Cross-Quality Face Recognition with Deep Models and Human Recognition

Na Zhang

Motivation

- To examine the performance of cross quality face recognition
- Compare with human performance of FR on cross-quality faces
- Focus on extremely difficult level of face images (those face images that deep model fails to recognize successfully)

Datasets Preparation

- Two datasets
  - IJB-A: 21,230 images (500 subjects)
  - FaceScrub: 78,650 images (530 subjects)
- Divide each dataset into three groups using same protocol (according to face quality score).
  - High quality set: image quality score >= 60
  - Middle quality set: image quality score in [30, 60)
  - Low quality set: image quality score < 30

- IJB-A
  - High quality set : 1,543 images (500 subjects)
  - Middle quality set : 13,491 images (483 subjects)
  - Low quality set: 6,196 images (489 subjects)
- FaceScrub
  - High quality set: 57,124 images (530 subjects)
  - Middle quality set: 21,164 images (530 subjects)
  - Low quality set: 362 images (232 subjects)

Considering high cost of time and memory of code running, trim FaceScrub dataset:

- Method
  - High quality set: randomly select 1/6 images of each subject
  - Middle quality set: randomly select half of each subject
  - Low quality set: unaltered
- Trimmed Version of FaceScrub
  - High: 10,089 images (530 subjects)
  - Middle: 10,444 images (530 subjects)
  - Low: 362 images (232 subjects)

20,895 images (530 subjects) in total

Method

(1) Deep Model based Face Verification

- Choose low quality sets of each dataset as query images
- Choose high quality sets of each dataset as gallery images
- Perform face verification experiment using four deep models
    - VGGFace
    - LightCNN
    - CenterLoss
    - FaceNet

(2) Human based Face Verification

- Choose the deep model with best performance among the four models in face verification experiments
- Find the best decision boundary for positive and negative pairs based on the selected deep model
- Randomly select those pairs that the selected deep model fails to recognize correctly
- Recruit humans to perform face verification on these selected pairs using a tool

Face Verification on Deep Models

- Perform face verification experiment
    - Low vs. High quality set
    - Middle vs. High quality set

- Calculate Cosine Similarity Score
- Python Programming Language adopted
    - Calculate the Verification Accuracy with respect to
        - FAR=0.01
        - FAR=0.001
        - FAR=0.0001

(FAR: false accept error; TAR: true accept error)

Program Procedures

- Read face features of all probe and gallery images
- Construct Similarity Matrix
    - Rows: probe images
    - Columns: gallery images
    - Values: cosine similarity scores
- Create Similarity Mask Matrix
    - Rows: probe images
    - Columns: gallery images
    - Values: -1 means two images in row and column is positive pair; 127 indicates negative pair
- Calculate accuracy with respect to FAR=0.01, 0.001, 0.0001

IJB-A

High quality set: 1,543 images
Middle quality set: 13,491 images
Low quality set: 6,196 images

□ Low to High Matching

- ✔ Positive pairs: 18,978
- ✔ Negative pairs: 9,541,450

□ Middle to High Matching

- ✔ Positive pairs: 41,642
- ✔ Negative pairs: 20,774,971

□ Low to Middle:

- ✔ Positive pairs:
- ✔ Negative pairs:

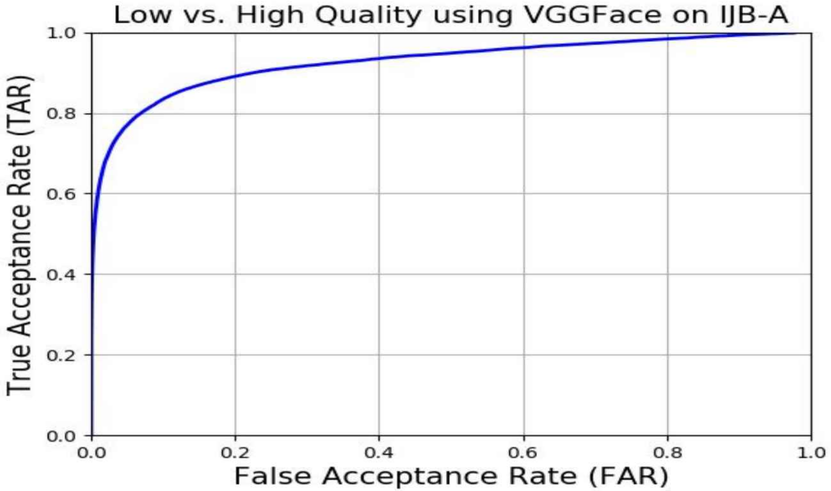
FaceScrub

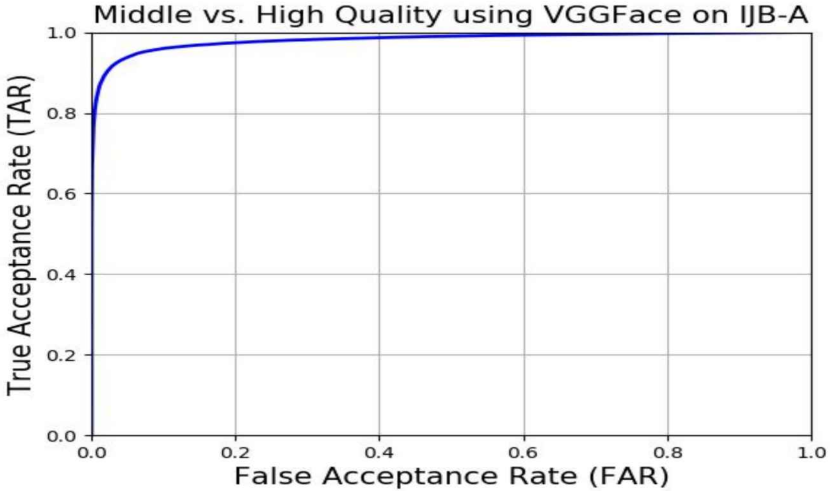
High quality set: 10,089 images
Middle quality set: 10,444 images
Low quality set: 362 images

