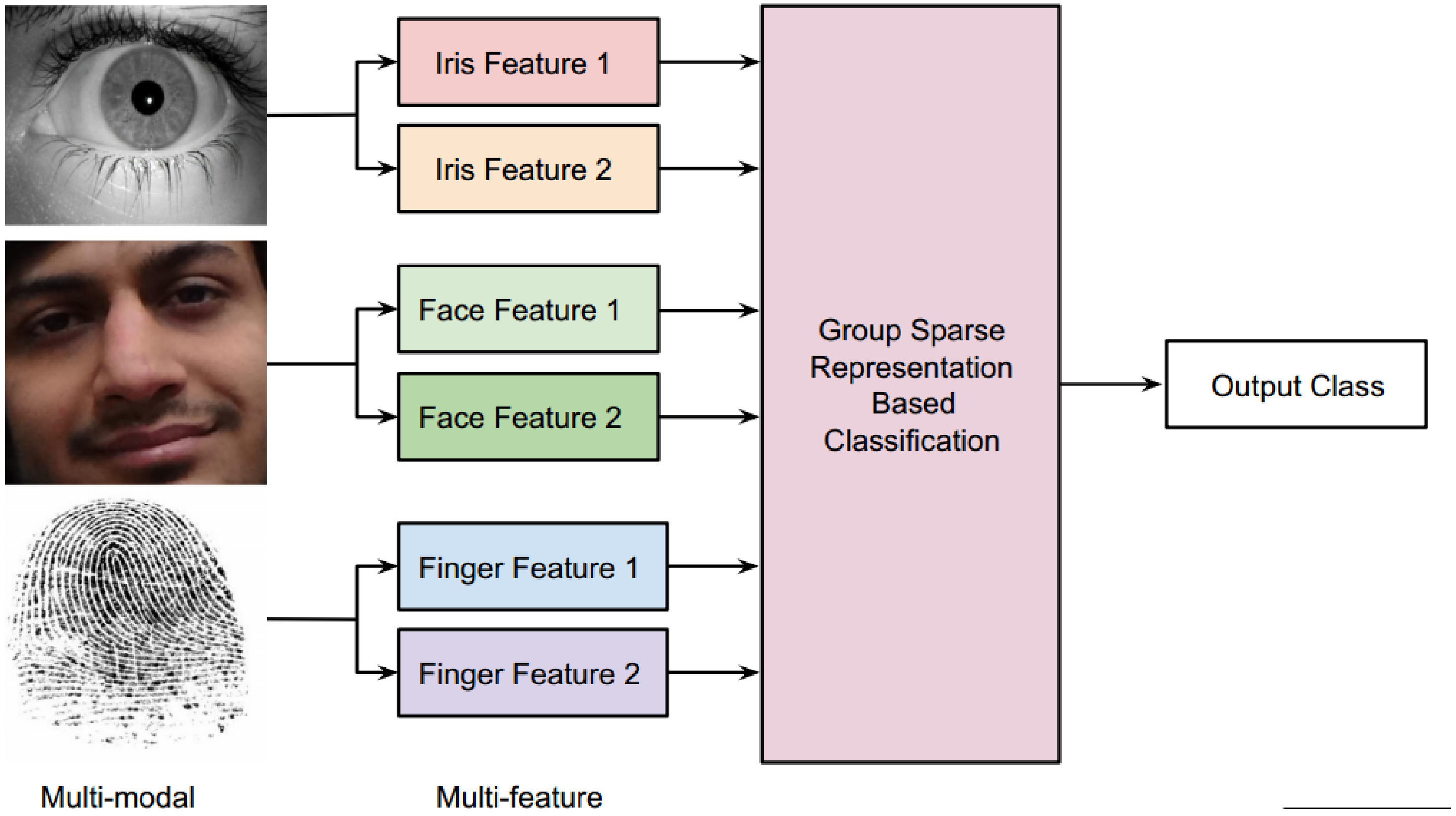
G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, Group Sparse Representation based Classification for Multi-feature Multimodal Biometrics, Information Fusion, Volume 32(B), Pages 3-12, 2016

Proposed Group Sparse Classifier



G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, Group Sparse Representation based Classification for Multi-feature Multimodal Biometrics, Information Fusion, Volume 32(B), Pages 3-12, 2016

Group Sparse Classifier for multi-modal biometrics



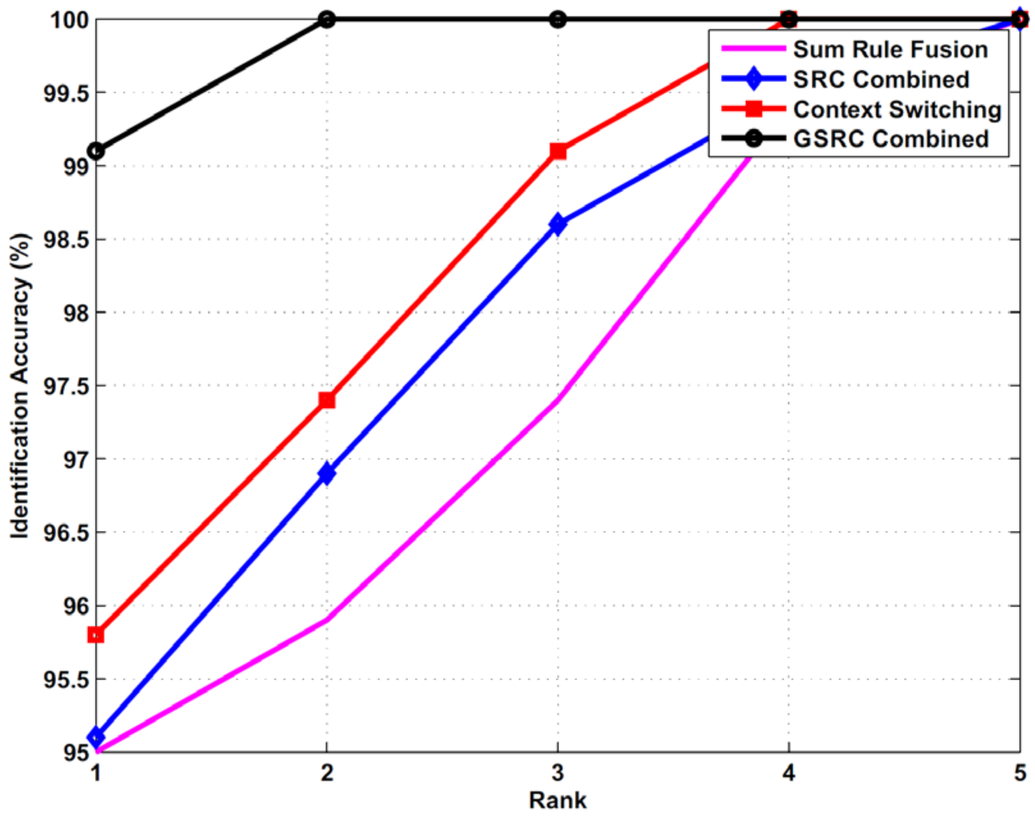
G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, Group Sparse Representation based Classification for Multi-feature Multimodal Biometrics, Information Fusion, Volume 32(B), Pages 3-12, 2016

Databases and protocol

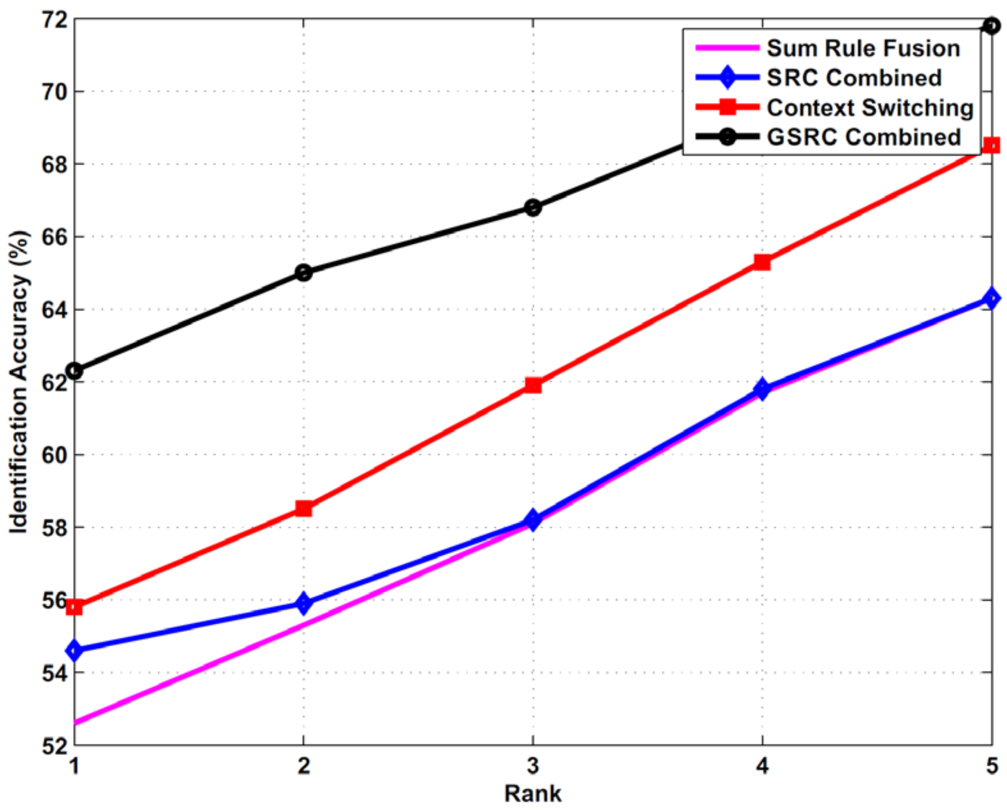
Database	Modalities	Subjects	Protocol	
			Training	Testing
WVU	Iris, fingerprint, palmprint, hand geometry, face video and voice, face	270	108 subjects	162 subjects
LEA	Face, fingerprint, iris	18,000	9000 subjects	9000 subjects

Protocol taken from: S. Bharadwaj, H. S. Bhatt, R. Singh, M. Vatsa, and A. Noore. QFuse: Online Learning Framework for Adaptive Biometric System. Pattern Recognition, 48(11):3428 – 3439, 2015

Results



WVU



LEA

Wright et al. 2009, Bhardwaj et al. 2015, Goswami et al. 2014

Group Sparse Classifier: outcomes

Journal Article

- G. Goswami, P. Mittal, A. Majumdar, M. Vatsa, and R. Singh, Group Sparse Representation based Classification for Multi-feature Multimodal Biometrics, Information Fusion, Volume 32(B), Pages 3-12, 2016.

Conference Article

- G. Goswami, R. Singh, M. Vatsa, A. Majumdar, Kernel Group Sparse Representation based Classifier for Multimodal Biometrics, 30th International Joint Conference on Neural Networks, 2017.

Video face recognition

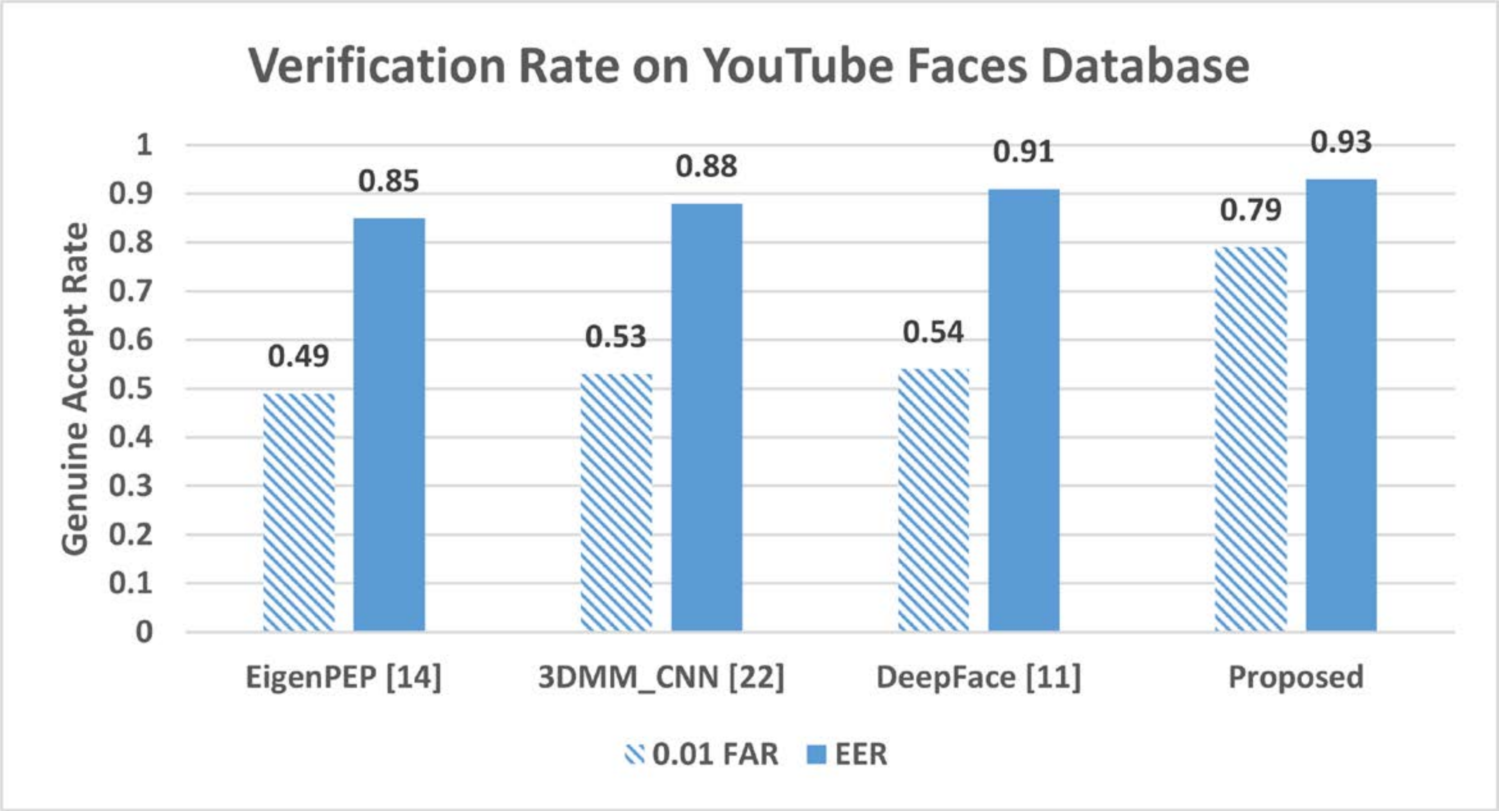
Objective: Extract the most “useful” frames from face videos and extract discriminative information from these frames using data-driven learned representations

Video face recognition

- More informative and also more challenging
- Lot of information, but how much is relevant?

