Online Open World Face Recognition From Video Streams
ID:23202

Federico Pernici, Federico Bartoli, Matteo Bruni and Alberto Del Bimbo
MICC - University of Florence - Italy
http://www.micc.unifi.it
The effectiveness of data in Deep Learning

- Performance increases linearly with orders of magnitude of training data [Chen2017].
However...

- Linear improvement in performance requires **exponential number** of labelled examples.

[Sun2017: Revisiting the Unreasonable Effectiveness of Data ICCV2017]
The cost of annotation

• The **cost of annotation** remains the most critical fact in Supervised Learning.

• Crowdsourcing...
  
  • 1M images with 1000 categories at 1 cent per question $10M.
  
  • ImageNet used several heuristics (e.g., hierarchy of labels) to reduce the space of questions, reducing the cost to the order of $100K
Learning from video streams

An attracting alternative:

• learn objects appearance from **video streams** with no supervision, both exploiting
  • the **large quantity** of video available in the Internet and
  • the fact that **adjacent video frames** contain **semantically similar** information (weak supervision).
Practical Problem...

- **Online Open World** Face Recognition from video streams
  - It is not possible to predict a priori how many face objects to recognize (i.e. the number of **classes is unknown**).
  - The system must be able to **detect known/unknown** classes.
  - There are no labels.
  - The system must be able to **add** the detected **unknown classes** to the model (Open World).
  - The system cannot be retrained from scratch (it must be works forever).

- The problem appears to present a daunting challenge for deep learning (**catastrophic forgetting**).
Problem details...

• New face identities...
• Wrong identity associations...
• False positives... (not a novel class)

Unconstrained videos are typically made of shots
Problem details

• The Learner operates in two steps.
  • First, it automatically labels the data in the next frame.
  • Second, it uses this labeled data to train the classifier.

• Errors may introduce noisy labels (wrong identities).

• **Noisy labels** may impair irreversibly the learning process as time advance.
Our solution: exploit a Memory module

- The appearance in video streams typically evolves over time:
  - **Data** can no longer be assumed as independent and identically distributed (i.i.d.)

- **Store** the past experience in a **memory** module (i.e. Hippocampus) [Schaul2015].
  - If appearances are never forgotten (Infinite Memory), it is possible to limit the non stationary effects [Cornuéjols2006].
  - This also makes it possible to mix more and less recent information.

[Schaul2015: Prioritized Experience Replay]
System Overview

• Main components:
  • Face detection (GPU)
  • Descriptor extraction (GPU)
  • Matching (GPU)
  • Memory (GPU)
  • Memory Controller
Face Detection and Description

• Faces are **detected** using the Tiny Faces method [Peiyun2017]
  • The method uses a CNN with the ResNet101 architecture

• Detected faces are **represented** according CNN activations (the face descriptor) extracted from the VGGface CNN [Parkhi2015]
Main Idea: quick learning using Memory

• The memory module is used for fast learning and consists of the following triples:

\[ \mathcal{M}(t) = \left\{ (x_i, \text{Id}_i, e_i) \right\}_{i=1}^{N(t)} \]

• The **eligibility** \( e_i \) is a scalar quantity in \([0,1]\) associated to each descriptor \( x_i \) (i.e. CNN activations)
  • It captures the **redundancy** of a descriptor with respect to the other descriptors in the memory.
  • Each descriptor has an associated identity \( \text{Id}_i \).
Intuition: Memory and Eligibilities

• Faces appearance model is extended using the video **exemplars** collected while tracking.

• To control **redundancy** the eligibilities $e_i$ of matching descriptors are time updated according to:
  \[ e_i(t + 1) = \eta_i e_i(t) \]
  where $\eta_i$ take into account descriptor distance (i.e. spatial redundancy).

• Descriptors are **removed** when their corresponding eligibilities $e_i$ drops below a given threshold.

• The eligibility is:
  • Low for **ordinary** «events»
  • High for **rare** «events»

• **Unmatched** descriptors are **added** to the memory with a novel Id and $e=1$. 

**Appearance Learned Offline (i.e. VggFace Deep Learning)**

The extended appearance learned from video

**Video data exemplars**
Discriminative Matching

• **Video temporal coherence:**
  • Faces in consecutive frames have little differences.
  • Similar descriptors will be stored in the memory (Repeated Temporal Structure).

