Deep multi-task learning for gait-based biometrics

- Manuel J. Marín-Jiménez (Univ. of Córdoba)
- Francisco M. Castro (Univ. of Málaga)
- Nicolás Guil (Univ. of Málaga)
- Fernando de la Torre (Carnegie Mellon University)
- Rafael Medina-Carnicer (Univ. of Córdoba)
Outline

1. Problem definition

2. Our approach
   i. Input data
   ii. Deep Multi-task Model

3. Experiments and results

4. Conclusions and future work

ICIP 2017 - mjmarin@uco.es
Motivation

- “Who’s he?”
- “I cannot distinguish his face”
- “But he walks like Peter”

We are good at identifying people at a distance

- But, why?
- Because each person has his/her own gait pattern → gait signature
The problem

Objective: identify people based on the way they walk
→ Gait recognition

Input: video sequence
Output: identity

+ Gender
+ Age
Previous approaches

Context: video surveillance, control access,…

Gait Energy Image (GEI)
[Han PAMI06]

Chrono-Gait Image (CGI)
[Wang PAMI12]

Pyramidal Fisher Motion (PFM)
[Castro IJPRAI17]

*Images extracted from their corresponding papers: [Han PAMI06], [Wang PAMI12], [Castro IJPRAI17]

ICIP 2017 - mjmarin@uco.es
Outline

1. Problem definition

2. Our approach
   i. Input data
   ii. Deep Multi-task Model

3. Experiments and results

4. Conclusions and future work
Multi-task CNN

- Goal: identify people + age + gender
Outline

1. Problem definition

2. Our approach
   i. Input data
   ii. Deep Multi-task Model

3. Experiments and results

4. Conclusions and future work
Input data

- Optical flow channels: \{OF-x, OF-y\} @ 80x60 pix

Input data

- Optical flow channels: \{OF-x, OF-y\}
- Fixed length: 25 frames (~1 gait cycle)
- Crop frames to 1:1 aspect ratio
- Person centred in middle-frame

Input data

- Multiple subsequences are extracted

Outline

1. Problem definition

2. Our approach
   i. Input data
   ii. Deep Multi-task Model

3. Experiments and results

4. Conclusions and future work
Deep Multi-task (DMT) model

- **DMT loss** function:
  identification loss (main) + auxiliary losses.

\[
\mathcal{L}_{\text{DMT}}(g(v, \theta), Y) = \mathcal{L}_m(\hat{y}^m, y^m) + \sum_{t=1}^{T} \lambda_t \mathcal{L}_t(\hat{y}^t, y^t)
\]
Main task: identification

- Identification loss function: softmax log-loss

\[ \mathcal{L}_m(\hat{y}, c) = -\hat{y}_c + \log \sum_{k=1}^{C} e^{\hat{y}_k} \]

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
Aux task: gender recognition

- Gender loss function:
  softmax log-loss (two classes)

\[
\mathcal{L}_g (\hat{\mathbf{y}}, c) = -\hat{y}_c + \log \sum_{k=1}^{C} e^{\hat{y}_k}
\]
Aux task: age estimation

- Age loss function:
  Tukey’s biweight loss [Black96] → regression

\[
\mathcal{L}_a(\hat{y}_i, y_i) = \rho(r_i^{\text{MAD}})
\]

\[
\rho(r_i) = \begin{cases} 
\frac{c^2}{6} \left[1 - (1 - \frac{r_i}{c})^2\right]^3, & \text{if } |r_i| \leq c \\
\frac{c^2}{6}, & \text{otherwise}
\end{cases}
\]

\[
r_i^{\text{MAD}} = \frac{y_i - \hat{y}_i}{1.4826 \times \text{MAD}}
\]

\[
\text{MAD} = \text{median}_{k \in \{1, \ldots, S\}} \left( \left| r_k - \text{median}_{j \in \{1, \ldots, S\}} (r_j) \right| \right)
\]

\[c = 4.6851\]

Aux task: identity verification

- Verification loss function:
  L2 distance with margin [Hadsell06]

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

+1: same id
-1: different id

CNN architectures

**Convolutional:**
- spatial filters *(trainable, local)* + ReLU

**Fully-connected**

**Softmax:**
- values in (0,1) and adds up to 1

### CNN type A
- **conv1**: 7x7x64, stride 1, pool 2x2
- **conv2**: 5x5x128, stride 2, pool 2x2
- **conv3**: 3x3x512, stride 1, pool 2x2
- **conv4**: 2x2x512, stride 1
- **full5**: 256
- **softmax**

- N units (classes)

### CNN type B
- **conv1**: 7x7x64, stride 1, norm, pool 2x2
- **conv2**: 5x5x128, stride 2, pool 2x2
- **conv3**: 3x3x512, stride 1, pool 2x2
- **conv4**: 2x2x2048, stride 1
- **full5**: 1024
- **dropout**: 0.1
- **softmax**

- N units (classes)

**Gait signature**

ICIP 2017 - mjmarin@uco.es
Outline

1. Problem definition

2. Our approach
   i. Input data
   ii. Deep Multi-task Model

3. Experiments and results

4. Conclusions and future work
Dataset

**TUM-GAID** dataset [Hofmann JVCIR14]

- 305 subjects (100 train + 50 val + **155 test**)
- Scenarios:
  - Normal (N) + elapsed time (**TN**)
  - Bag (B) + elapsed time (**TB**)
  - Coating shoes (S) + elapsed time (**TS**)
- **Identity, age and gender** labels

Implementation details

• DMT training:
  - *task-wise early stopping* criterion [Zhang14]
  - back-propagation → SGD+momentum
  - batch: 256 samples
  - learning rate: 0.01 (reduced 0.1 if val error stuck)
  - max epochs: 30

• From subseqs to sequence-level decision:
  - Majority voting (e.g. on SVM scores)
  - Product of Softmax scores

Filters learnt

Convolutional filters @ first layer

- $a$: spatial derivatives
- $b$: temporal derivatives
Experiment: aux tasks contrib.