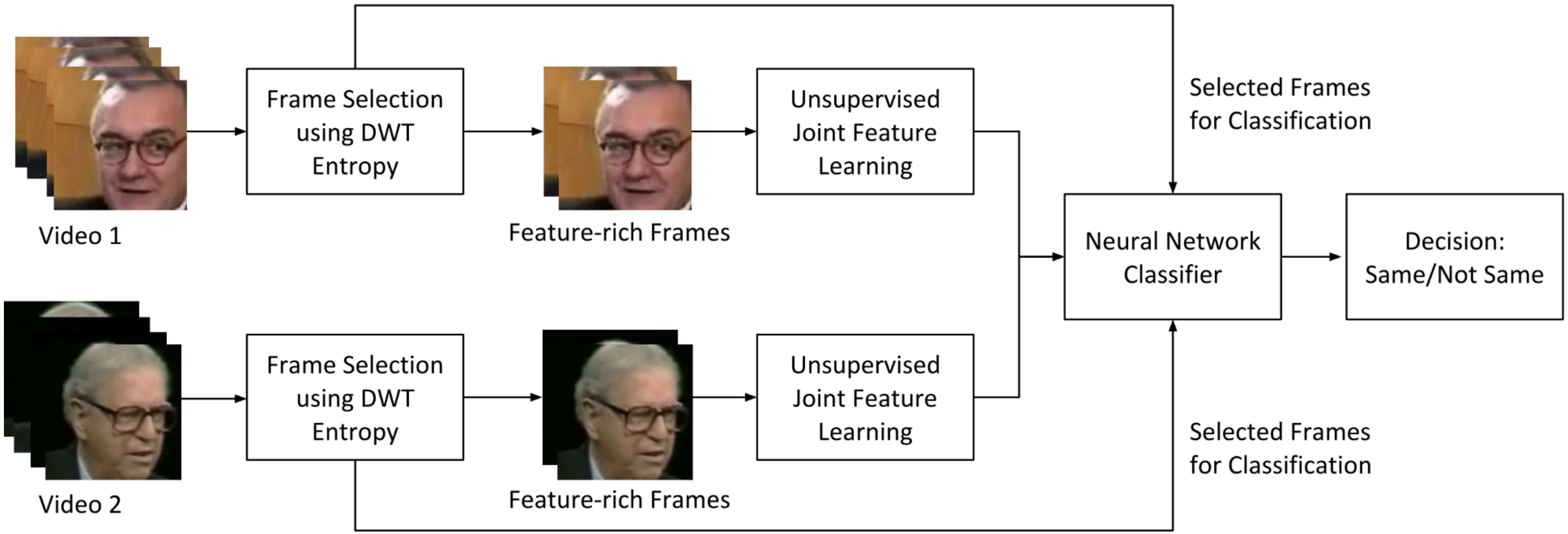
Image sources: the YTF (L. Wolf, T. Hassner and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In IEEE Conference on Computer Vision and Pattern Recognition, pages 529–534, 2011) and PaSC (J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In IEEE Conference on Biometrics: Theory, Applications and Systems, pages 1–8, 2013) databases.

Gaps in existing video face recognition



Li et al. 2014, Taigman et al. 2014, Tran et al. 2016, Goswami et al. 2017

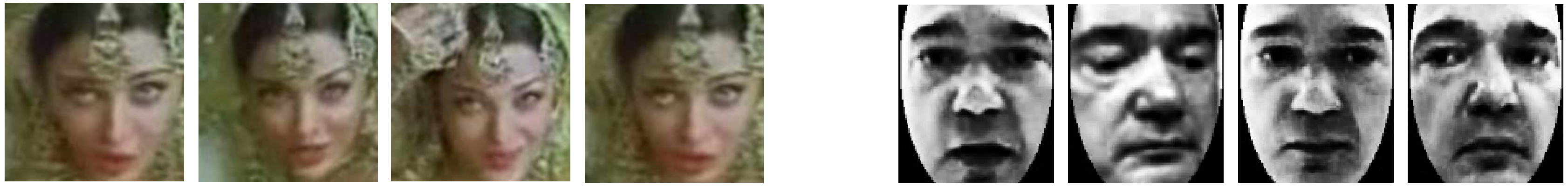
Proposed algorithm



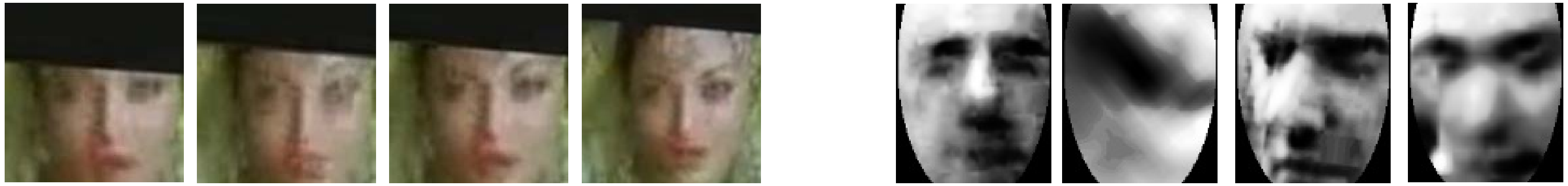
G. Goswami, M. Vatsa, and R. Singh. Face verification via learned representation on feature-rich video frames. IEEE Transactions on Information Forensics and Security, 12:1686–1698, 2017

Feature-richness based frame selection

Most feature-rich



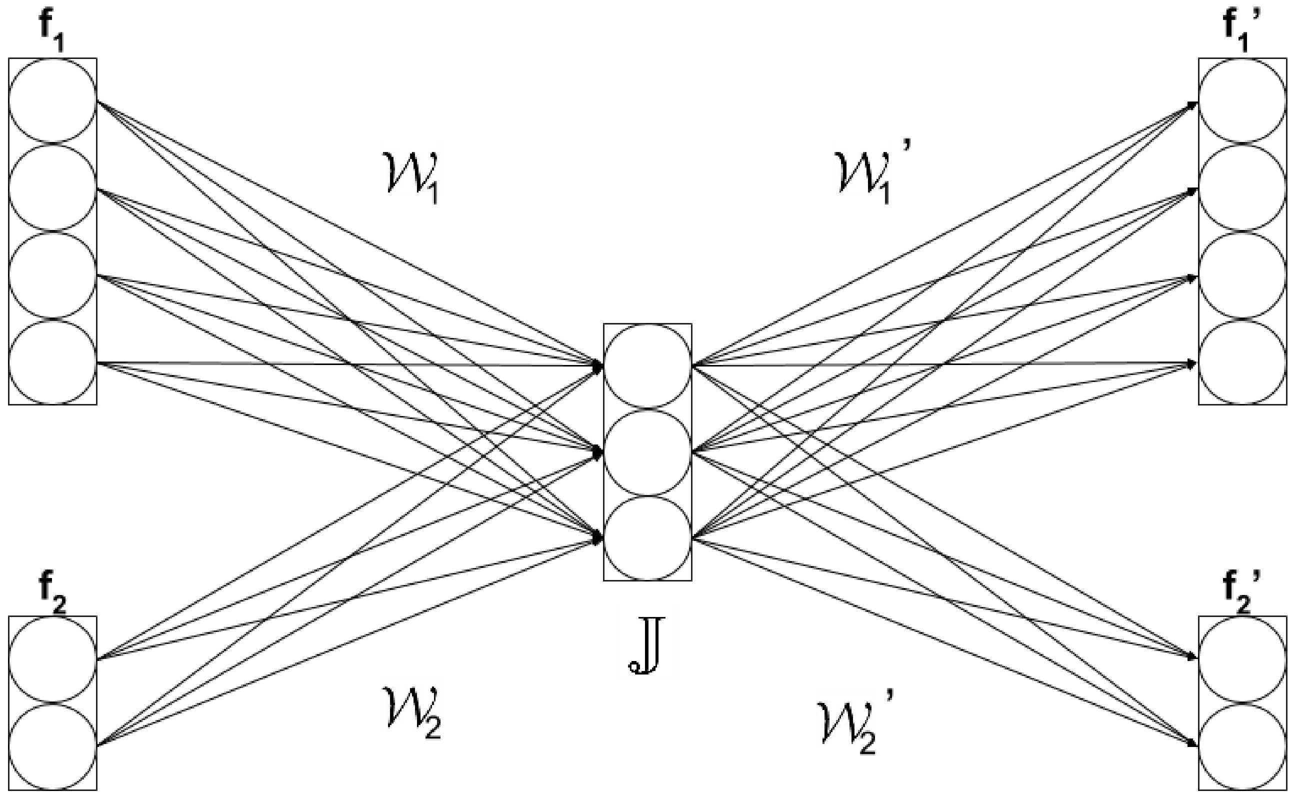
Least feature-rich



- Helps avoid bad frames
- Recognition oriented frame selection

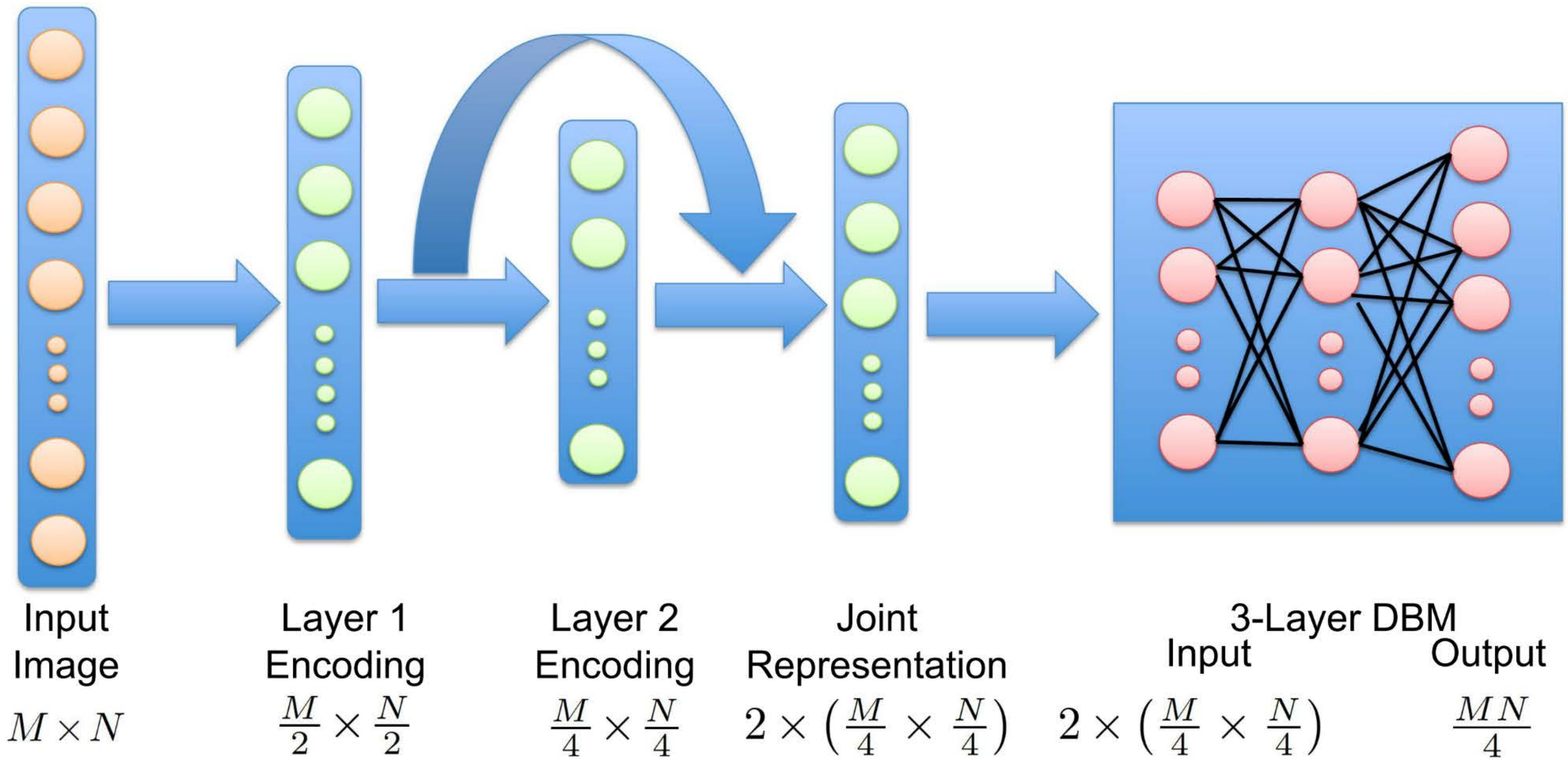
Image sources: the YTF (L. Wolf, T. Hassner and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In IEEE Conference on Computer Vision and Pattern Recognition, pages 529–534, 2011) and PaSC (J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In IEEE Conference on Biometrics: Theory, Applications and Systems, pages 1–8, 2013) databases.

Deep learning architecture: joint representation framework



G. Goswami, M. Vatsa, and R. Singh. Face verification via learned representation on feature-rich video frames. IEEE Transactions on Information Forensics and Security, 12:1686–1698, 2017

Deep learning architecture: overview



G. Goswami, M. Vatsa, and R. Singh. Face verification via learned representation on feature-rich video frames. IEEE Transactions on Information Forensics and Security, 12:1686–1698, 2017

Deep learning architecture: SDAE+DBM

- SDAE provides low-level features that are robust to noise in the data
- DBM can then extract progressively higher level features better suited for recognition
- Updated RBM loss function:

$$\mathcal{L}_{new} = \mathcal{L} + \mathcal{A} \parallel W \parallel_1 + \mathcal{B} \parallel W \parallel_{\tau}$$

- L-1 norm ensures sparsity in features whereas trace-norm ensures low-rankness

Databases

Database	No. of		Average no. of	
	Subjects	Videos	Videos per subject	Frames per video
YouTube Faces	1595	3425	2	181
PaSC (Handheld)	265	1401	4 to 7	235
PaSC (Control)	265	1401	4 to 7	239

- Pre-defined benchmark protocol
- Face detection and alignment using provided bounding box data

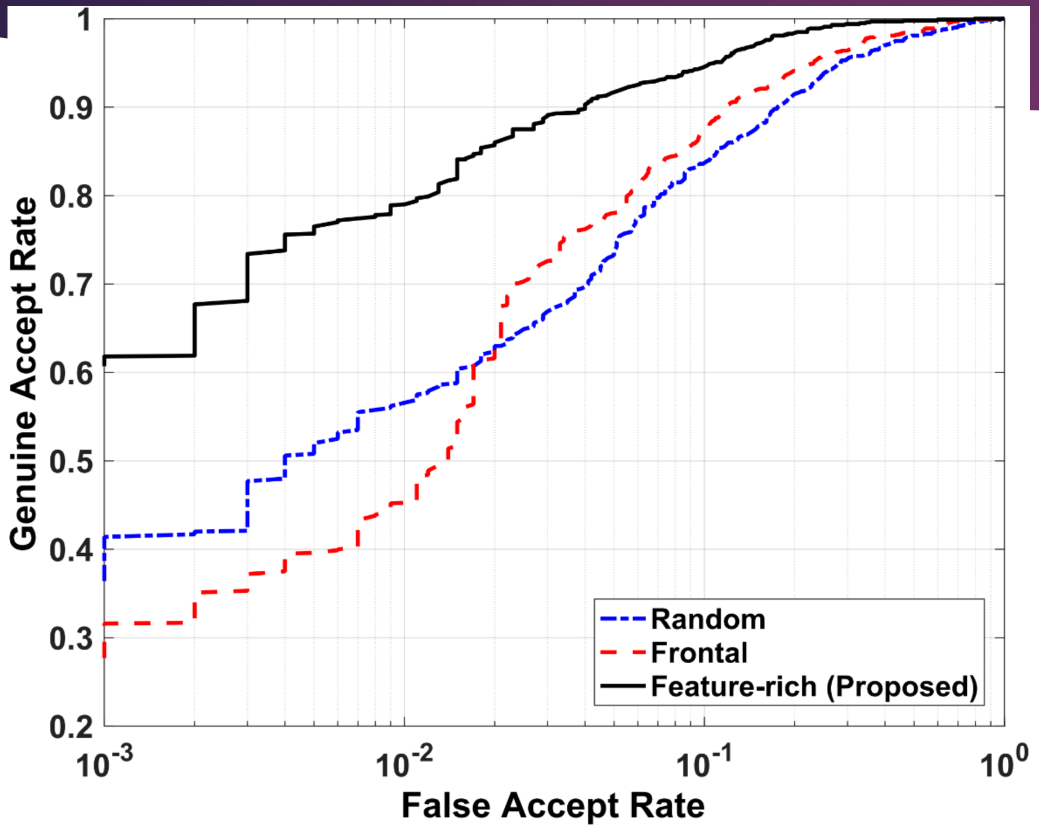
L. Wolf, T. Hassner and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In IEEE Conference on Computer Vision and Pattern Recognition, pages 529–534, 2011.
J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In IEEE Conference on Biometrics: Theory, Applications and Systems, pages 1–8, 2013.