- □ Low to High Matching
  - ✔ Positive pairs: 6,676
  - ✔ Negative pairs: 3,645,542
- □ Middle to High Matching
  - ✔ Positive pairs: 193,745
  - ✔ Negative pairs: 105,175,771
- □ Low to Middle:
  - ✔ Positive pairs:
  - ✔ Negative pairs:

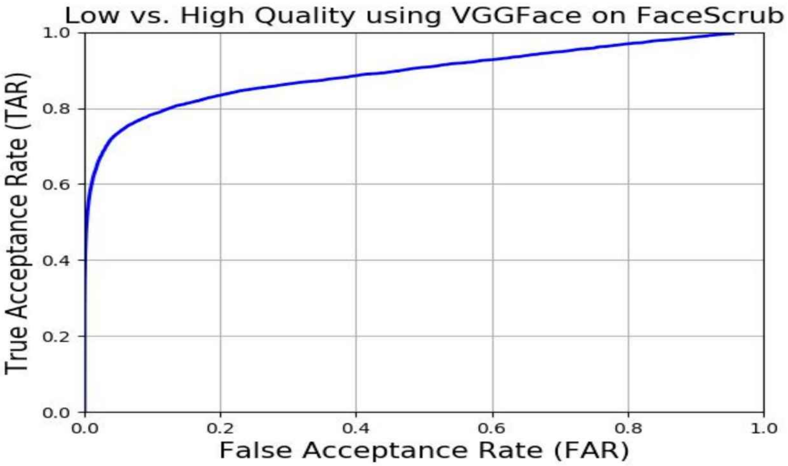
Deep Feature Matching:

- VGGFace on IJB-A:

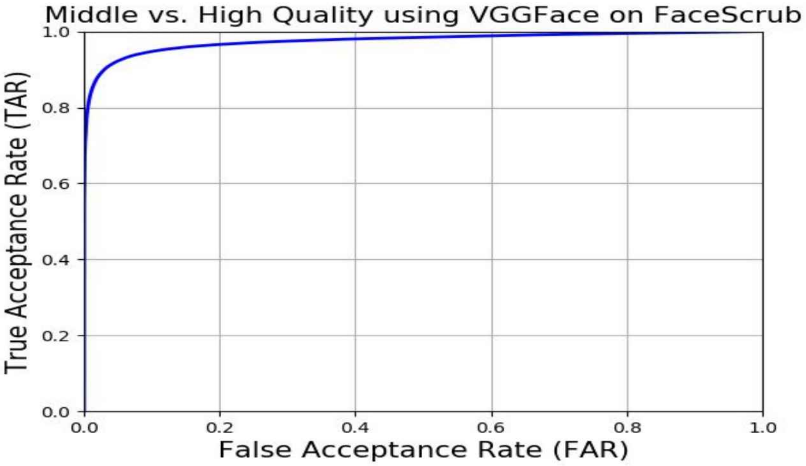




- VGGFace on FaceScrub:

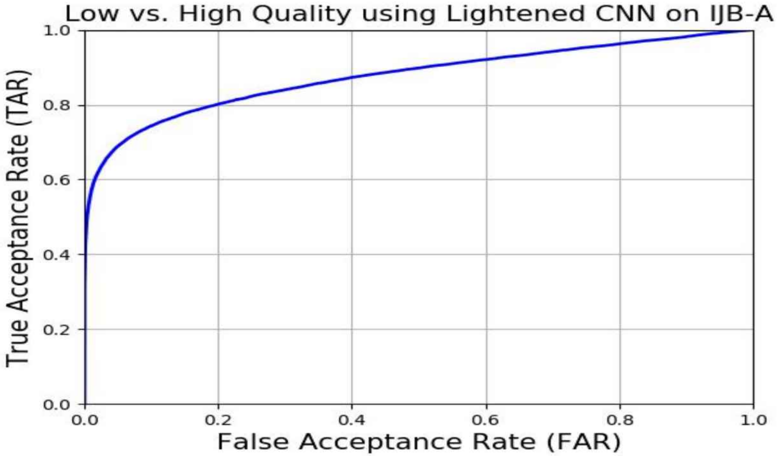


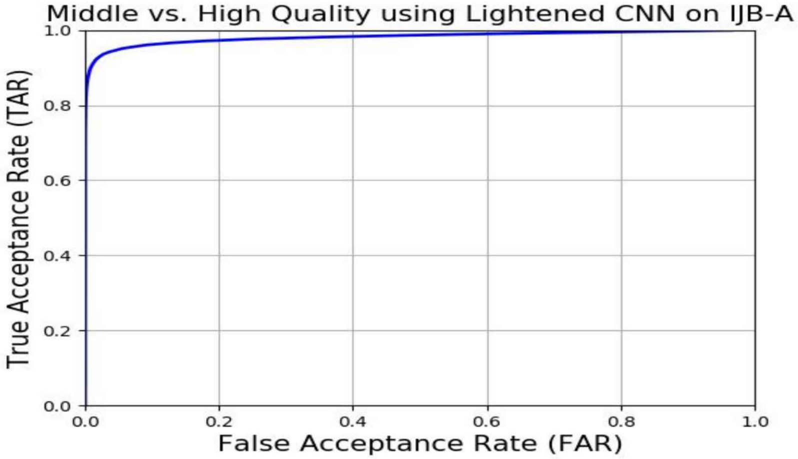
Low vs. High Quality using VGGFace on FaceScrub