- Auxiliary tasks **speed up convergence** of the main task.
Experiment: aux tasks contrib.

- Auxiliary tasks **speed up convergence** of the main task.

\[ \text{AUC in parenthesis: lower is better} \]
Experiment: identification

- Identification results

Accuracy: higher is better

SMP: softmax product
State-of-the-art: comparison

CNN type B

<table>
<thead>
<tr>
<th>Layer</th>
<th>Filter Size</th>
<th>Output Channels</th>
<th>Stride</th>
<th>Normalization</th>
<th>Pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>7x7x64</td>
<td></td>
<td>1</td>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>conv2</td>
<td>5x5x128</td>
<td></td>
<td>2</td>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>conv3</td>
<td>3x3x512</td>
<td></td>
<td>1</td>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>conv4</td>
<td>2x2x2048</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>full5</td>
<td></td>
<td>1024</td>
<td></td>
<td>dropout 0.1</td>
<td></td>
</tr>
<tr>
<td>softmax</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N units (classes)</td>
</tr>
</tbody>
</table>

Castro et al. IWANN2017

<table>
<thead>
<tr>
<th>Layer</th>
<th>Filter Size</th>
<th>Output Channels</th>
<th>Stride</th>
<th>Normalization</th>
<th>Pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv1</td>
<td>7x7x96</td>
<td></td>
<td>1</td>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>conv2</td>
<td>5x5x192</td>
<td></td>
<td>2</td>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>conv3</td>
<td>3x3x512</td>
<td></td>
<td>1</td>
<td></td>
<td>2x2</td>
</tr>
<tr>
<td>conv4</td>
<td>2x2x4096</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>full5</td>
<td></td>
<td>4096</td>
<td></td>
<td>dropout</td>
<td></td>
</tr>
<tr>
<td>full6</td>
<td></td>
<td>2048</td>
<td></td>
<td>dropout</td>
<td></td>
</tr>
<tr>
<td>softmax</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Experiment: identification

<table>
<thead>
<tr>
<th>Method</th>
<th>N</th>
<th>B</th>
<th>S</th>
<th>Avg</th>
<th>TN</th>
<th>TB</th>
<th>TS</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>80x60</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ours-1 (SVM)</td>
<td>100</td>
<td>97.1</td>
<td>97.1</td>
<td>98.1</td>
<td>53.1</td>
<td>59.4</td>
<td>50</td>
<td>54.2</td>
</tr>
<tr>
<td>Ours-2 (SVM)</td>
<td>99.7</td>
<td>96.5</td>
<td>97.4</td>
<td>97.9</td>
<td>56.3</td>
<td>56.3</td>
<td>56.3</td>
<td>56.3</td>
</tr>
<tr>
<td>Ours-2 (7-NN)</td>
<td>99.7</td>
<td>97.4</td>
<td>99.7</td>
<td><strong>98.9</strong></td>
<td>59.4</td>
<td>62.5</td>
<td>68.8</td>
<td><strong>63.6</strong></td>
</tr>
<tr>
<td><strong>Castro17b-CNN (SVM)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Castro17b-CNN (SVM)</td>
<td>99.7</td>
<td>97.1</td>
<td>97.1</td>
<td>98</td>
<td>59.4</td>
<td>50</td>
<td>62.5</td>
<td>57.3</td>
</tr>
<tr>
<td>Hofmann et al</td>
<td>99.4</td>
<td>27.1</td>
<td>52.6</td>
<td>59.7</td>
<td>44</td>
<td>6</td>
<td>9</td>
<td>19.7</td>
</tr>
<tr>
<td>RSM</td>
<td>100</td>
<td>79</td>
<td>97</td>
<td>92</td>
<td>58</td>
<td>38</td>
<td>57</td>
<td>51.3</td>
</tr>
<tr>
<td><strong>Castro17a-PFM</strong></td>
<td>99.7</td>
<td>99</td>
<td>99</td>
<td><strong>99.2</strong></td>
<td>78.1</td>
<td>56.3</td>
<td>46.9</td>
<td><strong>60.4</strong></td>
</tr>
</tbody>
</table>

Ours-1: id+age0.2 (vB)
Ours-2: id+verif0.1 (vB)
7-NN: 7-Nearest Neighbour with PCA-128


Experiment: gender

- Gender recognition results

Accuracy: higher is better
Experiment: age

• Age estimation results

Mean Absolute Error: lower is better
Experiment: verification

- Identity verification results

EP: accuracy at equilibrium point
AUC: area under the precision-recall curve
Outline

1. Problem definition

2. Our approach
   i. Input data
   ii. Deep Multi-task Model

3. Experiments and results

4. Conclusions and future work
Conclusions and future work

• DMT speeds up convergence of the main task.

• Accuracy of identification (main task) improves.

• CNN filters of the first layers are useful for several tasks.

• Other modalities: gray, depth,...

• Other tasks
THANK YOU
FOR YOUR ATTENTION

Video

OF-x

OF-y

Manuel Jesús Marín Jiménez  @mjmarinj