Impact of frame selection

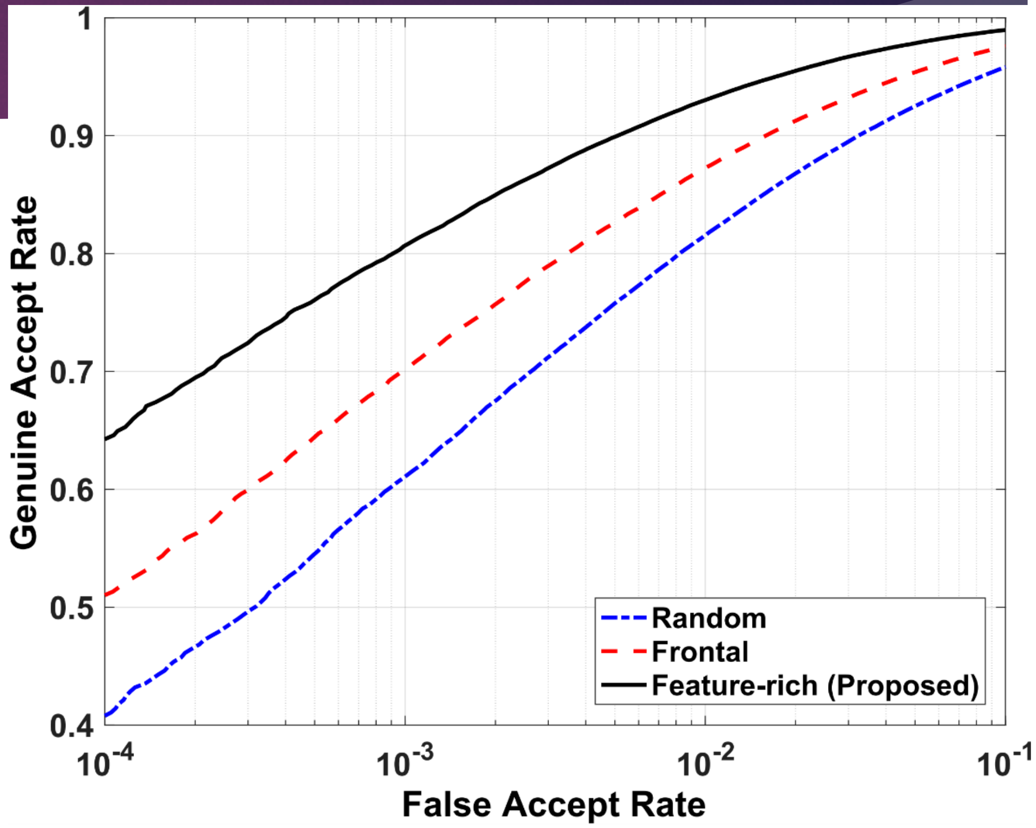
Frame Selection	Algorithm	GAR at 0.01 FAR		
		YTF	PaSC (Handheld)	PaSC (Control)
All		0.74	0.89	0.92
Image Quality	BRISQUE	0.62	0.82	0.84
	NIQE	0.62	0.83	0.82
	SSEQ	0.62	0.82	0.82
Memorability	MDLFace	0.69	0.89	0.94
Proposed feature-richness	25 frames	0.75	0.91	0.94
	50 frames	0.77	0.91	0.93
	Adaptive	**0.79**	**0.93**	**0.96**

Mittal et al. 2012, Mittal et al. 2013, Liu et al. 2014, Goswami et al. 2014

Impact of frame selection



YouTube
Faces



PaSC
(Handheld)

L. Wolf, T. Hassner and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In IEEE Conference on Computer Vision and Pattern Recognition, pages 529–534, 2011.
J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In IEEE Conference on Biometrics: Theory, Applications and Systems, pages 1–8, 2013.

Results and comparison

Algorithm	External Data	Layers	YTF (at EER)	PaSC (at 1% FAR) Control	PaSC (at 1% FAR) Handheld
Trunk-Branch Ensemble CNNs with Batch Normalization [39]#	2.68 Million$	18 + 11 + 11*	94.9	98.0	97.0
VGG Face [153]+	2.62 Million	21	**97.4**	91.3	87.0
GoogLeNet [186] features with aggregation [216]	3 Million	22	95.5	-	-
CNN-3DMM Estimation [191]	0.49 Million	101	88.8	-	-
Proposed SDAE-DBM Joint Representation	No	9	93.4	95.9	93.1
	YTF + PaSC	9	95.0	96.6	96.1
	2.48 Million	9	95.4	**98.1**	**97.2**

Parkhi et al. 2015, Tran et al. 2016, Yang et al. 2016, Ding and Tao 2017, Goswami et al. 2017

Video face recognition: outcomes

Journal Article

- G. Goswami, M. Vatsa, and R. Singh, Video Face Verification via Learned Representation on Feature-Rich Frames, IEEE Transactions on Information Forensics and Security, Volume 12(7), Pages 1686-1698, 2017.

Conference Article

- G. Goswami, R. Bhardwaj, M. Vatsa, and R. Singh, MDLFace: Memorability augmented deep learning for video face recognition, IEEE/IAPR International Joint Conference on Biometrics, 2014. **(Oral Presentation)**

Book Chapter

- T.I. Dhamecha, G. Goswami, R. Singh, and M. Vatsa, On Frame Selection for Video Face Recognition, in Advances in Face Detection and Facial Image Analysis, Springer International Publishing, 2016, pp. 279-297.

Adversarial attacks on deep learning

Objectives:

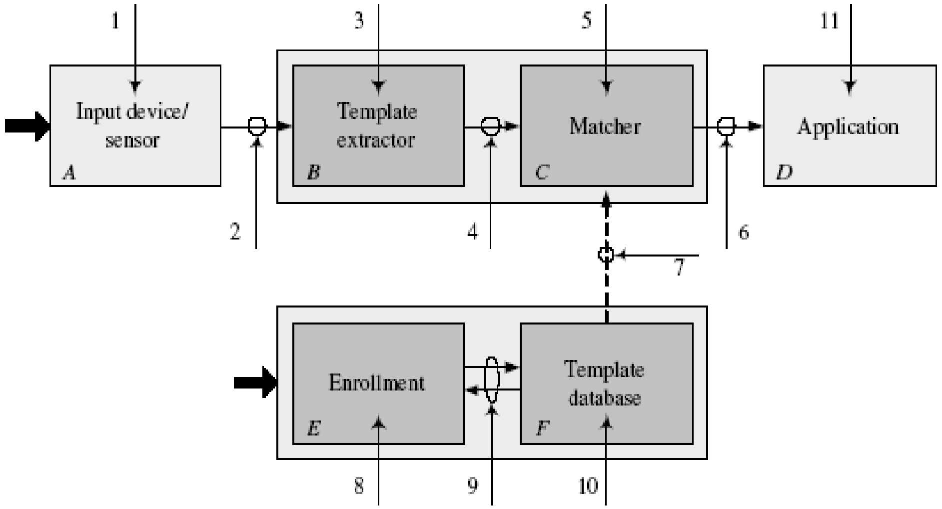
- Assess the impact of adversarial attacks on deep learning based face recognition algorithms
- Create methods to detect and mitigate the effect of such attacks

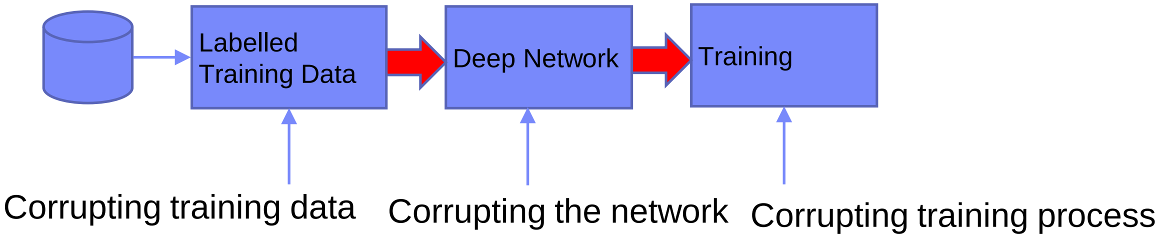
Robustness of Models

- Generalization and Robustness are important for DL
- Sensitivity towards “distribution drift” is a research challenge
- DL models have some singularities and limitations
- These can be exploited by an adversary to “fool” a DL system

Shallow Learning Attack Model (Pre-DL Era)

Deep Learning Attack Models (DL Era)



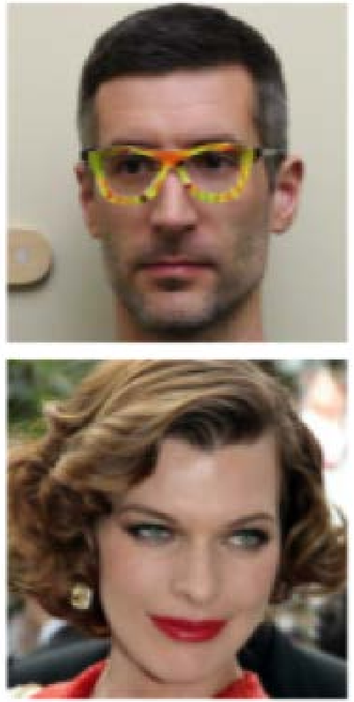


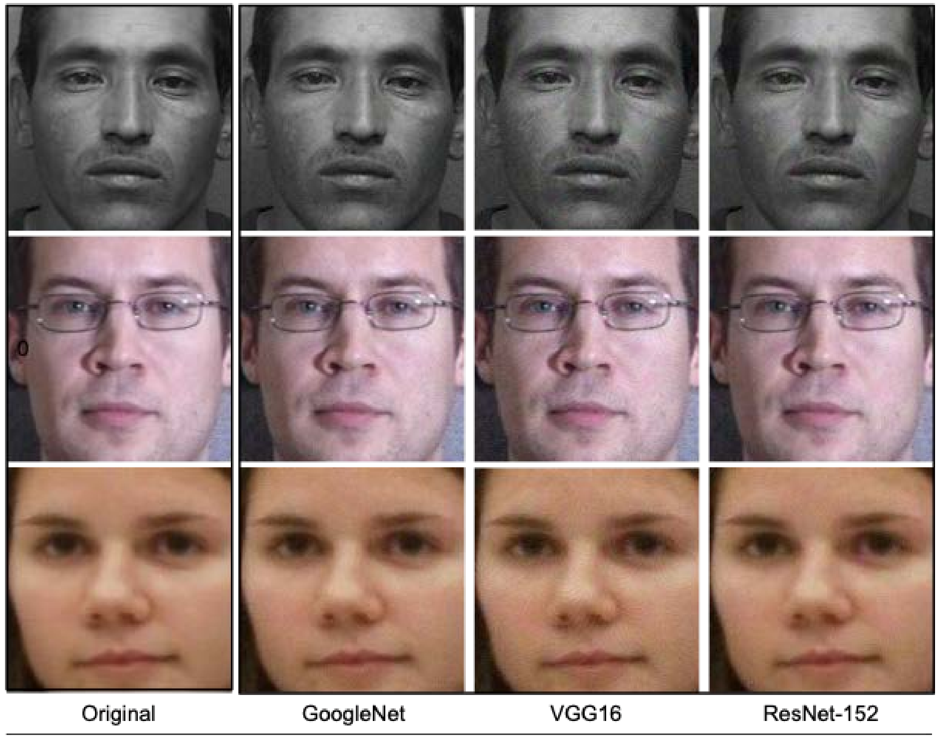
Formidable adversaries:

- Thieves
- Hackers
- Users
- Customers
- Employees
- Merchants
- Competitors
- Competitors' governments

Ratha et al. 2003

Digital Adversarial Attacks





CCS, 2016

Universal Attack, CVPR 2017

Who are these celebrities?

Non-existing identities

Distortions

(a) Original

(b) xMSB

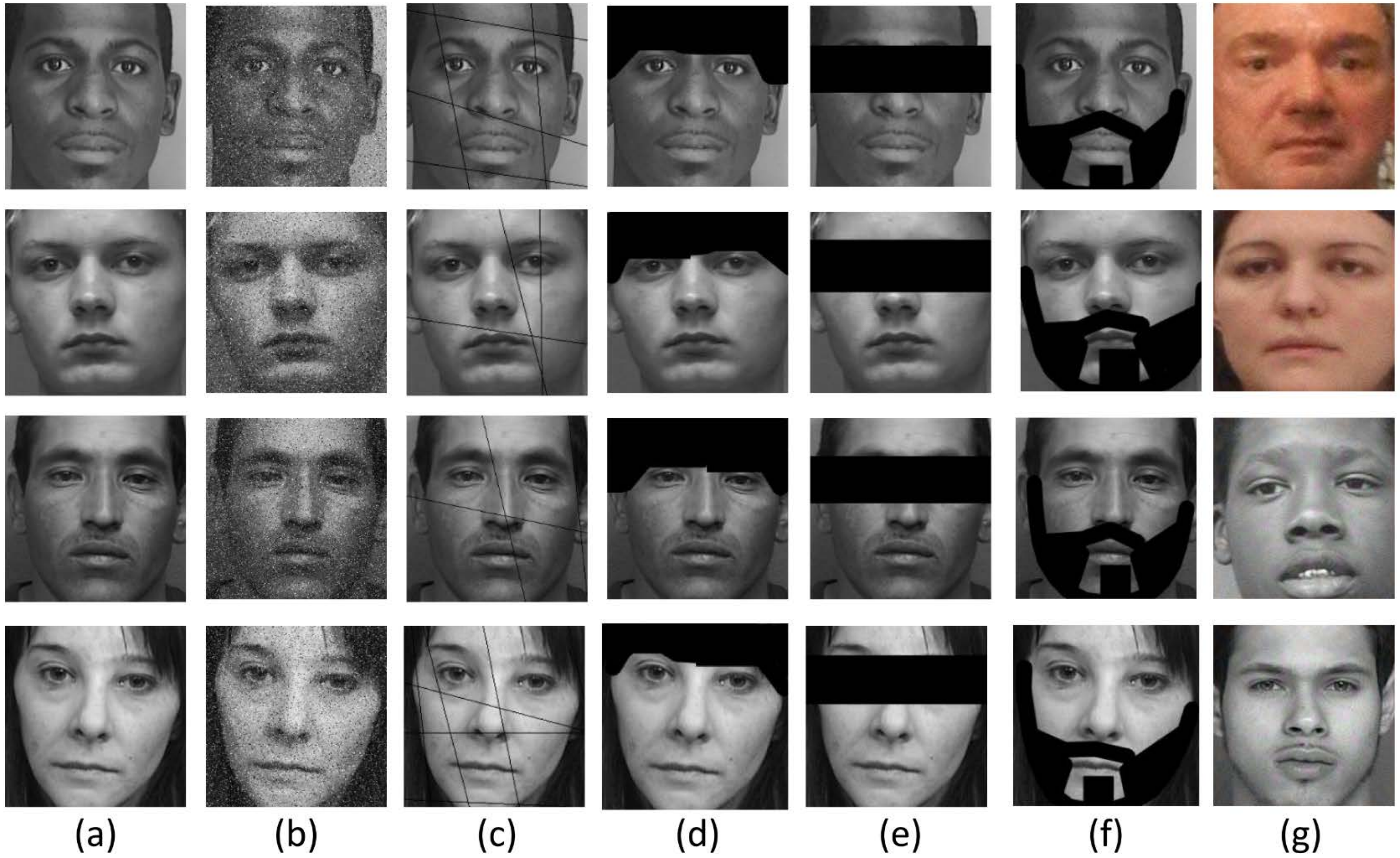
(c) Grids

(d) Forehead and Brow Occlusion (FHBO)

(e) Eye Region Occlusion (ERO)

(f) Beard-like distortion

(g) Universal adversarial perturbation



S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. Universal adversarial perturbations. In IEEE Conference on Computer Vision and Pattern Recognition, 2017.