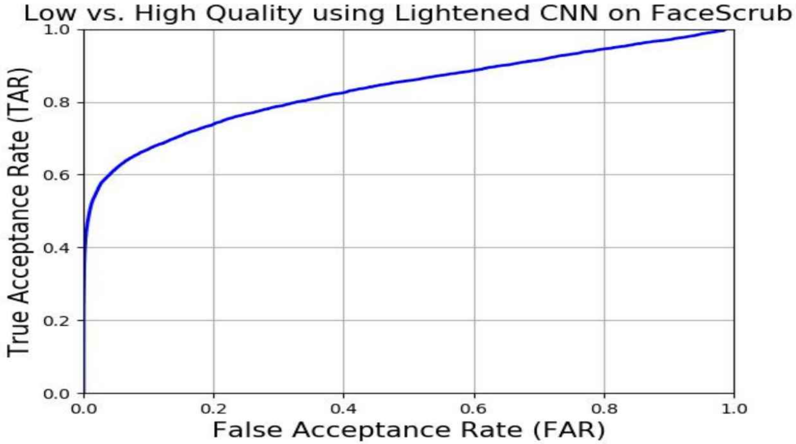
Middle vs. High Quality using VGGFace on FaceScrub

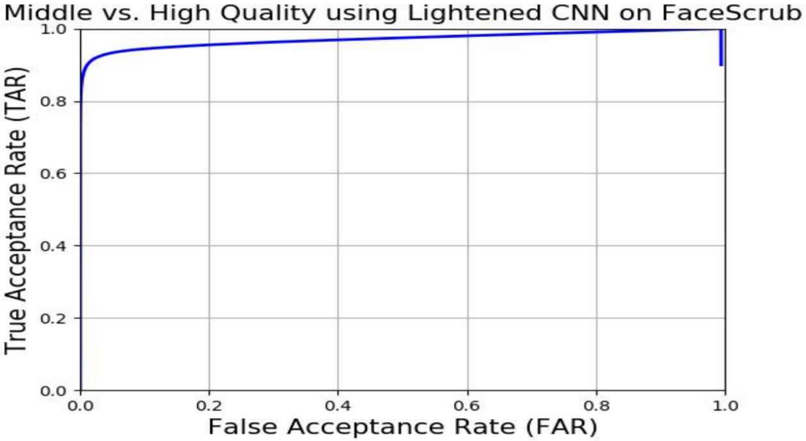
- LightCNN on IJB-A:



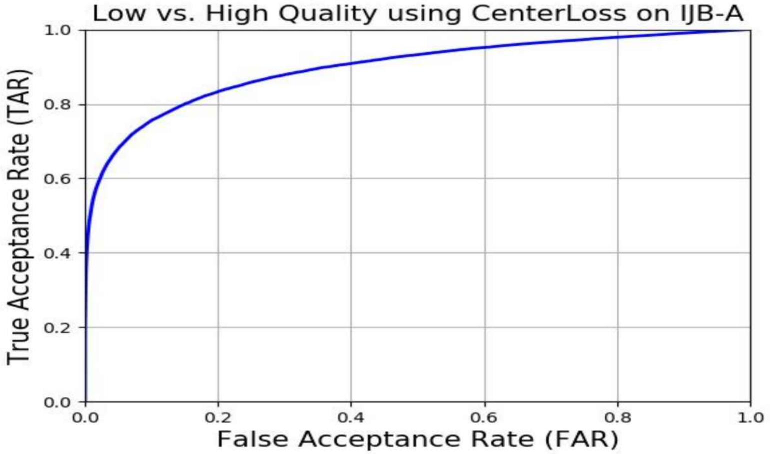


- LightCNN on FaceScrub:

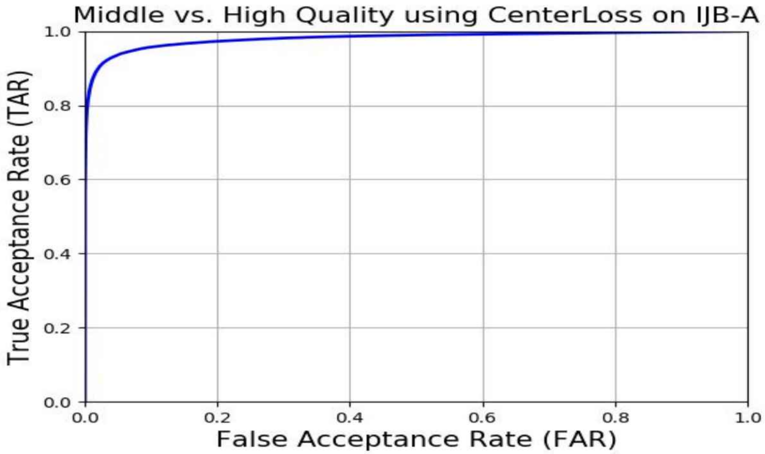

Low vs. High Quality using Lightened CNN on FaceScrub


Middle vs. High Quality using Lightened CNN on FaceScrub

- CenterLoss on IJB-A:

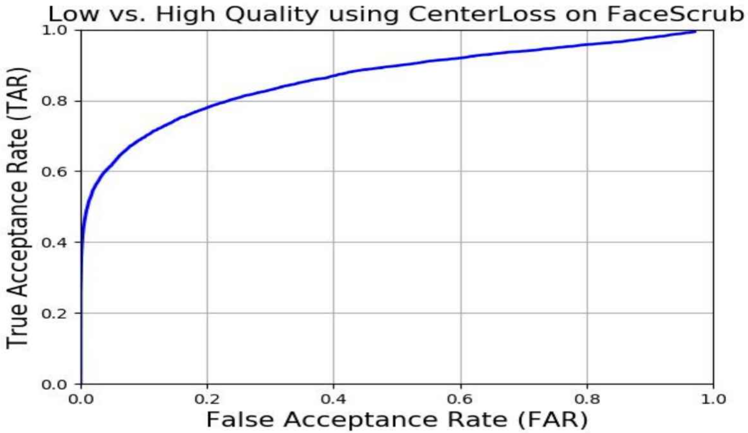


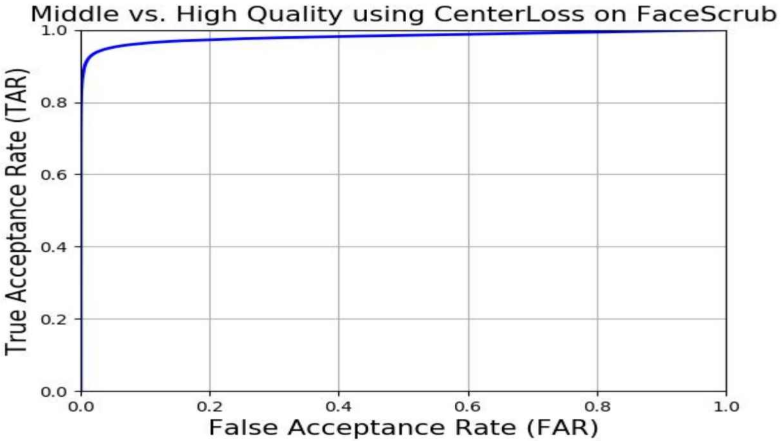
Low vs. High Quality using CenterLoss on IJB-A



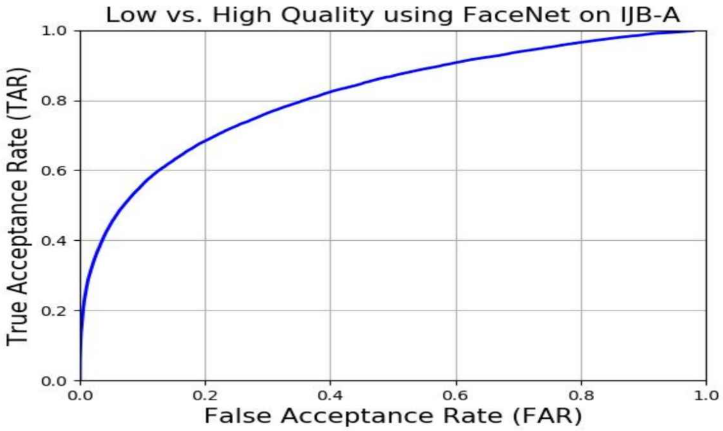
Middle vs. High Quality using CenterLoss on IJB-A

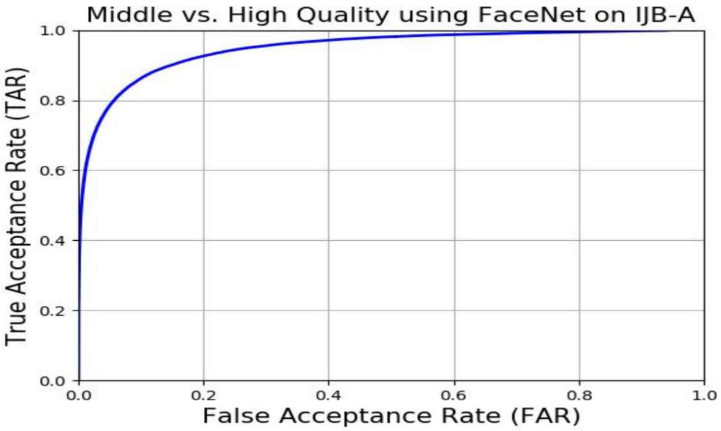
- CenterLoss on FaceScrub:



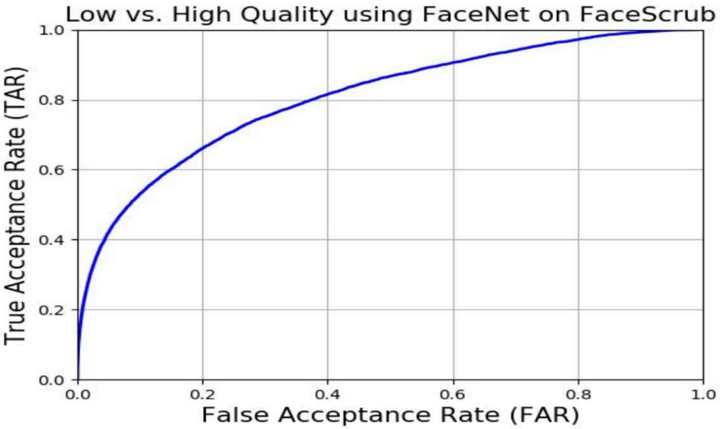


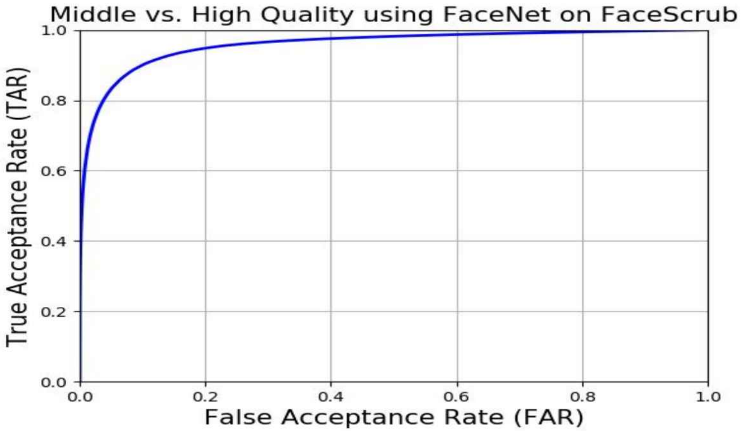
- FaceNet on IJB-A:



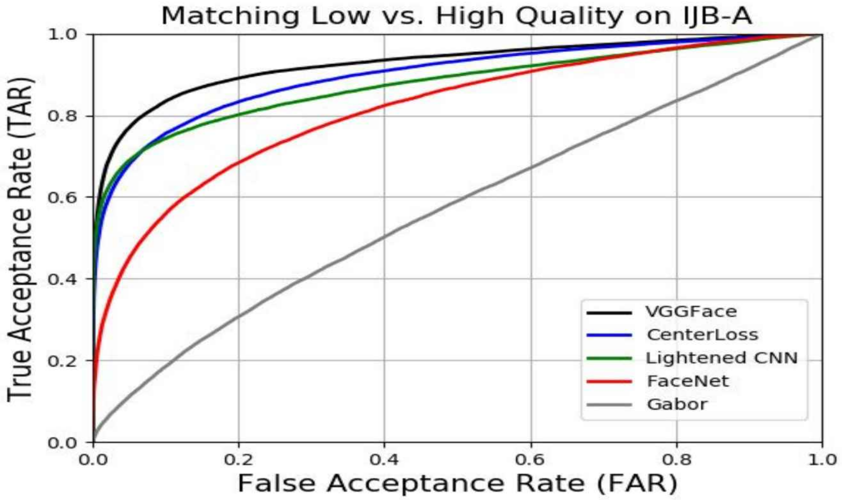


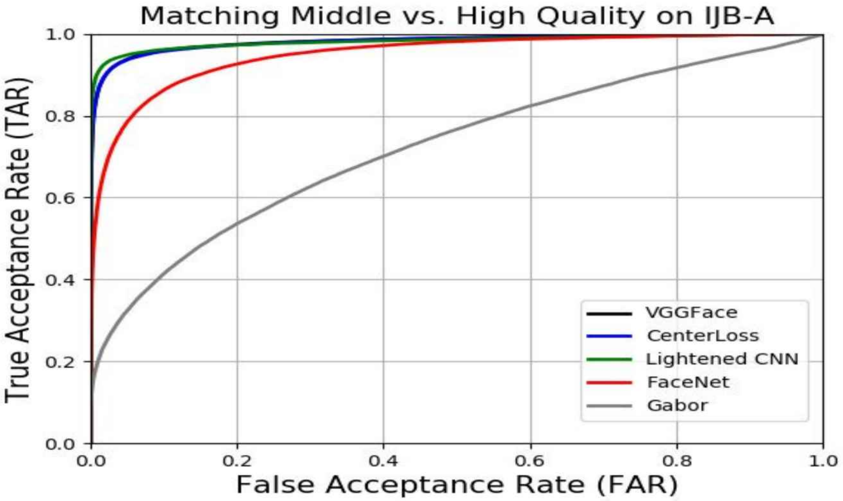
- FaceNet on FaceScrub:



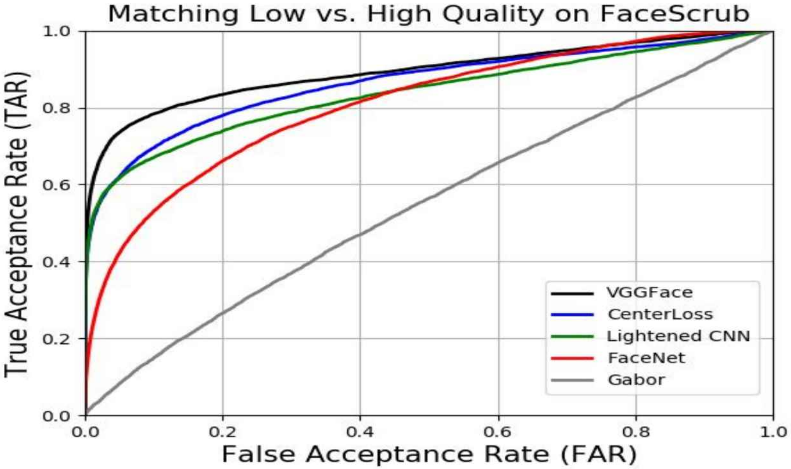


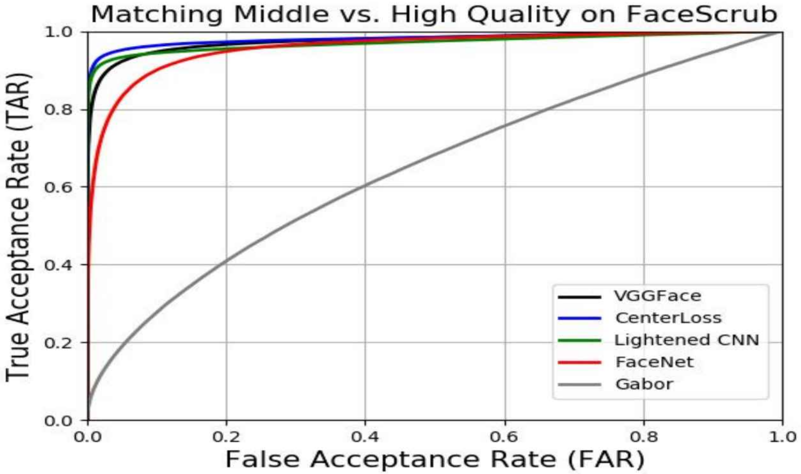
IJB-A


Matching Low vs. High Quality on IJB-A
True Acceptance Rate (TAR)
False Acceptance Rate (FAR)
0.0
0.2
0.4
0.6
0.8
1.0
VGGFace
CenterLoss
Lightened CNN
FaceNet
Gabor


