CSE 252C: Advanced Computer Vision

Manmohan Chandraker

Lecture 9: Face Recognition
Cooperative Face Recognition

- People stand in front of a camera with good illumination conditions.
  - Border pass, access control, attendance
Unconstrained Face Recognition

- Images captured with less user cooperation, in more challenging conditions
  - Video surveillance, hand held system
Partial Faces in Unconstrained Scenes

[Liao et al., Partial Face Recognition, PAMI 2013]
Face Recognition on LFW Benchmark

- Human performance: 99.20%
- Local Binary Patterns: 95.17%
Face Recognition on LFW Benchmark

- Human performance: 99.20%
- Local Binary Patterns: 95.17%
- DeepFace: 97.35%
Face Recognition on LFW Benchmark

- Human performance: **99.20%**
- Local Binary Patterns: 95.17%
- DeepFace: 97.35%
- DeepID2: 99.15%
Face Recognition on LFW Benchmark

- Human performance: **99.20%**
- Local Binary Patterns: 95.17%
- DeepFace: 97.35%
- DeepID2: 99.15%
- FaceNet: **99.63%**
Progress in Face Recognition

- Human performance: 99.20%
- Local Binary Patterns: 95.17%
- DeepFace: 97.35%
- DeepID2: 99.15%
- FaceNet: 99.63%
Modules in Face Recognition Pipelines

- Input Image
- Face Detection
- Face Alignment
- Anti-Spoofing

Training:
- Training data after processing
- Feature extraction
- Loss function

Testing:
- Test data after processing
- Feature extraction
- Face matching

Deep face recognition
Axes for Studying Face Recognition

Face Verification:
  - Specify whether pair of images belong to the same person

13K images, 5.7K people

Standard benchmark in the community

Test protocols depending upon availability of outside training data

[ G. Huang, M. Ramesh, T. Berg and E. Learned-Miller]
IJB-A Dataset

(a) Frontal, Cooperative subject, Controlled environment
(b) Near frontal, uncooperative, minimal environment variations (e.g., LFW)
(c) Full variation in pose, illumination, environment

Automated detection ability:
- Human performance
- Near human performance
- Cannot detect consistently

Automated recognition ability:
- Human performance
- Near human performance
- Cannot recognize

[Klare et al., CVPR 2015]
MS-1M Dataset

(a) Professions

(b) Nationality

(c) Date of Birth

(d) Gender

[Guo et al., ECCV 2016]
Face Datasets

Variation in number of faces per identity

Variation in approximate labeling noise levels

[Guo et al., ECCV 2016]
## Face Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Identities</th>
<th>Images</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFW</td>
<td>5,749</td>
<td>13,233</td>
<td>Small, used for testing, saturated</td>
</tr>
<tr>
<td>Celeb Faces</td>
<td>10177</td>
<td>202,599</td>
<td>Many identities, attribute labels</td>
</tr>
<tr>
<td>VGG-Face</td>
<td>2622</td>
<td>1,635,159</td>
<td>Many examples per class, somewhat noisy</td>
</tr>
<tr>
<td>IJB-A</td>
<td>500</td>
<td>5k images, 2k videos</td>
<td>Challenging pose, lighting, quality</td>
</tr>
<tr>
<td>MS-1M</td>
<td>80k</td>
<td>7M</td>
<td>Largest public dataset for training (currently)</td>
</tr>
<tr>
<td>Facebook</td>
<td>4030</td>
<td>4.4M</td>
<td>Proprietary, used in DeepFace</td>
</tr>
<tr>
<td>Google</td>
<td>8M</td>
<td>200M</td>
<td>Proprietary, used in FaceNet</td>
</tr>
</tbody>
</table>