Deep networks

- OpenFace:
  - Open source implementation of FaceNet with 3,733,968 parameters
  - Trained using FaceScrub and CASIA-WebFace datasets
- VGG:
  - Deep neural network with 11 convolutional layers
  - Trained on 2.6 million face images pertaining to 2,622 subjects
- LightCNN:
  - Deep neural network with 5 convolutional layers
  - Combined database of 99,891 subjects

B. Amos, B. Ludwiczuk, J. Harkes, P. Pillai, K. Elgazzar, and M. Satyanarayanan. OpenFace: Face Recognition with Deep Neural Networks. http://github.com/ cmusatyalab/openface.
O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In British Machine Vision Conference, 2015.
X. Wu, R. He, Z. Sun, and T. Tan. A lightCNN for deep face representation with noisy labels. arXiv preprint arXiv:1511.02683, 2015.

Databases

- PaSC: Still-to-still protocol with 4,688 images belonging to 293 subjects. 2344 x 2344 score matrix
- MEDS: MEDS-II database with 1,309 faces of 518 subjects. 858 x 858 score matrix for all frontal face images.

(a) PaSC

(b) MEDS

J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In IEEE Conference on Biometrics: Theory, Applications and Systems, pages 1–8, 2013.
Multiple Encounters Dataset (MEDS), http://www.nist.gov/itl/iad/ig/sd32.cfm, National Institute of Standards and Technology, 2011.

Results: PaSC database

System	PaSC					
	Original	Grids	xMSB	FBO	ERO	Beard
OpenFace	39.38	10.13	10.13	14.97	6.53	22.56
VGG	31.19	3.14	1.26	15.24	8.79	24.04
COTS	40.32	24.26	19.11	13.02	0	6.15
LightCNN	60.1	24.6	29.5	31.9	24.4	38.1
L-CSSE	61.2	43.1	36.9	29.4	39.1	39.8

A. Majumdar, R. Singh, and M. Vatsa. Face recognition via class sparsity based supervised encoding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6):1273–1280, 2017.
J. R. Beveridge, P. J. Phillips, D. S. Bolme, B. A. Draper, G. H. Given, Y. M. Lui, M. N. Teli, H. Zhang, W. T. Scruggs, K. W. Bowyer, P. J. Flynn, and S. Cheng. The challenge of face recognition from digital point-and-shoot cameras. In IEEE Conference on Biometrics: Theory, Applications and Systems, pages 1–8, 2013.

Results: MEDS database

System	MEDS					
	Original	Grids	xMSB	FHBO	ERO	Beard
COTS	40.3	24.3	19.1	13.0	0	6.2
OpenFace	39.4	10.1	10.1	14.9	6.5	22.6
VGG-Face	54.3	3.2	1.3	15.2	8.8	24.0
LightCNN	60.1	24.6	29.5	31.9	24.4	38.1
L-CSSE	61.2	43.1	36.9	29.4	39.1	39.8

A. Majumdar, R. Singh, and M. Vatsa. Face recognition via class sparsity based supervised encoding. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6):1273–1280, 2017.
Multiple Encounters Dataset (MEDS), http://www.nist.gov/itl/iad/ig/sd32.cfm, National Institute of Standards and Technology, 2011.

Existence of adversaries

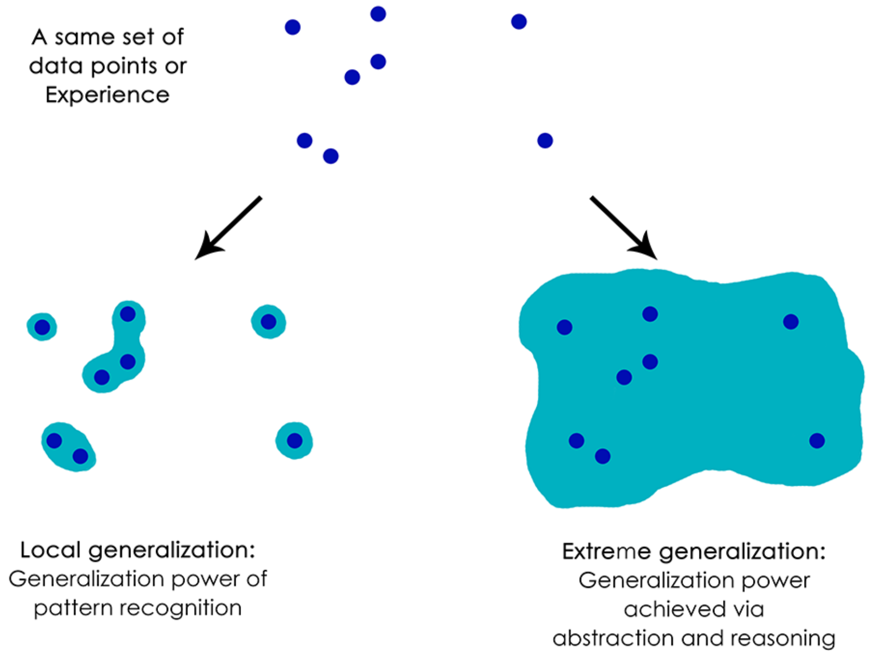
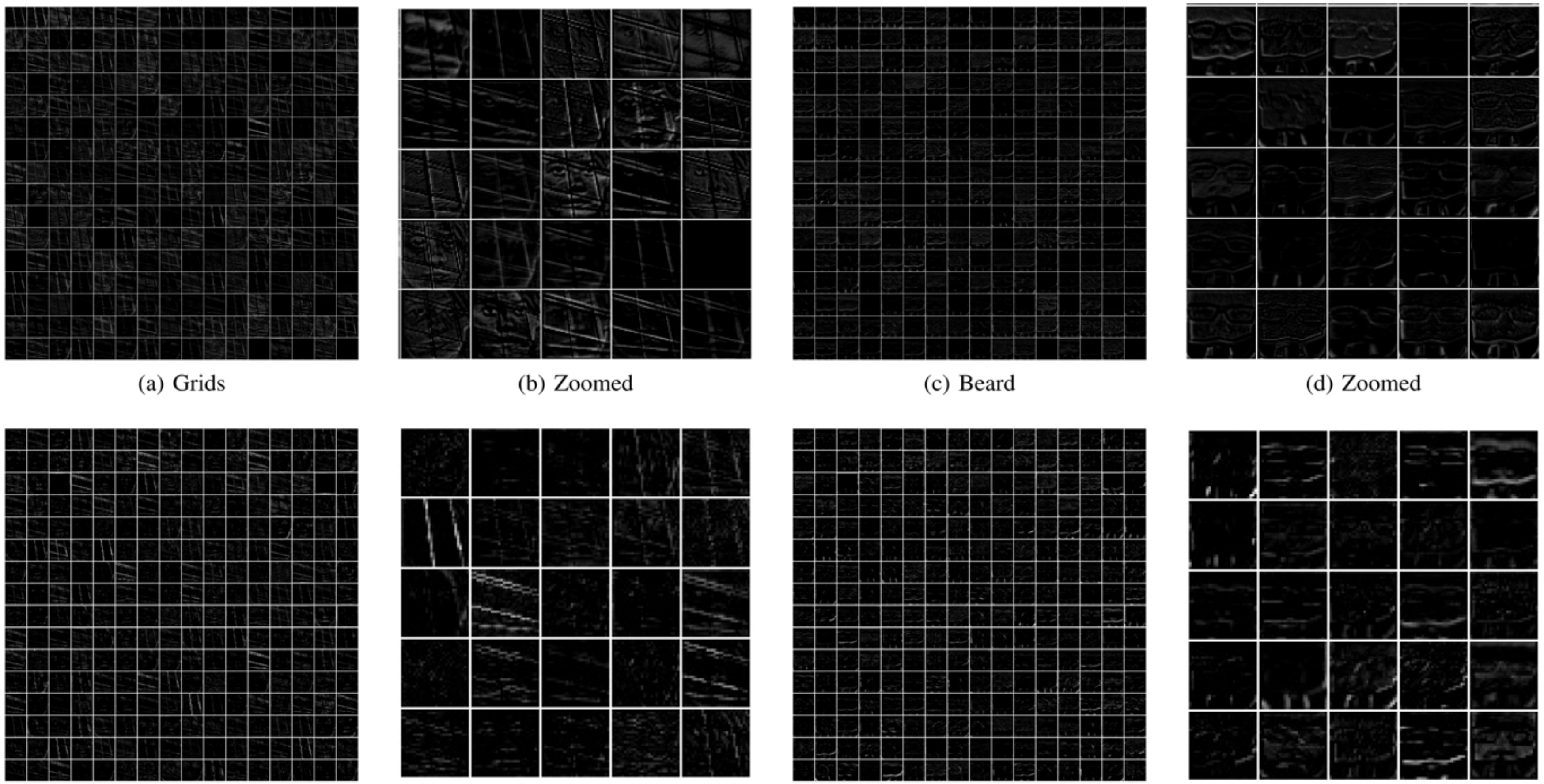


Image courtesy: https://blog.keras.io/the-limitations-of-deep-learning.html

Information captured in intermediate layers

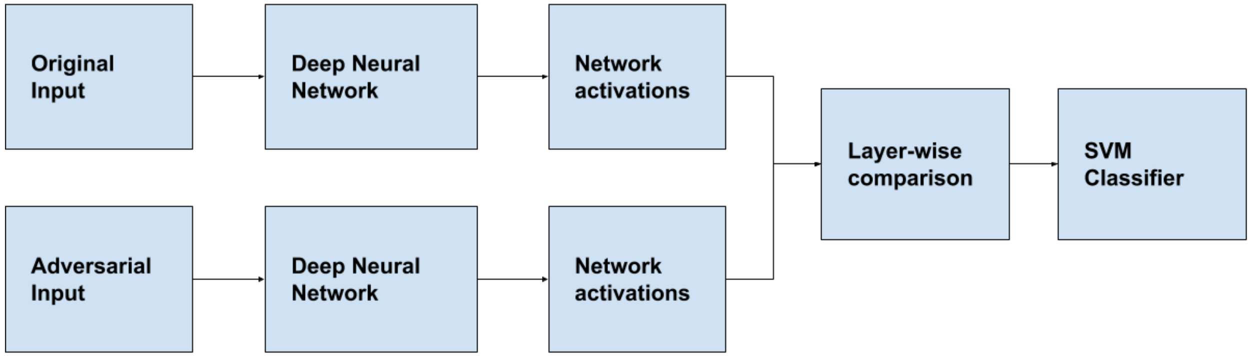


(a) Grids (b) Zoomed (c) Beard (d) Zoomed

(e) Grids (f) Zoomed (g) Beard (h) Zoomed

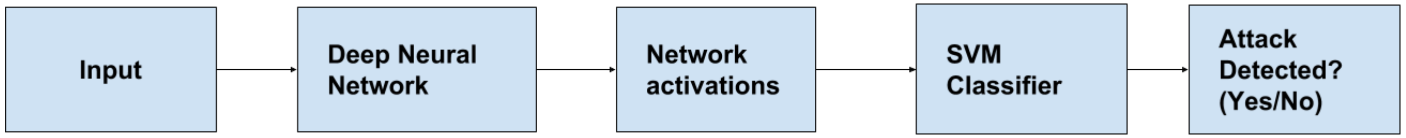
G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks, Thirty-Second AAAI Conference on Artificial Intelligence, 2018

Detecting adversarial attacks: training



- We save the mean output for undistorted images during training: $\mu_i = \frac{1}{N_{train}} \Sigma_{j=1}^{N_{train}} \phi_i\left(I_j\right)$
- Using a distorted set, the detection module learns the Canberra distances of the intermediate activations: $\Psi_i(I, \mu) = \Sigma_z^{\lambda_i} \frac{|\phi_i(I)_z - \mu_{iz}|}{|\phi_i(I)_z| + |\mu_{iz}|}$
- Using these distance metrics as feature vectors, one distance for each layer, a SVM classifier is trained to classify each image as normal/adversarial

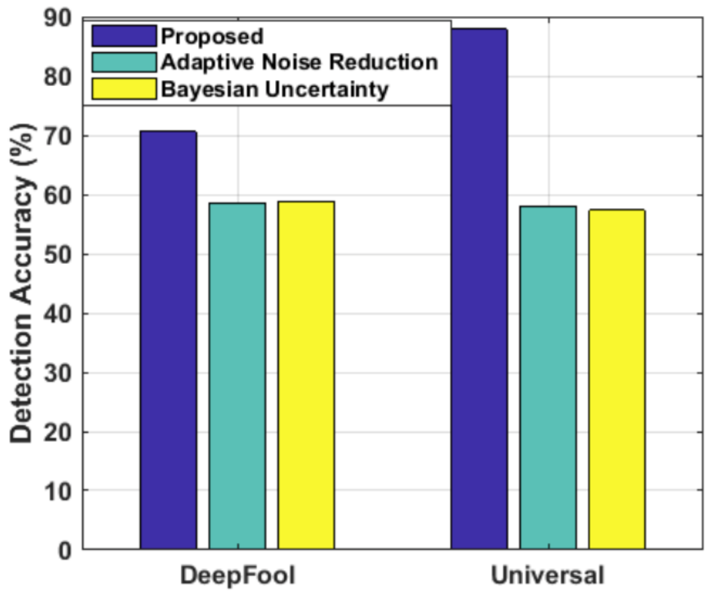
Detecting adversarial attacks: testing

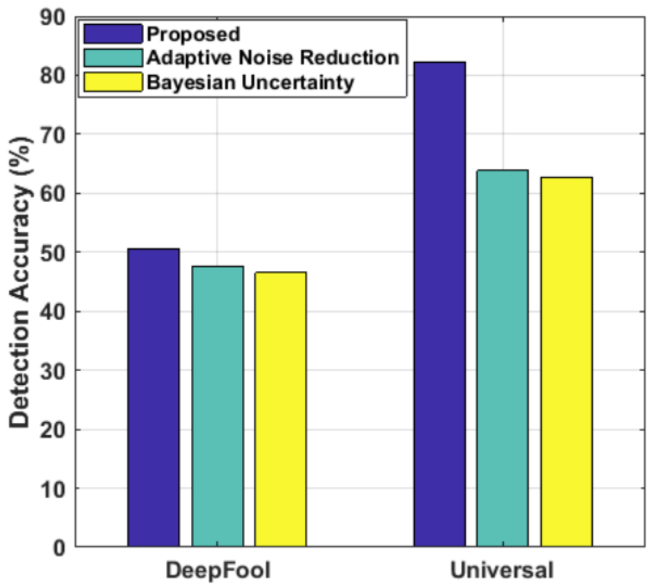


- Each input is characterized by the activations in the intermediate layers of the deep network
- The distances of these activations are computed using the pre-computed mean for the undistorted images during training
- The feature vector obtained using these distances are used to perform two-class classification with the SVM classifier