Matching Middle vs. High Quality on IJB-A
True Acceptance Rate (TAR)
False Acceptance Rate (FAR)
0.0
0.2
0.4
0.6
0.8
1.0
VGGFace
CenterLoss
Lightened CNN
FaceNet
Gabor

FaceScrub


Matching Low vs. High Quality on FaceScrub
True Acceptance Rate (TAR)
False Acceptance Rate (FAR)
0.0
0.2
0.4
0.6
0.8
1.0
VGGFace
CenterLoss
Lightened CNN
FaceNet
Gabor


Matching Middle vs. High Quality on FaceScrub
True Acceptance Rate (TAR)
False Acceptance Rate (FAR)
0.0
0.2
0.4
0.6
0.8
1.0
VGGFace
CenterLoss
Lightened CNN
FaceNet
Gabor

Verification Result

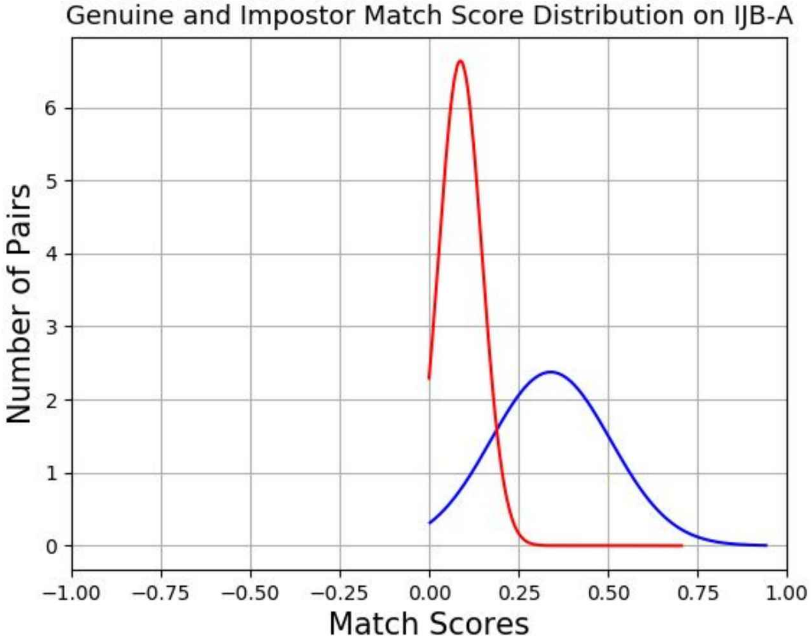
Dataset	Model	Low to High			Middle to High		
		FAR=0.01	0.001	0.0001	0.01	0.001	0.0001
IJB-A	VGGFace	0.605	0.367	0.194	0.858	0.675	0.491
	Lightened CNN	0.566	0.402	0.269	0.905	0.808	0.678
	CenterLoss	0.521	0.313	0.164	0.859	0.692	0.499
	FaceNet	0.257	0.100	0.033	0.586	0.330	0.165
FaceScrub	VGGFace	0.595	0.389	0.231	0.837	0.662	0.468
	Lightened CNN	0.503	0.330	0.148	0.896	0.811	0.668
	CenterLoss	0.493	0.341	0.215	0.914	0.814	0.652
	FaceNet	0.219	0.075	0.019	0.633	0.350	0.162

Choose the Best Deep Model on Low vs. High Matching

- IJB-A
  - VGGFace
- FaceScrub
  - VGGFace

Decision Boundary: IJB-A, VGGFace

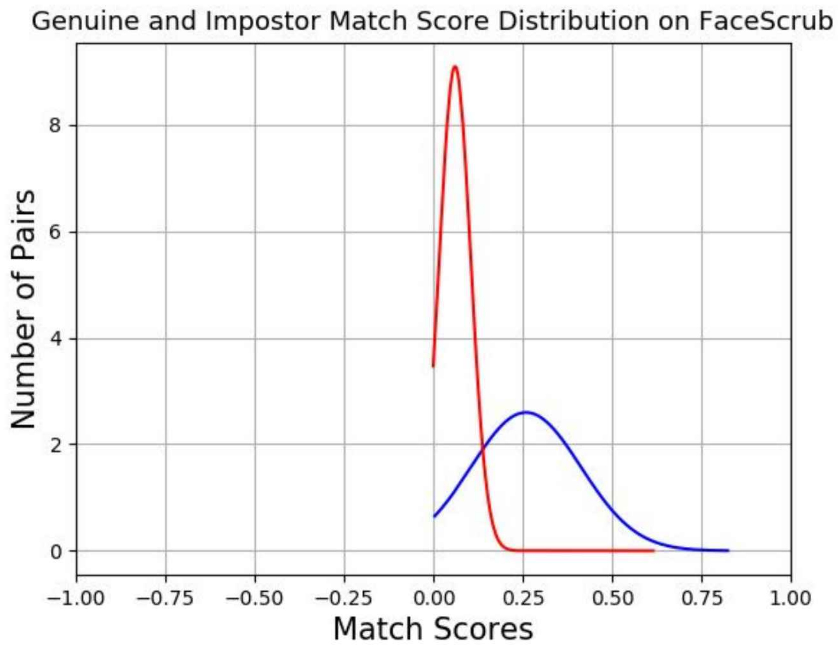
- Matching Score Threshold:
  - 0.188121