---

### Timeline of Face Datasets

- **1994**: ORL
- **1996**: FRGC, AR
- **1998**: CAS-PEAL
- **2000**: FRGCv2 (3D), MORPH (cross-age)
- **2004**: BU-3DFE (3D), Multi-PIE (cross-age)
- **2006**: LFW (cross-age), Bosporus (3D)
- **2008**: CUFs (sketch-photo), CUFsf (sketch-photo)
- **2010**: FG-NET (cross-age), CASIA-HFB (cross-age)
- **2012**: CASIA-NR-VIS (2.0), CASIA WebFace (50K, 10K)
- **2014**: CASIA NR-VIS (3.0), (cross-age)
- **2016**: VGGFace (2.6M, 2.6K)
- **2018**: Facebook (4.4M), Google (8M)

### Examples

- 1994: ORL
- 1996: FRGC, AR
- 1998: CAS-PEAL
- 2000: FRGCv2 (3D), MORPH (cross-age)
- 2004: BU-3DFE (3D), Multi-PIE (cross-age)
- 2006: LFW (cross-age), Bosporus (3D)
- 2008: CUFs (sketch-photo), CUFsf (sketch-photo)
- 2010: FG-NET (cross-age), CASIA-HFB (cross-age)
- 2012: CASIA-NR-VIS (2.0), CASIA WebFace (50K, 10K)
- 2014: CASIA NR-VIS (3.0), (cross-age)
- 2016: VGGFace (2.6M, 2.6K)
- 2018: Facebook (4.4M), Google (8M)
Axes for Studying Face Recognition

**Loss**
- Euclidean distance
- Angular margin
- Softmax variations

**Architecture**
- Backbone Networks
- Assembled Networks

**Data Process**
- One to many augmentation
- Many to one normalization

**Data**
- MS-Celeb-1M
- VGGFace 2
- CASIA-Webface
- ... (other datasets)
- IJB-A
- FG-Net
- CP/CA/S L-LFW
- ... (other datasets)

**Specific scenario**
- Low shot
- Anti-spoofing
- Cross age
- Cross pose
- NIV-VIS
- Make-up
- Template-based
- 3D
- Video
- Photosketch
Face Processing

**Augmentation**

- **One-to-Many Augmentation**: mitigate difficulty of diverse data collections
  - Generate 3D pose-variant faces from frontal inputs, use for training
  - Use GANs or other methods to generate faces with diverse attributes

**Normalization**

- **Many-to-One Normalization**: reduce variation in test-time inputs
  - Generate frontal face from pose-variant input
  - Use GANs or methods to generate faces with neutral attributes

---

[Zhu et al., “3DDFA”]

[Yin et al., “FF-GAN”]
**Face Processing**

- **One-to-Many Augmentation**: mitigate difficulty of diverse data collections
  - Generate 3D pose-variant faces from frontal inputs, use for training
  - Use GANs or other methods to generate faces with diverse attributes

- **Many-to-One Normalization**: reduce variation in test-time inputs
  - Generate frontal face from pose-variant input
  - Use GANs or methods to generate faces with neutral attributes
Axes for Studying Face Recognition

### Loss
- Euclidean distance
- Angular margin
- Softmax variations

### Architecture
- Backbone Networks
- Assembled Networks

### Data Process
- One to many augmentation
- Many to one normalization

### Data
- MS-Celeb-1M
- VGGFace 2
- CASIA-Webface
- LJB-A
- FG-Net
- CP/CA/S L-LFW

### Specific scenario
- Low shot
- Video
- Cross age
- Anti-spoofing
- Cross pose
- Video
- NIV-VIS
- Make-up
- Photo-sketch
- Template-based
- 3D
Network Architectures

**Backbone networks**

- **AlexNet** (12/2012)
- **VGGNet** (2014)
- **GoogleNet** (6/2015)
- **ResNet** (6/2016)
- **SENet** (9/2017)

- **Deepface** (7/2014) (AlexNet)
- **Facenet** (6/2015) (GoogleNet)
- **SphereFace** (7/2017) (ResNet)
- **VGGFace2** (11/2017) (SENet)