Detection results: comparison and observations

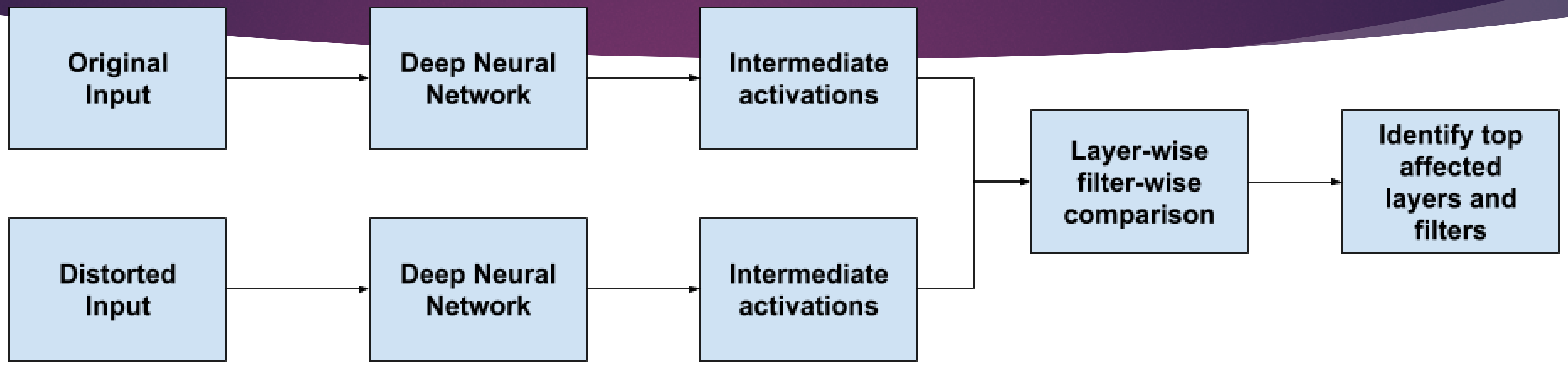
Distortion	MEDS					PaSC				
	Face Quality [23]	BIQI [24]	SSEQ [25]	LightCNN	VGG	Face Quality [23]	BIQI [24]	SSEQ [25]	LightCNN	VGG
Beard	60.0	64.0	43.2	92.2	86.8	56.2	47.4	49.9	89.5	99.8
ERO	61.8	64.3	38.1	91.9	86.0	56.2	48.7	51.2	90.6	99.7
FBO	56.7	63.2	43.9	92.9	84.4	53.5	52.5	51.4	81.7	99.8
Grids	60.7	63.7	44.4	68.4	84.4	55.8	51.1	39.0	89.7	99.9
xMSB	54.3	66.6	40.9	92.9	85.4	55.0	61.0	16.1	93.2	99.8





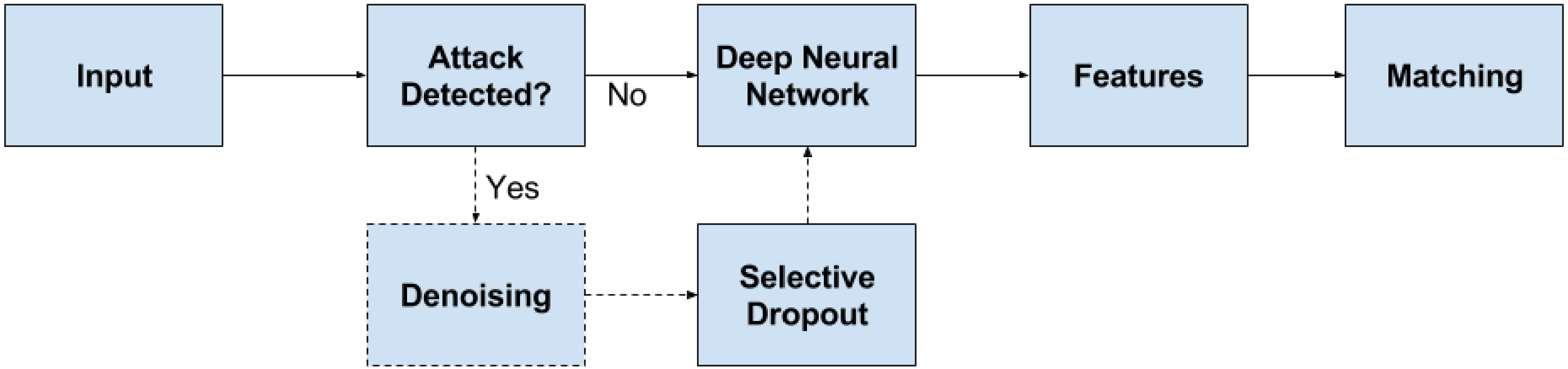
Moorthy et al. 2010, Liu et al. 2014, Parkhi et al. 2015, Wu et al. 2015, Dezfooli et al. 2015, Chen et al. 2015, Dezfooli et al. 2017, Goswami et al. 2017

Mitigating adversarial attacks: training



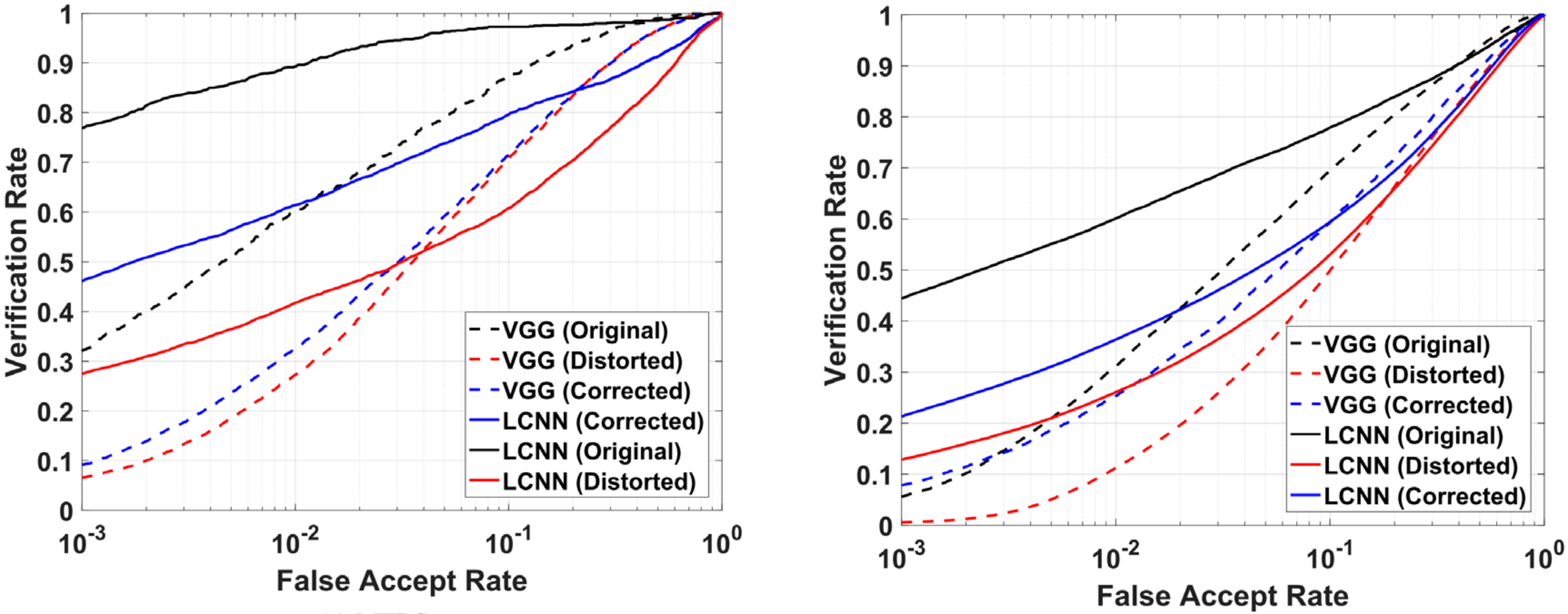
- The mitigation module learns layer-wise filter-wise scores: $\epsilon_{ij} = \Sigma_{k=1}^{N_{dis}} \|\phi_{ij}(I_k) - \phi_{ij}(I_k^{'})\|$
- ϵ_{ij} denotes the score for the j^{th} filter in the i^{th} layer
- These results are stored for the network and used at runtime to perform selective dropout of the most affected K filters from the top N layers.
- N and K are learned using a grid search based optimization

Mitigating adversarial attacks: testing



- Weights of the top most affected N layers and K filters are set to 0 to limit the propagation of adversarial perturbations through the network
- Optionally, apply domain/sample specific noise removal before performing selective dropout to further improve results

Mitigation results



(a) MEDS

(b) PaSC

G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks, Thirty-Second AAAI Conference on Artificial Intelligence, 2018

Detecting and mitigating adversarial attacks: outcomes

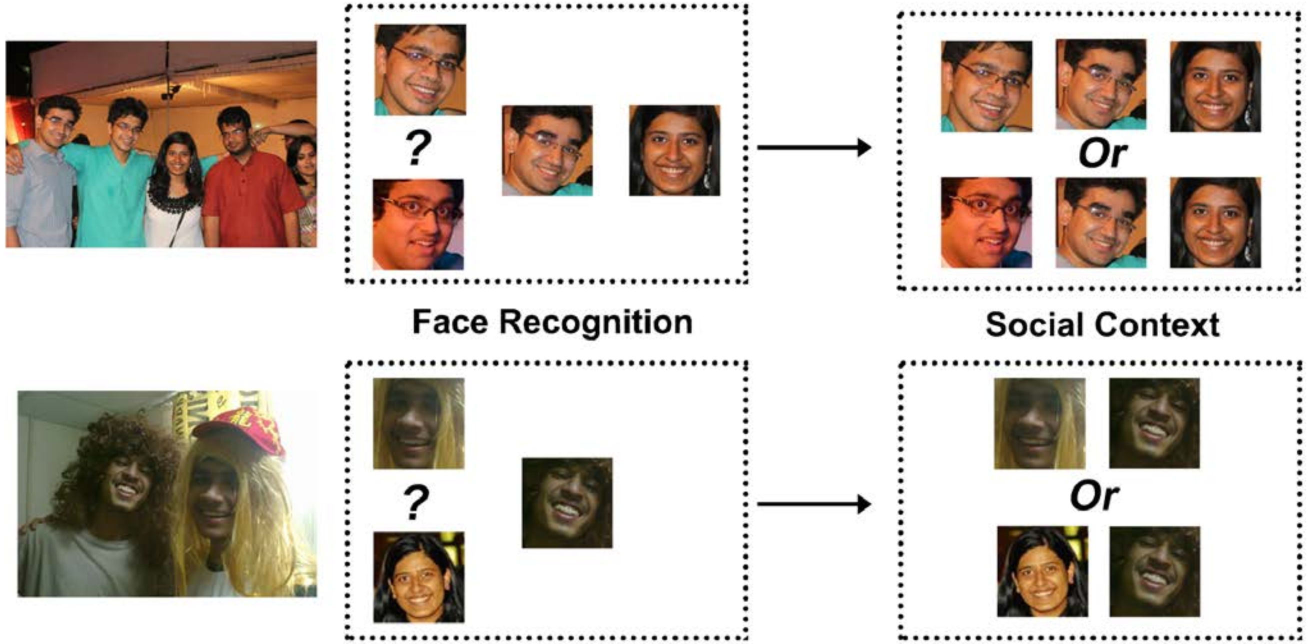
- US Patents
  - 2 US patents filed
- Conference Article
  - G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, Unravelling Robustness of Deep Learning based Face Recognition Against Adversarial Attacks, Thirty-Second AAAI Conference on Artificial Intelligence, 2018. **(Oral Presentation)**
- Journal Article
  - G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa, Detecting and Mitigating Adversarial Perturbations for Robust Face Recognition, International Journal of Computer Vision. *(Submitted after revision)*

Concluding remarks

2012-14	• First RGB-D face recognition algorithm using handcrafted features
2014-16	• Group Sparse Representation based Classifier
2014-17	• First algorithm to break the 90% verification rate barrier on the PaSC database
2016-18	• First study of adversarial attacks on face recognition • First attack mitigation algorithm

Concluding remarks

Input → Feature Extraction Methodology → Matching Representations ⇢ Augmentation ⇢ Decision



R. Bhardwaj, G. Goswami, R. Singh and M. Vatsa, Harnessing Social Context for Improved Face Recognition, IAPR International Conference on Biometrics, 2015.