Choose Consine Similarity Score as Match Score

Decision Boundary: FaceScrub, VGGFace

- Matching Score Threshold:
  - 0.138071



Positive and Negative Pairs

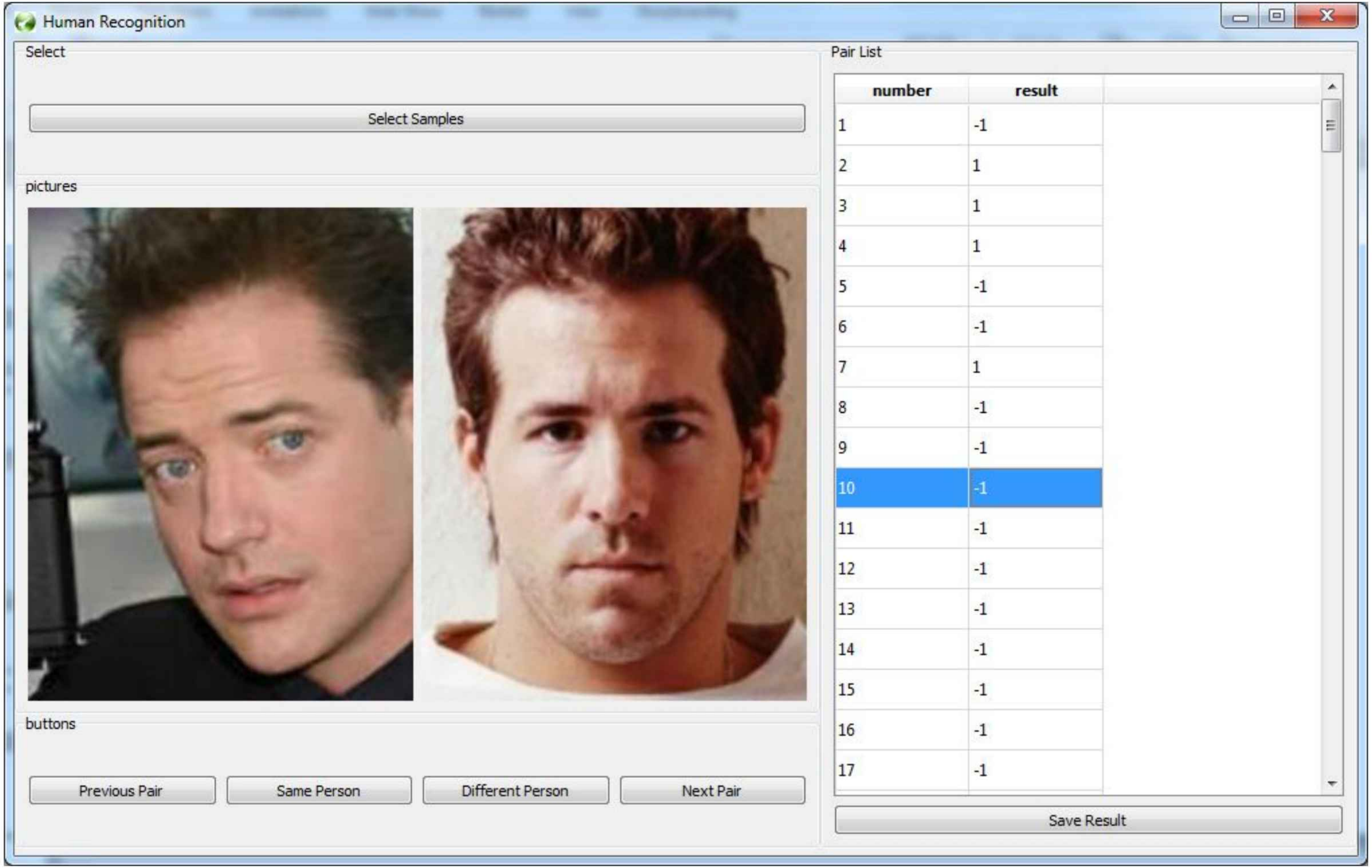
- Use threshold of each dataset to filter all pairs
- Filtered Pairs
  - IJB-A
    - ✔ Positive pairs: Match Score < 0.188121
    - ✔ Negative pairs: Match Score >= 0.188121
  - FaceScrub
    - ✔ Positive pairs: Match Score < 0.138071
    - ✔ Negative pairs: Match Score >= 0.138071
- Randomly select 100 positive pairs and 100 negative pairs from each dataset
- In this case, deep model recognition rate is 0% correct

Experiment

- We recruit a number of participants to visually check all face pairs to determine if each face pair showed in front of them belong to the same identity or different identities.

- For convenience, we developed a tool based on Python language to aid participants perform this experiment

Tool



number	result
1	-1
2	1
3	1
4	1
5	-1
6	-1
7	1
8	-1
9	-1
10	-1
11	-1
12	-1
13	-1
14	-1
15	-1
16	-1
17	-1

Participants

- A total of 20 participants
  - Male: 14
  - Female: 6
- Some participants has much experience on face images quality
- Some know about face image quality
- And others have never worked on facial image analysis using a computer

Procedure

- For each dataset
  - There are 100 positive pairs and 100 negative pairs
  - Randomize all the pairs (200 pairs)
  - Divide all the pairs into four subsets, each contains 50 pairs
- Finally, we get 8 subsets in total
- Participants view two images side by side for each subset
- When finish one subset, participants are asked to do next subset after a pretty good rest
- Participants have unlimited time to finish it

- Participants are asked to rate each pair of images
  - 1: same subject
  - -1: different subjects

Result

- We divide all participants into three groups
  - Group1: Have much experience on face image quality
    3 participants
  - Group2: Working on some facial image analysis tasks
    4 participants
  - Group3: Never worked on facial image analysis with a computer
    13 participants
- For each group
  - Majority Voting to get result of each images pair
  - Draw ROC curve and confusion matrix
  - Calculate Accuracy