**Multi-tasked networks**

[Peng et al., ICCV 2017]
Axes for Studying Face Recognition

Loss
- Euclidean distance
- Angular margin
- Softmax variations

Architecture
- Backbone Networks
- Assembled Networks

Data Process
- One to many augmentation
- Many to one normalization

Data
- MS-Celeb-1M
- VGGFace 2
- CASIA-Webface
- IJB-A
- FG-Net
- CP/CA/S L-LFW

Specific scenario
- Low shot
- Anti-spoofing
- Cross age
- Cross pose
- Make-up
- 3D
- Photo-sketch
- NIV-VIS
- ...
Loss Functions

Large-margin losses and softmax variants

- DeepID2 (contrastive loss)
- DeepID (softmax)
- DeepID3 (contrastive loss)
- Deepface (softmax)
- DeepID2+ (contrastive loss)
- FaceNet (triplet loss)
- VGGface (triplet+softmax)
- L-softmax (large margin)
- TSE (triplet loss)
- Range loss
- L2 softmax (feature normalization)
- TPE (triplet loss)
- Center loss (center loss)
- Marginal loss
- vMF loss (weight and feature normalization)
- Normface (feature normalization)
- A-softmax (large margin)
- Cosface (large margin)
- CoCo loss (feature normalization)
- Arcface (large margin)
- Center invariant loss (center loss)
- AMS loss (large margin)