Accomplishments

- 2 US patents filed in 2018
- 10 journal papers including 2 IEEE TIFS, 3 Information Fusion, 1 PR, 1 PloS ONE
- 10 conference articles including AAAI, ICPR, IJCB, BTAS
- 3 book chapters
- Recipient of the IBM Ph.D. fellowship
- One semester at the IBM TJ Watson Research Center, NY with Dr. Nalini Ratha
- Recipient of the best poster award at BTAS 2013 for “On RGB-D face recognition using Kinect”
- Recipient of the IJCB 2014 Best Doctoral Consortium presentation award and the IDRBT Doctoral Colloquium award

THANK YOU

QUESTIONS ARE WELCOME

BACKUP SLIDES SECTION

Face as a biometric

- Advantages:
  - Does not require cooperation from the subject
  - Does not require specialized capture process and/or equipment
  - Only biometric available in surveillance scenarios
  - Sketch recognition
- Disadvantages:
  - Affected by many covariates
  - Changes with time and age
  - High inter-identity similarity

Progression of face recognition: 2007-2010

- 2007: Introduction of the LFW benchmark database
- Proposal of new hand-crafted features
- Fusion of different hand-crafted features
- Metric learning to combine hand-crafted features
- Focus on 2D still image based face recognition

Year	Best algorithm (LFW)
2008	Ensemble of LBP, Gabor, TPLBP, and FPLBP features
2009	Information Theoretic Metric Learning + LBP, SIFT, TPBLP, and FPLBP
2010	Cosine Similarity Metric Learning + LBP, Gabor, and intensity

Progression of face recognition: 2011-2013

- Introduction of YouTube Faces video database
- Shift of focus to varying forms of face recognition
- Continuation of using hand-crafted feature ensembles and metric learning methodologies

Year	Best algorithm (LFW)	Best algorithm (YTF)
2011	Large scale feature-search with neuro-morphic feature representations	Matched Background Similarity + LBP, CSLBP, and FPLBP
2012	Distance Metric Learning with Eigen-value Optimization + SIFT	
2013	SIFT + Fisher Vectors + Joint-Metric Similarity Learning	Sparse Coding + Whitened PCA + Pair-wise constrained Multiple Metric Learning

Progression of face recognition: 2014-present

- Shift of focus to deep learning and data-driven learning
- Introduction of large scale face databases
- Consideration of robustness for systems with very high performance in a database constrained environment

Year	Best algorithm (LFW)	Best algorithm (YTF)
2014	Deep Convolutional Neural Networks (DeepFace, DeepID)	
2015	FaceNet: Deep Convolutional Neural Network	
2016	LBPNet: Local Binary Pattern Network (LBP + CNN)	Discriminative 3D Morphable Models with a very Deep Convolutional Neural Network
2017-2018	Probabilistic Elastic Part Model + LBP and SIFT	Feature-richness based frame selection with SDAE + DBM based joint feature representation

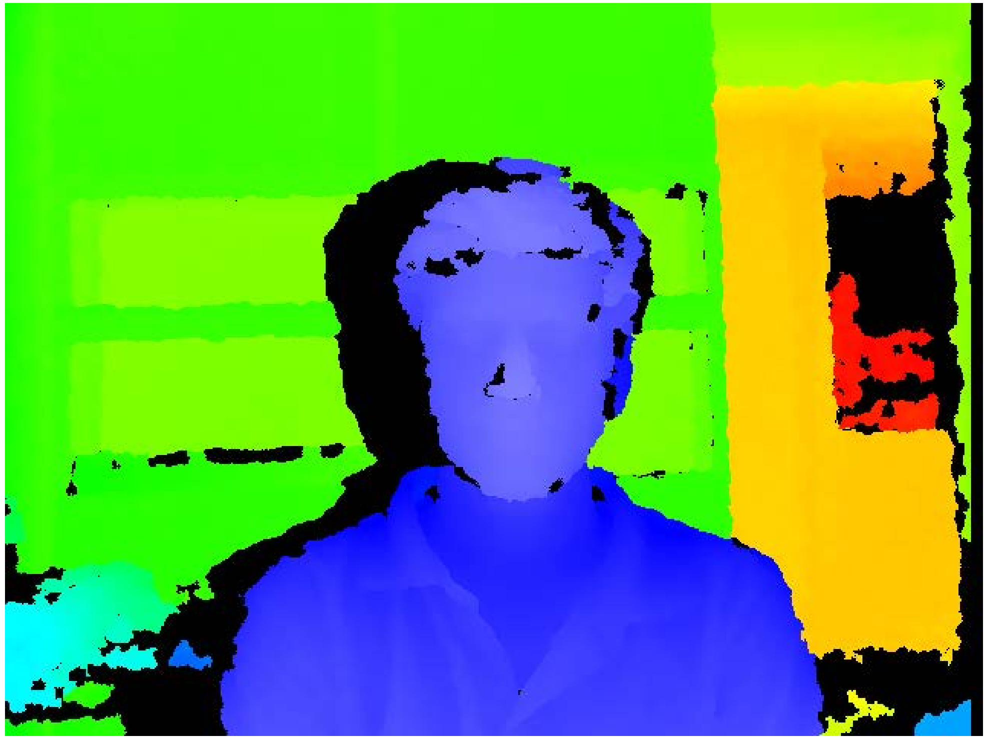
What is Kinect?

KINECT for XBOX 360.

- Originally designed as a motion sensing device for use with the Xbox 360 gaming console.
- Provides RGB image, depth map, IR image, and voice (For images: 640x480 resolution).
- Low cost sensor.

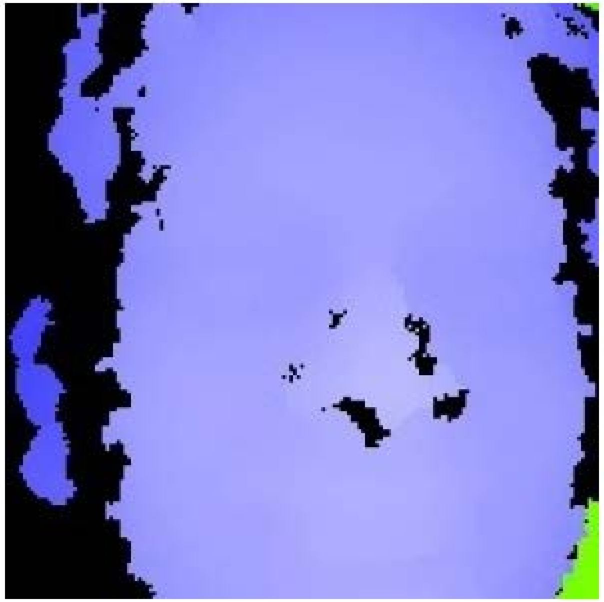
Example RGB-D image obtained using Kinect

RGB Image



Depth Map

Comparison of depth data from different sources


Kinect


Minolta 3D Scanner

How to make use of depth information?

- Traditional: Fit a 3D model based on depth data
- Proposed: Extract features from depth data and combine with visible spectrum features
- Useful for maintaining invariance to expression and illumination.
- High intra-class similarity, can be used to provide stability to the feature descriptor.
- Depth map returned by Kinect is somewhat noisy, with holes. Use in addition with RGB data.
- Information needs to be enhanced before using in feature description : use Entropy map
- Extract geometric attributes

Face recognition algorithm pipeline

WHAT	HOW	WHY
Preprocessing	Interpolation and resizing	Holes and spikes present in depth map
Feature Extraction	Entropy, Saliency, and HOG	Explained in next slide
Matching	Chi-square distance	Ideal for matching histograms
Decision	Score level fusion	To combine different feature sources

Preprocessing

- Face detection using Viola Jones detector in visible spectrum
- Resize to 100 X 100
- Divide image in blocks of 25 X 25
- If a pixel is a hole/spike, rectify using linear interpolation from 3X3 neighborhood

Visual entropy

- Characterizes the variance in pixel intensities in a neighborhood
- Entropy (H) of an image neighborhood **x:**

$$H(\mathbf{x}) = -\sum_{i=1}^{n} p(x_i) log_2 p(x_i)$$

- Encodes the uniqueness of the image at a local level

Visual saliency

- Models visual attention
- In terms of features, it models the feature activations of the image as occurs in the visual cortex of mammals.
- Computed using Itti Koch’s method: Center-surround differences, color, intensity, and orientation features
- Computed only for the visible spectrum image (RGB)
- Provides intra-class stability

Histogram of oriented gradients

- Computes the gradient of the image and creates the gradient orientation histogram
- Popular feature descriptor used in object recognition
- Features: Robust to illumination, succinct representation, controllable granularity
- Used to extract a matcher-friendly representation of the different feature maps

Feature extraction components

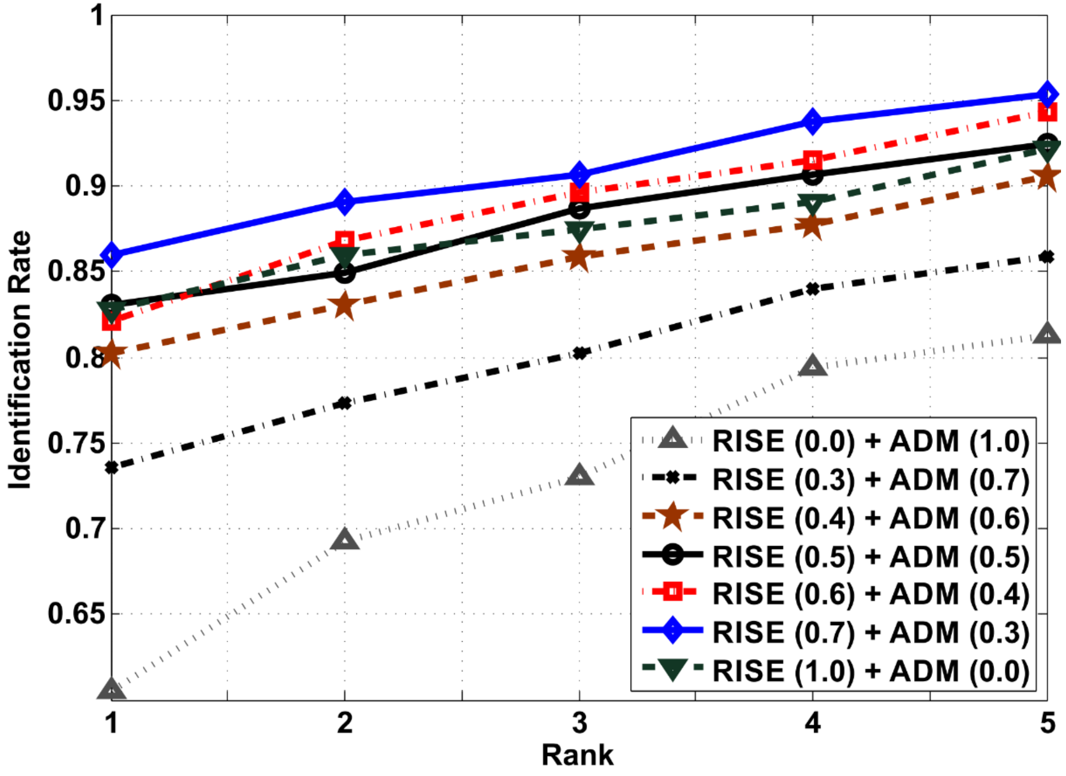
WHAT	WHY
Visual entropy	No feature descriptors exist for depth information, entropy encodes the facial depth variations and texture for RGB image
Visual saliency	Additional feature source which is in accordance to human visual system and provides discriminative information
HOG (Histogram of Oriented Gradients)	Feature histograms are more robust during matching compared to feature maps.

Matching two face images using ADM

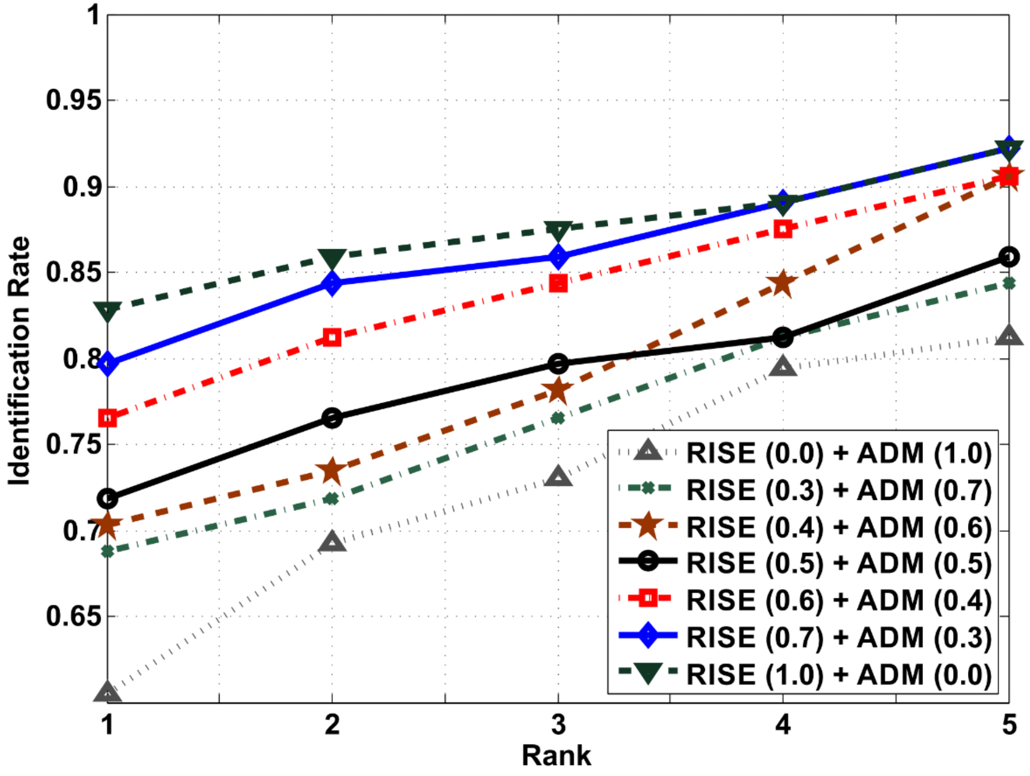
$$\Phi = \sum_{i=1}^{N} w_i \times (A_i - a_i)^2$$

- A_i = attributes of gallery image
- a_i = attributes of probe image
- w_i = weight of i^{th} attribute (optimized using parameter sweep)
- N = total no. of attributes
- φ = ADM match score between gallery and probe image

Combining RISE and ADM



Match score level fusion

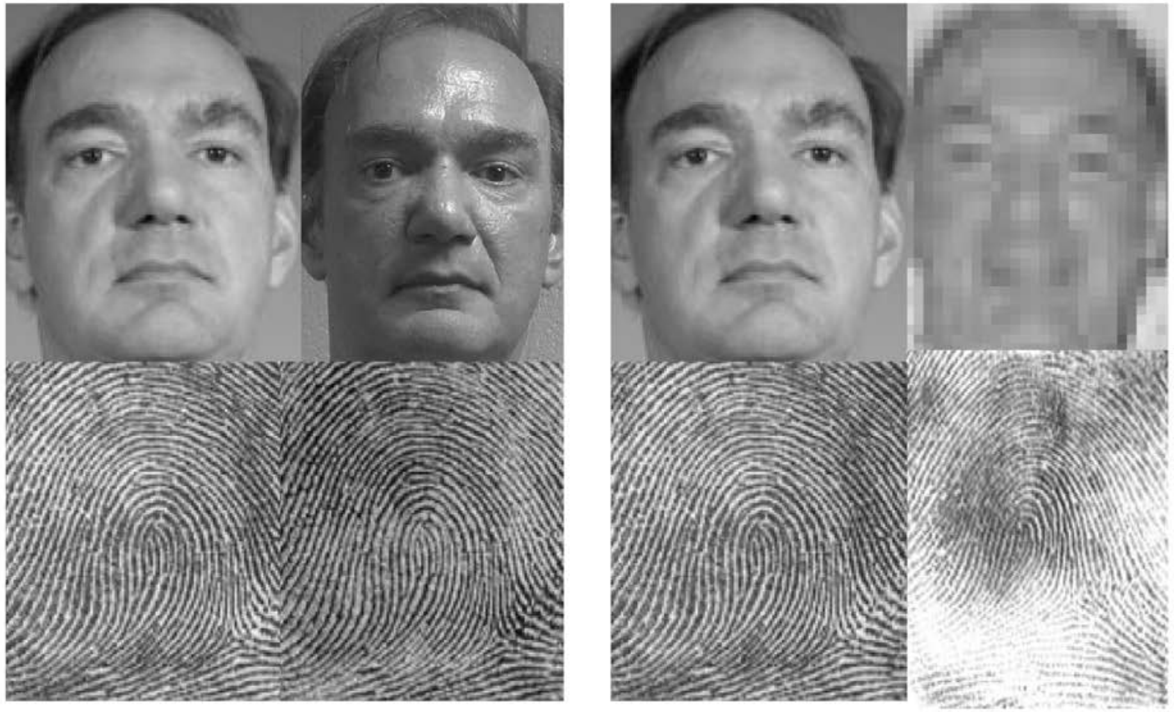


Rank level fusion

Why do we need multi-modal biometrics?