IJB-A

IJB-A: All

Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	81%	19%	84%
	Negative	13%	87%	

IJB-A: Group1

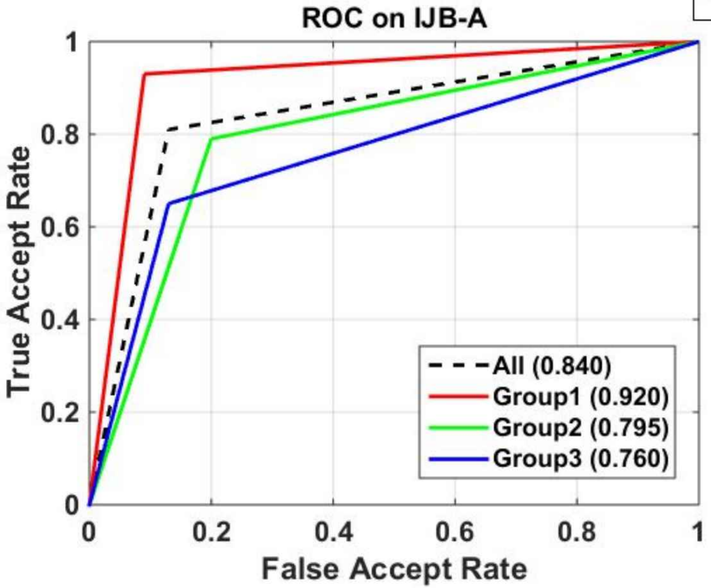
Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	93%	7%	92%
	Negative	9%	91%	

IJB-A: Group2

Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	79%	21%	79.5%
	Negative	20%	80%	

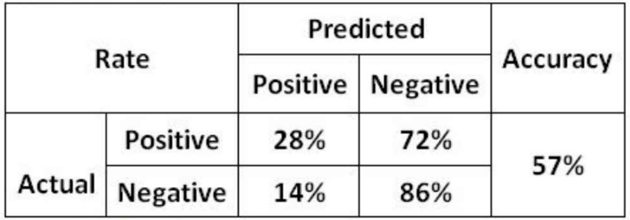
IJB-A: Group3

Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	65%	35%	76%
	Negative	13%	87%	



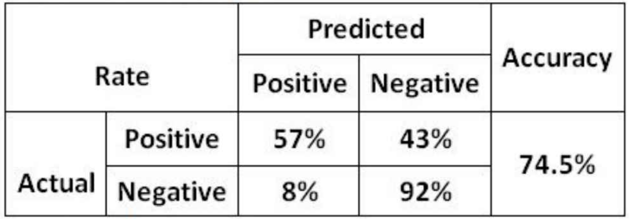
FaceScrub

FaceScrub: All



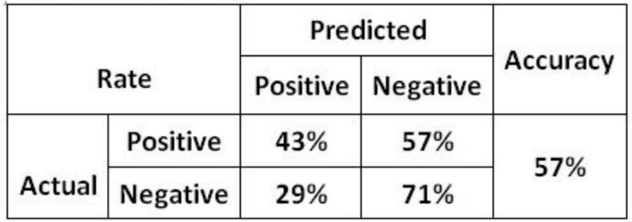
Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	28%	72%	57%
	Negative	14%	86%	

FaceScrub: Group1



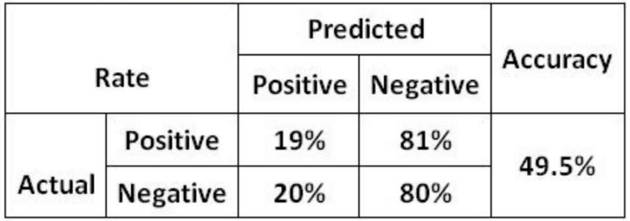
Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	57%	43%	74.5%
	Negative	8%	92%	

FaceScrub: Group2



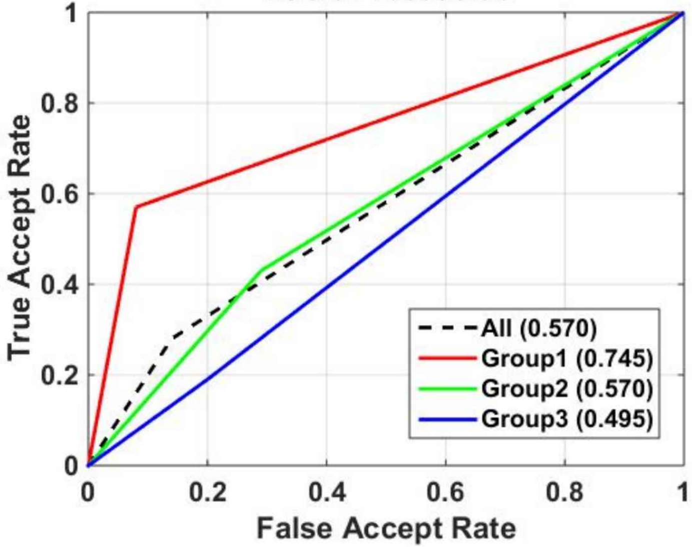
Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	43%	57%	57%
	Negative	29%	71%	

FaceScrub: Group3



Rate		Predicted		Accuracy
		Positive	Negative	
Actual	Positive	19%	81%	49.5%
	Negative	20%	80%	

ROC on FaceScrub



Conclusion

- People has experience of face recognition performs better than those has not.
- People has higher accuracy in recognition of negative pairs than that of positive pairs.
- Hard to recognize positive pairs since quality is low; for negative pairs, it is easier to view them as negative(different persons)

- Accuracy on Facescrub is lower than IJB-A
- FaceScrub low quality images has lower quality than IJB-A’s (quality score!)