Timeline:
- 2014: Softmax loss, Contrastive loss
- 2015: Learning, Triplet loss
- 2016: Center loss
- 2017: Feature and weight normalization
- 2018: Large margin loss
<table>
<thead>
<tr>
<th>Method</th>
<th>Public. Time</th>
<th>Loss</th>
<th>Architecture</th>
<th>Number of Networks</th>
<th>Training Set</th>
<th>Accuracy ± Std(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFace [195]</td>
<td>2014</td>
<td>softmax</td>
<td>Alexnet</td>
<td>3</td>
<td>Facebook (4.4M, 4K)</td>
<td>97.35 ± 0.25</td>
</tr>
<tr>
<td>DeepID2 [187]</td>
<td>2014</td>
<td>contrastive loss</td>
<td>Alexnet</td>
<td>25</td>
<td>CelebFaces+ (0.2M, 10K)</td>
<td>99.15 ± 0.13</td>
</tr>
<tr>
<td>DeepID3 [188]</td>
<td>2015</td>
<td>contrastive loss</td>
<td>VGGNet-10</td>
<td>50</td>
<td>CelebFaces+ (0.2M, 10K)</td>
<td>99.53 ± 0.10</td>
</tr>
<tr>
<td>FaccNet [176]</td>
<td>2015</td>
<td>triplet loss</td>
<td>GoogleNct-24</td>
<td>1</td>
<td>Google (500M, 10M)</td>
<td>99.63 ± 0.09</td>
</tr>
<tr>
<td>Baidu [124]</td>
<td>2015</td>
<td>triplet loss</td>
<td>CNN-9</td>
<td>10</td>
<td>Baidu (1.2M, 18K)</td>
<td>99.77</td>
</tr>
<tr>
<td>VGGface [149]</td>
<td>2015</td>
<td>triplet loss</td>
<td>VGGNet-16</td>
<td>1</td>
<td>VGGface (2.6M, 2.6K)</td>
<td>98.95</td>
</tr>
<tr>
<td>light-CNN [225]</td>
<td>2015</td>
<td>softmax</td>
<td>light CNN</td>
<td>1</td>
<td>MS-Celeb-1M (8.4M, 100K)</td>
<td>98.8</td>
</tr>
<tr>
<td>Center Loss [218]</td>
<td>2016</td>
<td>center loss</td>
<td>Lenet+7</td>
<td>1</td>
<td>CASIA-WebFace, CACD2000, Celebrity+ (0.7M, 17K)</td>
<td>99.28</td>
</tr>
<tr>
<td>L-softmax [126]</td>
<td>2016</td>
<td>L-softmax</td>
<td>VGGNet-18</td>
<td>1</td>
<td>CASIA-WebFace (0.49M, 10K)</td>
<td>98.71</td>
</tr>
<tr>
<td>Range Loss [261]</td>
<td>2016</td>
<td>range loss</td>
<td>VGGNet-16</td>
<td>1</td>
<td>MS-Celeb-1M, CASIA-WebFace (5M, 100K)</td>
<td>99.52</td>
</tr>
<tr>
<td>L2-softmax [157]</td>
<td>2017</td>
<td>L2-softmax</td>
<td>ResNet-101</td>
<td>1</td>
<td>MS-Celeb-1M (3.7M, 58K)</td>
<td>99.78</td>
</tr>
<tr>
<td>Normface [206]</td>
<td>2017</td>
<td>contrastive loss</td>
<td>ResNet-28</td>
<td>1</td>
<td>CASIA-WebFace (0.49M, 10K)</td>
<td>99.19</td>
</tr>
<tr>
<td>CoCo loss [130]</td>
<td>2017</td>
<td>CoCo loss</td>
<td>-</td>
<td>1</td>
<td>MS-Celeb-1M (3M, 80K)</td>
<td>99.86</td>
</tr>
<tr>
<td>vMF loss [75]</td>
<td>2017</td>
<td>vMF loss</td>
<td>ResNet-27</td>
<td>1</td>
<td>MS-Celeb-1M (4.6M, 60K)</td>
<td>99.58</td>
</tr>
<tr>
<td>SphereFace [125]</td>
<td>2017</td>
<td>A-softmax</td>
<td>ResNet-64</td>
<td>1</td>
<td>CASIA-WebFace (0.49M, 10K)</td>
<td>99.42</td>
</tr>
<tr>
<td>CCL [155]</td>
<td>2018</td>
<td>center invariant loss</td>
<td>ResNet-27</td>
<td>1</td>
<td>CASIA-WebFace (0.49M, 10K)</td>
<td>99.12</td>
</tr>
<tr>
<td>AMS loss [205]</td>
<td>2018</td>
<td>AMS loss</td>
<td>ResNet-20</td>
<td>1</td>
<td>CASIA-WebFace (0.49M, 10K)</td>
<td>99.12</td>
</tr>
<tr>
<td>Cosface [207]</td>
<td>2018</td>
<td>cosface</td>
<td>ResNet-64</td>
<td>1</td>
<td>CASIA-WebFace (0.49M, 10K)</td>
<td>99.33</td>
</tr>
<tr>
<td>Arcface [42]</td>
<td>2018</td>
<td>Arcface</td>
<td>ResNet-100</td>
<td>1</td>
<td>MS-Celeb-1M (3.8M, 85K)</td>
<td>99.83</td>
</tr>
<tr>
<td>Ring loss [272]</td>
<td>2018</td>
<td>Ring loss</td>
<td>ResNet-64</td>
<td>1</td>
<td>MS-Celeb-1M (3.5M, 31K)</td>
<td>99.50</td>
</tr>
</tbody>
</table>
Learning Face Representations
Steps in Face Recognition

• Face Detection
  – Localize the face

• Face Alignment
  – Factor out 3D transformation

• Feature Extraction
  – Find compact representation

• Classification
  – Answer the question
Challenges in Face Alignment

• Infer 3D from 2D
  – Slight occlusion
  – Lighting condition
  – Head orientation
  – Non rigid deformation
DeepFace Alignment: Substep 1

- 2D feature point extraction
- 2D alignment $x_{anchor} = (S \ast R \ast T)x_{source}$
- Only for in plane alignment

[Taigman et al., DeepFace, CVPR 2014]
DeepFace Alignment: Substep 2

- 3D feature point extraction
- 3D alignment: piecewise affine transformation
- No perspective correction

\[
\min r^T \Sigma^{-1} r \\
\text{where } r = x_{2D} - x_{3D}P
\]

[Taigman et al., DeepFace, CVPR 2014]
Architecture

Layer 1-3: Intuition

- Convolution layers - extract low-level features (e.g. simple edges and texture)
- ReLU after each conv. layer
- Max-pooling: make convolution network more robust to local translations.