- Face:
  - Easiest to capture
  - Many challenging covariates
- Iris
  - More reliable
  - Special device and procedure to capture
- Fingerprint
  - Available from crime scenes
  - Latent fingerprints are highly difficult to match accurately

Why do we need multi-modal biometrics?



(a) (b)

Levels of fusion

- Sensor-level
- Feature-level
- Score-level
- Rank-level
- Decision-level

Sparse Representation based Classification

- Training samples from a class form a linear basis for test samples of the same class.

$$v_{test} = \alpha_{k,1} v_{k,1} + \alpha_{k,2} v_{k,2} + \cdots + \alpha_{k,n} v_{k,n} + \epsilon$$

$$v_{test} = V\alpha + \epsilon$$

$$V = \left[\underbrace{v_{1,1} | \ldots | v_{1,n}}_{v_1} \,|\, \underbrace{v_{2,1} | \ldots | v_{2,n}}_{v_2} \,|\, \ldots \underbrace{v_{c,1} | \ldots | v_{c,n}}_{v_c} \right]$$

Sparse Representation based Classification

$$\min_{\alpha} ||v_{test} - V\alpha||_2^2 + \lambda||\alpha||_1$$

- Solve the minimization problem
- For each class k,
  - Reconstruct a sample for each class
  - Find reconstruction error
- Assign sample to the class with minimum error

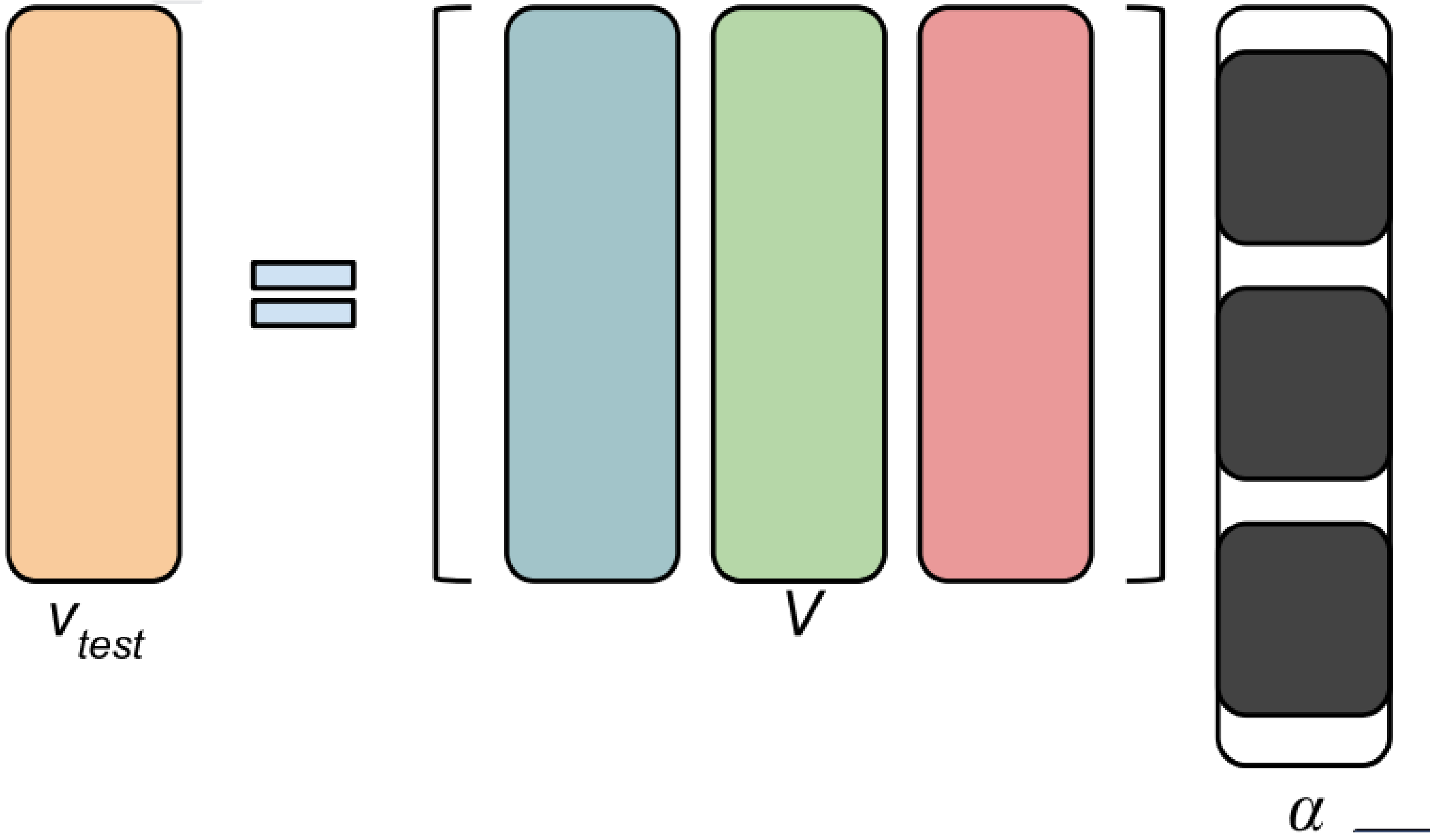
Block/Joint Sparse Classification

$$\min_{\alpha} ||v_{test} - V\alpha|| + \lambda||\alpha||_{2,1}$$

$$||\alpha||_{2,1} = \sum_{k=1}^{n} ||\alpha_k||_2$$

- The inner L2 norm ensures that all members of a particular class are selected, whereas the outer sum acts like a L1 norm and promotes a sparse solution (such that only a few classes are selected).
- Poor performance in face recognition.

The Sparse inverse problem



Proposed Group Sparse Classifier

$$v^i_{test} = V^i \alpha^i + \epsilon \forall i \in \{1 \ldots N\}$$

$$\begin{bmatrix} v^1_{test} \\ \cdots \\ \cdots \\ v^N_{test} \end{bmatrix} = \begin{bmatrix} V^1 & \ldots & 0 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ 0 & \ldots & V^N \end{bmatrix} \begin{bmatrix} \alpha^1 \\ \cdots \\ \cdots \\ \alpha^N \end{bmatrix} + \epsilon$$

$$\min_{\alpha} ||vtest - V\alpha||_2^2 + \lambda ||\alpha||_{2,1}$$

Proposed Group Sparse Classifier

$$v_{test}^{i,j} = V^{i,j}\alpha^{i,j} + \epsilon \forall j \in \{1 \ldots T_i\} \quad and \quad \forall i \in \{1 \ldots N\}$$

$$\begin{bmatrix} v_{test}^1 \\ \cdots \\ \cdots \\ v_{test}^N \end{bmatrix} = \begin{bmatrix} V^1 & \ldots & 0 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ 0 & \ldots & V^N \end{bmatrix} \begin{bmatrix} \alpha^1 \\ \cdots \\ \cdots \\ \alpha^N \end{bmatrix} + \epsilon$$

Deep learning architecture: joint framework

$$\operatorname{argmin}(\| \mathbf{f_1} - \mathbf{f}'_1 \|_2^2 + \| \mathbf{f_2} - \mathbf{f}'_2 \|_2^2 + \mathcal{R})$$

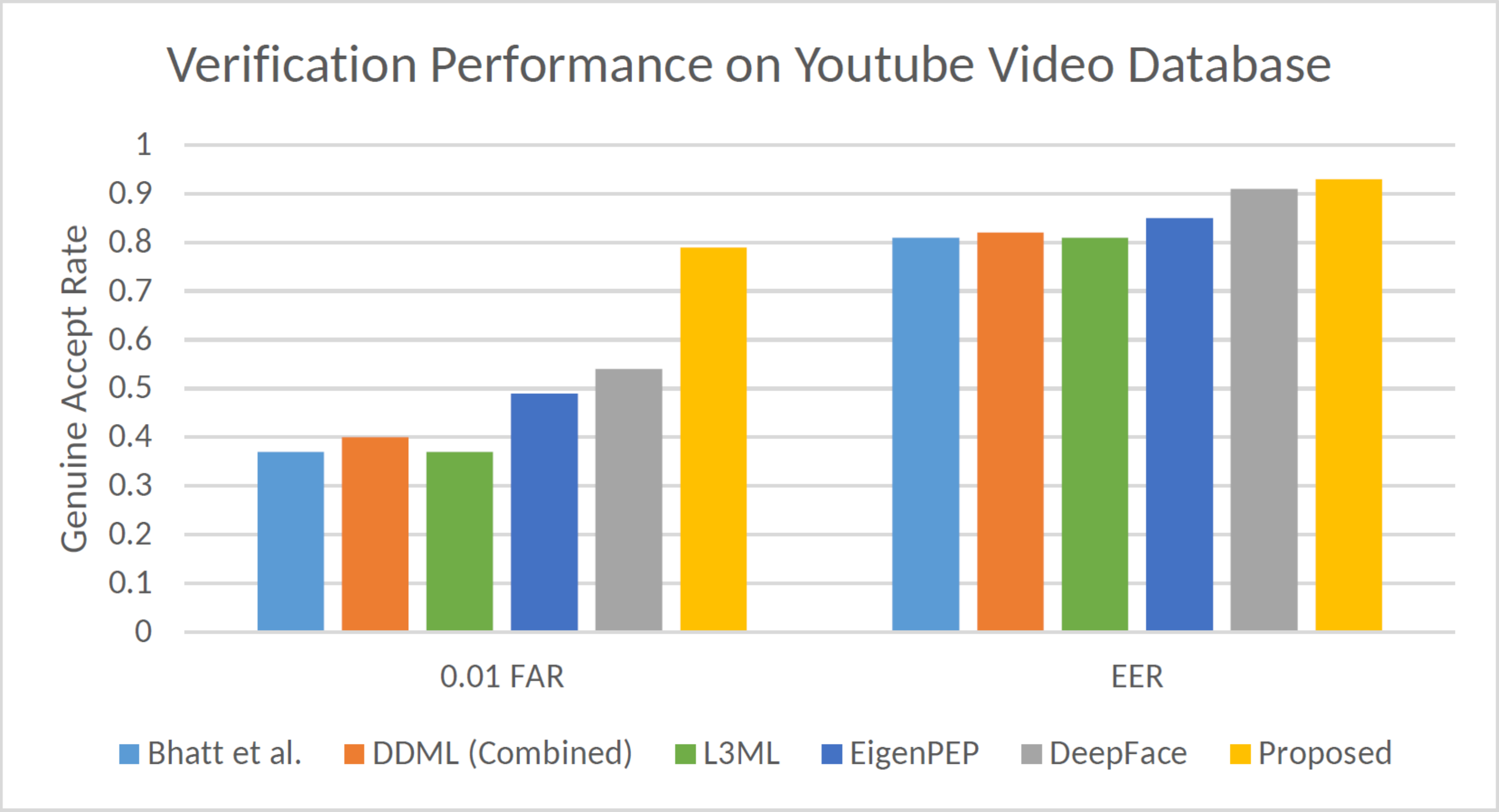
$$\operatorname{argmin}(\| \mathbf{f_1} - s(\mathcal{W}'_\mathbf{1}[s(\mathcal{W}_\mathbf{1}\mathbf{f_1})]) - s(\mathcal{W}'_\mathbf{1}[s(\mathcal{W}_\mathbf{2}\mathbf{f_2})]) \|_2^2 + \\ \| \mathbf{f_2} - s(\mathcal{W}'_\mathbf{2}[s(\mathcal{W}_\mathbf{2}\mathbf{f_2})]) - s(\mathcal{W}'_\mathbf{2}[s(\mathcal{W}_\mathbf{1}\mathbf{f_1})]) \|_2^2 + \mathcal{R})$$

$$\operatorname{argmin}\Big(\| \mathbf{f_1} - s(\mathcal{W}'_\mathbf{1}[s(\mathcal{W}_\mathbf{1}\mathbf{f_1})]) - s(\mathcal{W}'_\mathbf{1}[s(\mathcal{W}_\mathbf{2}\mathbf{f_2})]) \|_2^2 + \\ \| \mathbf{f_2} - s(\mathcal{W}'_\mathbf{2}[s(\mathcal{W}_\mathbf{2}\mathbf{f_2})]) - s(\mathcal{W}'_\mathbf{2}[s(\mathcal{W}_\mathbf{1}\mathbf{f_1})]) \|_2^2 + \\ (\lambda_1 \| \mathcal{W}_\mathbf{1} \|_2^2 + \lambda_2 \| \mathcal{W}_\mathbf{2} \|_2^2)\Big)_{dropout}$$

Deep learning architecture: experimental analysis

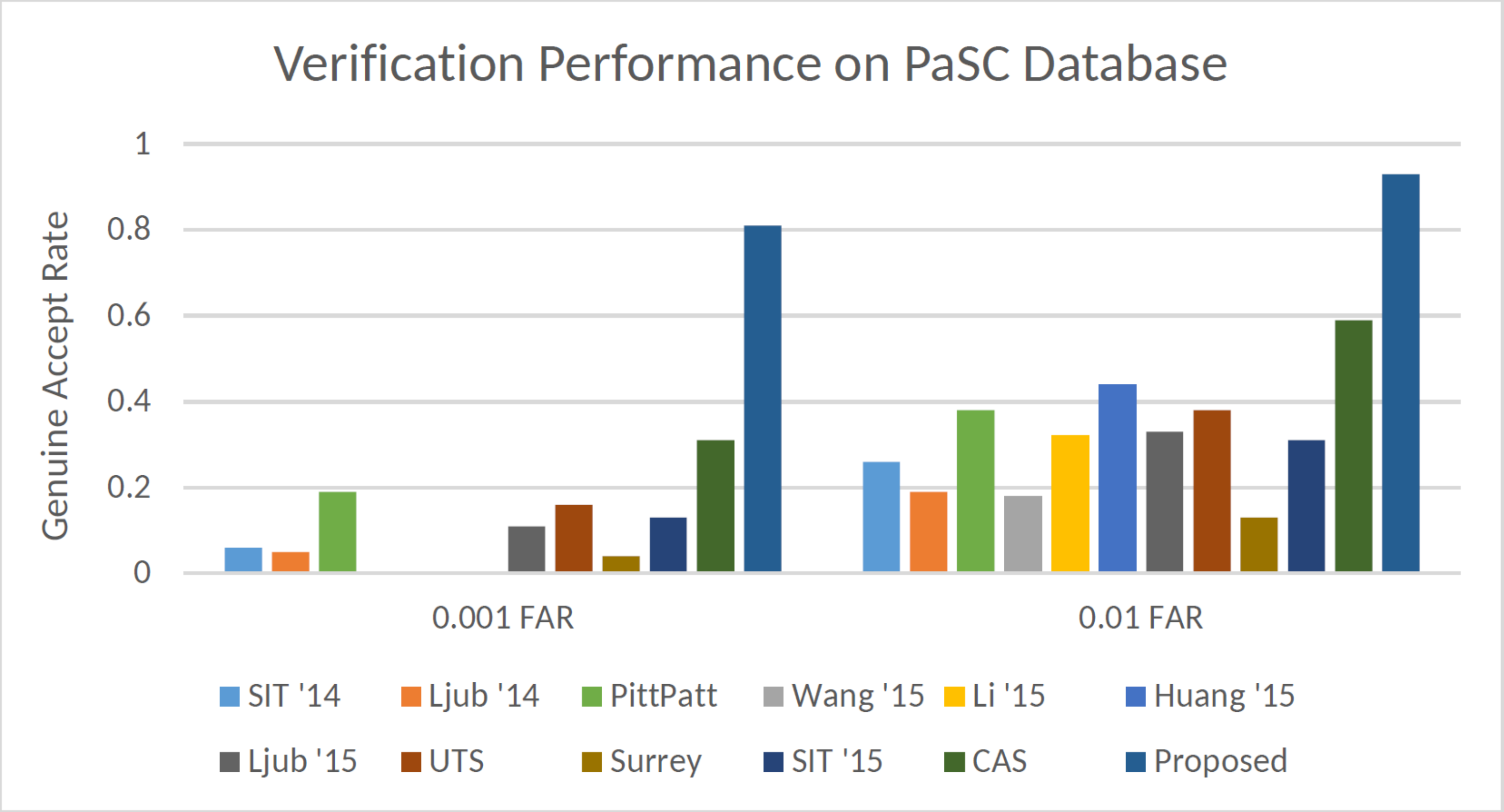
Modified Architecture	GAR at 0.01 FAR		
	YouTube	PaSC	
		Handheld	Control
1 layer Denoising Autoencoder only	0.21	0.09	0.12
2 layer SDAE only	0.39	0.28	0.39
DBM only	0.41	0.48	0.49
SDAE + DBM only	0.61	0.87	0.93
SDAE + DBM with joint representation	**0.79**	**0.93**	**0.96**