• **Distance Ratio test:** compares the distance to the closest neighbor with the distance to the second closest neighbor.
  • If they are far apart (d1/d2<thresh): OK.

• If repeated structure distances are comparable, the discriminative match cannot be assessed.
  • This limit is solved using **Reverse Nearest Neighbor (ReNN)**
Reverse Nearest Neighbour (ReNN)

- In ReNN **Roles** are **exchanged**
  - Each entry of the database is a query.
  - Faces in the current frame are the database.
ReNN and distance ratio

- This strategy exploits discriminatively the uniqueness of face in the current frame.

- The other important advantage ReNN is that all the descriptors $x_i$ of the repeated structure match with $o_1$:

  \[ \{o_1\} \leftrightarrow \{x_i\} \]

- This allows the automatic selection of the descriptors that need to be condensed into a more compact representation.
GPU based ReNN

• Reverse Nearest Neighbor under the distance ratio criterion can be **effectively accelerated** on the GPU.

• This is achieved using the *min* function twice in a GPUarray (Matlab, PyCuda).
  - Cuda Parallel Reduction is exploited.

• Complexity is almost constant as the number of descriptors in the memory increases (Nvidia Titan X Maxwell).
Asymptotic Stability

• Eligibility updating stabilizes around the pdf of each individual subject face.

• The eligibility updating rule:

\[ e_i(t + 1) = \eta_i e_i(t) \]

is a contraction (i.e. \( \eta_i < 1 \)), it converges to its unique fixed point.

• Toy problem with increasing difficulty...
Experimental Results

• We used the **Music-dataset** [Zhang2016].
• **8 music videos** downloaded from YouTube with annotations of 3,845 face tracks

• **Big Ban Theory** 1° season (Ep1,2,...,6).
• 6 videos, about 23 minutes each.
Experimental Results: drifting analysis

- **Ground Truth as detections**
- **Accuracy:**
  \[
  \text{MOTA} = 1 - \frac{\sum_i (\text{FN}_i + \text{FP}_i + \text{IDS}_i)}{\sum_i \text{GT}_i}
  \]
- **Fluctuations:** no information at the beginning.
- **Stability** is common to all the videos.
**Experimental Results: drifting analysis**

- **Ground Truth as detections**
- **Accuracy:**

\[
\text{MOTA} = 1 - \frac{\sum_i (\text{FN}_i + \text{FP}_i + \text{IDS}_i)}{\sum_i \text{GT}_i}
\]

- **Fluctuations:** no information at the beginning.
- **Stability is common to all the videos.**
**Comparison with Offline Methods**

Scores are based on Purity. Purity is a measure of the extent to which clusters contain a single class.

### MUSIC DATASET

<table>
<thead>
<tr>
<th>Videos</th>
<th>Apink</th>
<th>Bruno Mars</th>
<th>Darling</th>
<th>Girls Aloud</th>
<th>Hello Bubble</th>
<th>Pussycat Dolls</th>
<th>T-ara</th>
<th>Westlife</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>0.20</td>
<td>0.36</td>
<td>0.19</td>
<td>0.29</td>
<td>0.35</td>
<td>0.28</td>
<td>0.22</td>
<td>0.27</td>
</tr>
<tr>
<td>AlexNet</td>
<td>0.22</td>
<td>0.36</td>
<td>0.18</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.25</td>
<td>0.37</td>
</tr>
<tr>
<td>Pre-trained</td>
<td>0.29</td>
<td>0.50</td>
<td>0.24</td>
<td>0.33</td>
<td>0.34</td>
<td>0.31</td>
<td>0.31</td>
<td>0.37</td>
</tr>
<tr>
<td>VGG-Face</td>
<td>0.24</td>
<td>0.44</td>
<td>0.20</td>
<td>0.31</td>
<td>0.29</td>
<td>0.46</td>
<td>0.23</td>
<td>0.27</td>
</tr>
<tr>
<td>Siamese</td>
<td>0.48</td>
<td>0.88</td>
<td>0.46</td>
<td>0.67</td>
<td>0.54</td>
<td>0.77</td>
<td>0.69</td>
<td>0.54</td>
</tr>
<tr>
<td>Triplet</td>
<td>0.60</td>
<td>0.83</td>
<td>0.49</td>
<td>0.67</td>
<td>0.60</td>
<td>0.77</td>
<td>0.68</td>
<td>0.52</td>
</tr>
<tr>
<td>SymTriplet</td>
<td><strong>0.72</strong></td>
<td>0.90</td>
<td>0.70</td>
<td>0.69</td>
<td><strong>0.64</strong></td>
<td>0.78</td>
<td>0.69</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>MuFTiR-tiny</strong></td>
<td>0.51</td>
<td><strong>0.96</strong></td>
<td><strong>0.73</strong></td>
<td><strong>0.89</strong></td>
<td>0.59</td>
<td><strong>0.97</strong></td>
<td><strong>0.72</strong></td>
<td><strong>0.98</strong></td>
</tr>
</tbody>
</table>