[Taigman et al., DeepFace, CVPR 2014]
**Architecture**

Layer 4-6: Intuition

- Apply filters to different locations on the map
- Similar to a conv. layer but spatially dependent

- Different regions of an aligned image have different local statistics
- Spatial stationarity motivation for convolution does not hold
- Large increase in number of parameters, but inference time is similar
- Can be done under two conditions
  - Aligned images with similar semantic concepts are being considered
  - A large training dataset is available

[Taigman et al., DeepFace, CVPR 2014]
Architecture

- Layer F7 is fully connected and generates 4096d vector
- Sparse representation of face descriptor
- 75% of outputs are zero

- Layer F8 is fully connected and generates 4030d vector
- F8 calculates probability with softmax: $p_k = \frac{\exp(o_k)}{\sum_h \exp(o_h)}$
- Cross-entropy loss function: $L = -\sum_k \log(p_k)$

[Taigman et al., DeepFace, CVPR 2014]
Verification and Identification Signals

Verification:
- Match two images of an individual across large appearance variations
- Favors tight clusters for each identity

Identification:
- Distinguish images of one identity from another identity
- Favors large distance between clusters
Verification and Identification Signals

Learn face representations from:

- Prediction becomes richer
- Prediction becomes more challenging
- Supervision becomes stronger
- Feature learning becomes more effective

- Predicting binary labels (verification)
- Predicting multi-class labels (identification)
- Predicting thousands of real-valued pixels (multi-view) reconstruction
Network Structure

- Locally connected layer at the top
  - Respond to facial features at preferred spatial locations
- Feature layer fully connected to last convolutional and locally connected layer
  - Multiscale information
  - Face representation: \( f = \text{Conv}(x, \theta_c) \)

[Sun et al., DeepID2, NIPS 2014]
Identification Signal

- **Identification**: connect feature layer to n-way softmax layer
  - Outputs a probability distribution over n classes
  - Train with a cross-entropy loss

\[
\text{Ident}(f, t, \theta_{id}) = -\sum_{i=1}^{n} -p_i \log \hat{p}_i
\]

- Feature  Target class  Target probability distribution  Predicted probability distribution

- **Goal**: is to correctly classify all identities simultaneously
  - Incentivize learning discriminative features across inter-personal variations

[Sun et al., DeepID2, NIPS 2014]
Verification Signal

- Verification: directly regularize the feature vector
  - Pairwise: Gather faces from same class, push those from different classes

\[
\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \begin{cases} 
\frac{1}{2} \| f_i - f_j \|_2^2 & \text{if } y_{ij} = 1 \\
\frac{1}{2} \max (0, m - \| f_i - f_j \|_2)^2 & \text{if } y_{ij} = -1
\end{cases}
\]

- Cosine similarity:

\[
\text{Verif}(f_i, f_j, y_{ij}, \theta_{ve}) = \frac{1}{2} (y_{ij} - \sigma(wd + b))^2 , \text{ binary } y_{ij}, \quad d = \frac{f_i \cdot f_j}{\| f_i \|_2 \| f_j \|_2}
\]

- Goal is to learn features that can be matched across intra-personal variations

[Sun et al., DeepID2, NIPS 2014]
Training Process

- Goal: learn features and parameters for convolution layer, identity, verification
  - High $\lambda$: more weight on verification

**input:** training set $\chi = \{(x_i, l_i)\}$, initialized parameters $\theta_c$, $\theta_{id}$, and $\theta_{ve}$, hyperparameter $\lambda$, learning rate $\eta(t)$, $t \leftarrow 0$