Results: YTF



Verification Performance on Youtube Video Database
Genuine Accept Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.01 FAR
EER
Bhatt et al.
DDML (Combined)
L3ML
EigenPEP
DeepFace
Proposed

Results: PaSC

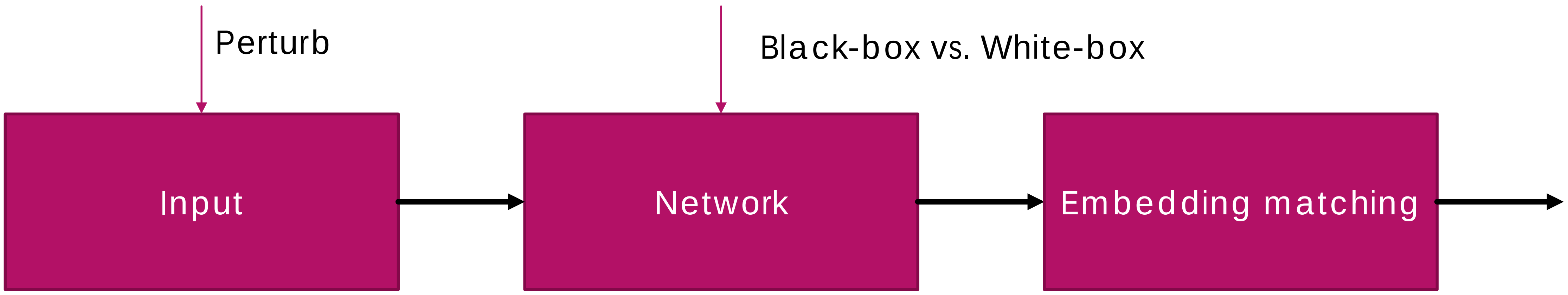


Adversarial attacks on deep learning

- Deep learning based methodologies have showcased state-of-the-art results in a variety of problems: handwritten digit recognition, object recognition, speech recognition, and more
- Despite high performance, deep networks are susceptible to adversarial attacks
- A methodology for addressing adversarial attacks is essential to make deep learning based algorithms robust and accurate in real-world applications

Adversaries for deep learning systems

- Input:
  - Perceptible vs Imperceptible input perturbations
  - Targeted attacks vs. non-targeted attacks
  - Image specific vs. Universal
- Network:
  - Black-box vs white-box



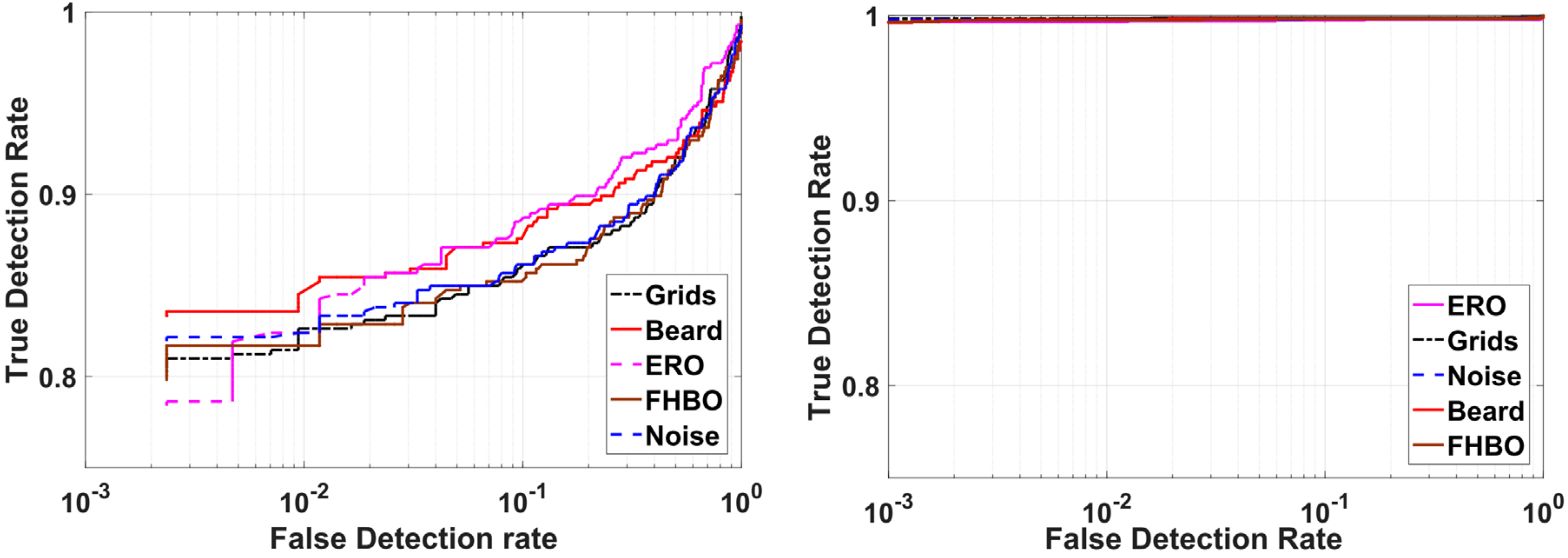
Contributions and highlights

- Existing:
  - Different methods of generating adversarial input examples for attacking deep networks that utilize network architecture information
  - Generative adversarial framework where training is improved using adversarial examples generated during training
    - Dependent on the particular network architecture
    - Requires special training methods and retraining for existing networks
- Proposed:
  - An attack methodology that doesn't need network architecture information
  - Generalized adversarial attack detection and mitigation approaches
    - Independent of the network architecture, plug and play for new networks
    - Requires training for only the detection and mitigation modules
    - Does not require network retraining or fine tuning

Adversarial attacks on face recognition

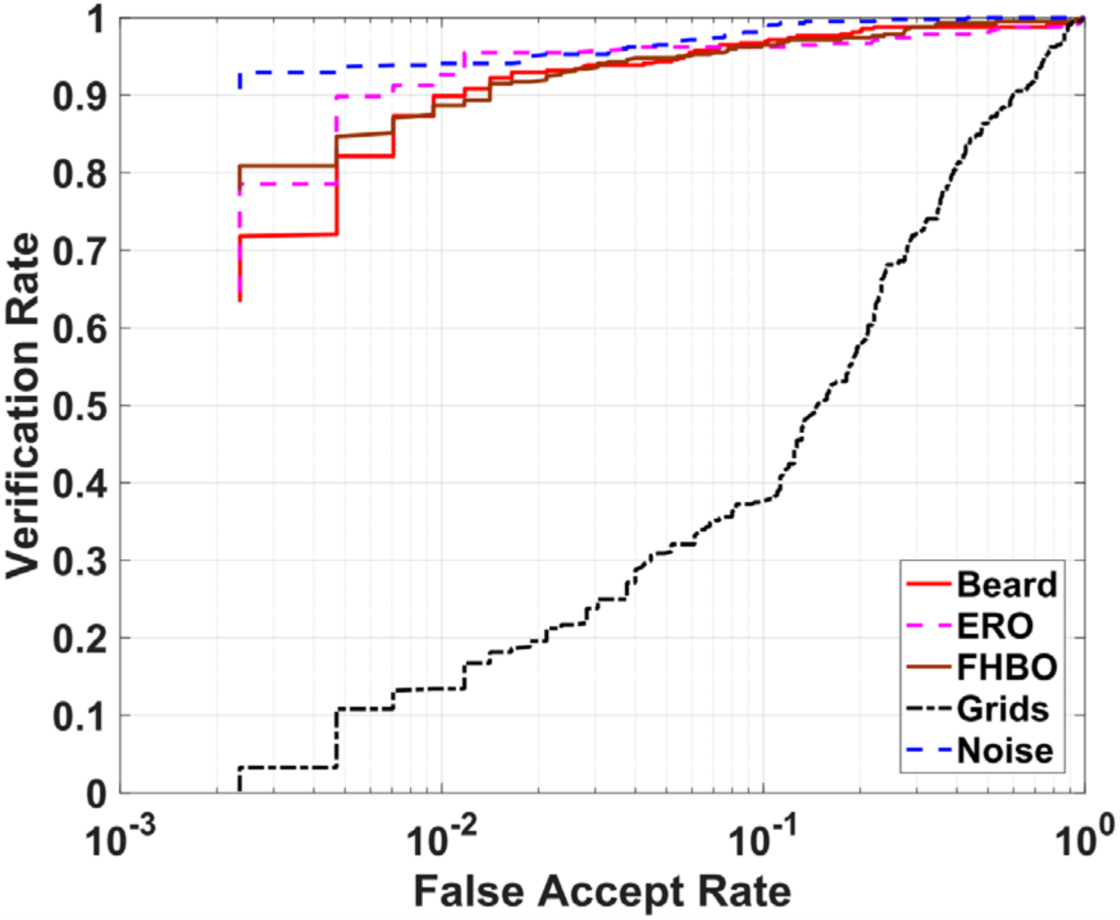
Scores depict real distance measures obtained for the pairs shown as reported by the OpenFace API and the VGG face network

Detection Results: VGG

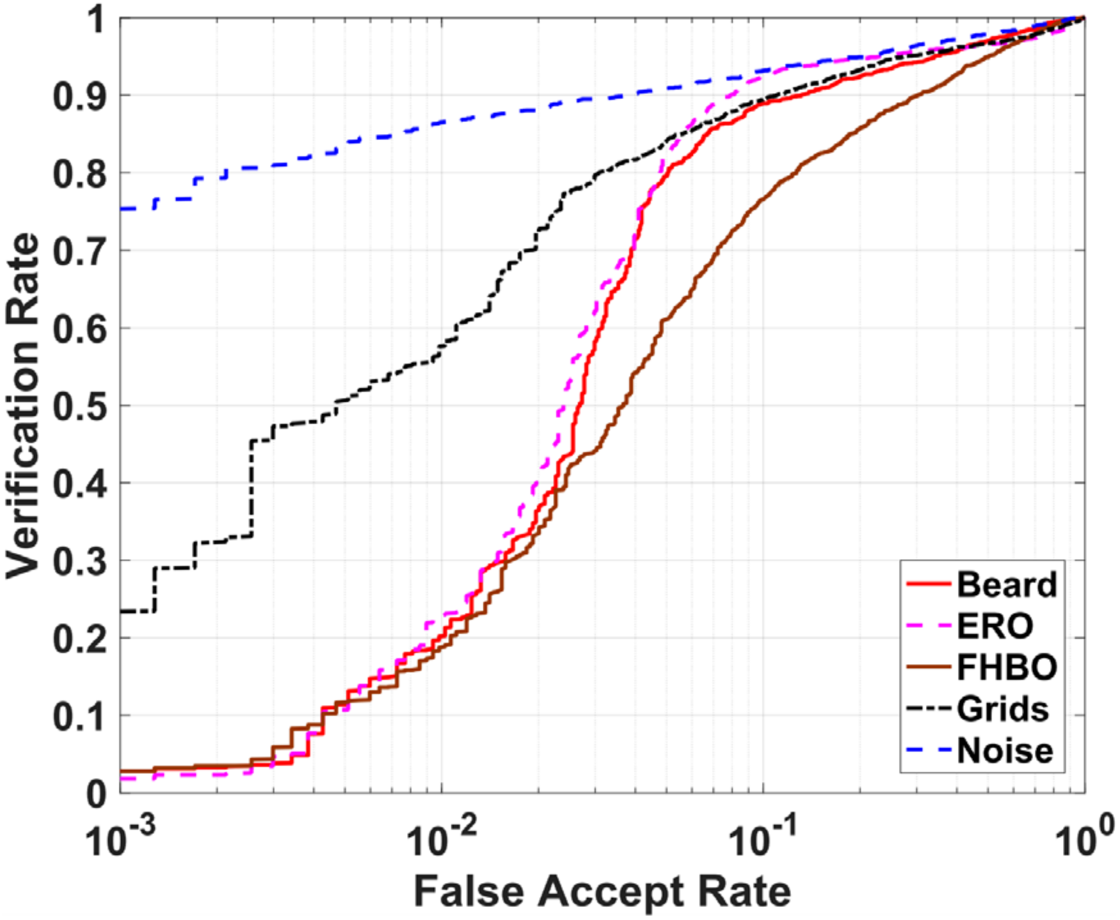


(a) MEDS (b) PaSC

Detection Results: LightCNN



(c) MEDS



(d) PaSC

Detection results: comparison and observations

Distortion	MEDS					PaSC				
	Face Quality [23]	BIQI [24]	SSEQ [25]	LightCNN	VGG	Face Quality [23]	BIQI [24]	SSEQ [25]	LightCNN	VGG
Beard	60.0	64.0	43.2	92.2	86.8	56.2	47.4	49.9	89.5	99.8
ERO	61.8	64.3	38.1	91.9	86.0	56.2	48.7	51.2	90.6	99.7
FBO	56.7	63.2	43.9	92.9	84.4	53.5	52.5	51.4	81.7	99.8
Grids	60.7	63.7	44.4	68.4	84.4	55.8	51.1	39.0	89.7	99.9
xMSB	54.3	66.6	40.9	92.9	85.4	55.0	61.0	16.1	93.2	99.8

- Quality based approaches are unable to perform well, especially for PaSC database which has inherently low quality images
- Texture based methods such as LBP and DSIFT features + SVM classifier yield 25% less accuracy compared to the proposed approach
- 60% of distorted images still pass face detection algorithms
- Lower accuracy with LightCNN might be due to lesser number of layers (and therefore, features)

Observations

- Deep learning based approaches lose two to three times more performance as opposed to non-deep learning based approach, showcasing lack of inherent robustness towards adversarial attacks
- Deep learning based approach appears to be more sensitive to noise in data
- Using intermediate layer outputs to detect attacks is highly accurate as compared to quality/texture based methods or a face detection based test