**while** not converge **do**

$t \leftarrow t + 1$ sample two training samples $(x_i, l_i)$ and $(x_j, l_j)$ from $\chi$

$f_i = \text{Conv}(x_i, \theta_c)$ and $f_j = \text{Conv}(x_j, \theta_c)$

$\nabla \theta_{id} = \frac{\partial \text{Ident}(f_i, l_i, \theta_{id})}{\partial \theta_{id}} + \frac{\partial \text{Ident}(f_j, l_j, \theta_{id})}{\partial \theta_{id}}$

$\nabla \theta_{ve} = \lambda \cdot \frac{\partial \text{Verif}(f_i, f_j, y_{ij}, \theta_{ve})}{\partial \theta_{ve}}$, where $y_{ij} = 1$ if $l_i = l_j$, and $y_{ij} = -1$ otherwise.

$\nabla f_i = \frac{\partial \text{Ident}(f_i, l_i, \theta_{id})}{\partial f_i} + \lambda \cdot \frac{\partial \text{Verif}(f_i, f_j, y_{ij}, \theta_{ve})}{\partial f_i}$

$\nabla f_j = \frac{\partial \text{Ident}(f_j, l_j, \theta_{id})}{\partial f_j} + \lambda \cdot \frac{\partial \text{Verif}(f_i, f_j, y_{ij}, \theta_{ve})}{\partial f_j}$

$\nabla \theta_c = \nabla f_i \cdot \frac{\partial \text{Conv}(x_i, \theta_c)}{\partial \theta_c} + \nabla f_j \cdot \frac{\partial \text{Conv}(x_j, \theta_c)}{\partial \theta_c}$

update $\theta_{id} = \theta_{id} - \eta(t) \cdot \theta_{id}$, $\theta_{ve} = \theta_{ve} - \eta(t) \cdot \theta_{ve}$, and $\theta_c = \theta_c - \eta(t) \cdot \theta_c$.

**end while**

**output** $\theta_c$

[Sun et al., DeepID2, NIPS 2014]
Balancing Identification and Verification

- Balance required between signals to learn good features
Balancing Identification and Verification

- Inter-class scatter: \[ \sum_{i=1}^{c} n_i \cdot (\bar{x}_i - \bar{x}) (\bar{x}_i - \bar{x})^T \]
- Intra-class scatter: \[ \sum_{i=1}^{c} \sum_{x \in D_i} (x - \bar{x}_i) (x - \bar{x}_i)^T \]
- Variance in scatter indicated by size of eigenvalues
  - Small number of eigenvectors: diversity of variation is low
  - Both diversity and magnitude of feature variance matters for recognition
Balancing Identification and Verification

- When only identification signal is used ($\lambda = 0$):
  - High diversity in both inter-personal and intra-personal features
  - Good for identification since it helps distinguish different identities
  - But large intra-personal variance is noise for verification
Balancing Identification and Verification

- When only identification signal is used ($\lambda = 0$):
  - High diversity in both inter-personal and intra-personal features
  - Good for identification since it helps distinguish different identities
  - But large intra-personal variance is noise for verification

- When only verification signal is used ($\lambda$ approaches $+\infty$):
  - Both intra-personal and inter-personal variance collapse to few directions
  - Good for verification, but cannot distinguish many classes in identification
Balancing Identification and Verification

- When both verification and identification signals are used ($\lambda = 0.05$):
  - Inter-personal variations stay high
  - Intra-personal variations reduce in diversity and magnitude
Balancing Identification and Verification

- Visualize features for 6 identities
- With only identification signal:
  - Cluster centers are well-separated, but large cluster size causes overlap
- With only verification signal:
  - Cluster sizes become small, but cluster centers also collapse
- With both signals:
  - Clusters sizes become small and cluster centers are reasonably separated