### BIG BANG THEORY

<table>
<thead>
<tr>
<th>Episodes</th>
<th>BBT01</th>
<th>BBT02</th>
<th>BBT03</th>
<th>BBT04</th>
<th>BBT05</th>
<th>BBT06</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG</td>
<td>0.37</td>
<td>0.32</td>
<td>0.38</td>
<td>0.35</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>AlexNet</td>
<td>0.47</td>
<td>0.32</td>
<td>0.45</td>
<td>0.35</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>Pre-trained</td>
<td>0.62</td>
<td>0.72</td>
<td>0.73</td>
<td>0.57</td>
<td>0.52</td>
<td>0.52</td>
</tr>
<tr>
<td>VGG-Face</td>
<td>0.91</td>
<td>0.85</td>
<td>0.83</td>
<td>0.54</td>
<td>0.65</td>
<td>0.46</td>
</tr>
<tr>
<td>Siamese</td>
<td>0.94</td>
<td>0.95</td>
<td>0.87</td>
<td>0.74</td>
<td>0.70</td>
<td>0.70</td>
</tr>
<tr>
<td>Triplet</td>
<td>0.94</td>
<td>0.95</td>
<td>0.92</td>
<td>0.74</td>
<td>0.68</td>
<td>0.70</td>
</tr>
<tr>
<td>SymTriplet</td>
<td>0.94</td>
<td>0.95</td>
<td>0.92</td>
<td>0.78</td>
<td>0.85</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>MuFTiR-tiny</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.85</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.94</strong></td>
</tr>
</tbody>
</table>
## Comparison with Offline Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Mode</th>
<th>IDS ↓</th>
<th>MOTA ↑</th>
<th>MOTP ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>mTLD [1]</td>
<td>Offline</td>
<td>1</td>
<td>-16.3</td>
<td>74.8</td>
</tr>
<tr>
<td>ADMM [2]</td>
<td>Offline</td>
<td>323</td>
<td>42.5</td>
<td>64.0</td>
</tr>
<tr>
<td>IHTLS [3]</td>
<td>Offline</td>
<td>312</td>
<td>45.7</td>
<td>64.0</td>
</tr>
<tr>
<td>Pre-Trained [4]</td>
<td>Offline</td>
<td>171</td>
<td>41.9</td>
<td>73.3</td>
</tr>
<tr>
<td>mTLD2 [1]</td>
<td>Offline</td>
<td>223</td>
<td>58.4</td>
<td>73.8</td>
</tr>
<tr>
<td>Siamese [4]</td>
<td>Offline</td>
<td>144</td>
<td>69.0</td>
<td>73.7</td>
</tr>
<tr>
<td>Triplet [4]</td>
<td>Offline</td>
<td>164</td>
<td>69.3</td>
<td>73.6</td>
</tr>
<tr>
<td>SymTriplet [4]</td>
<td>Offline</td>
<td>156</td>
<td>72.2</td>
<td>73.7</td>
</tr>
<tr>
<td>MuFTiR-tiny</td>
<td>Online</td>
<td>36</td>
<td>51.6</td>
<td>70.1</td>
</tr>
<tr>
<td>MuFTiR-tiny</td>
<td>Offline</td>
<td>34</td>
<td>51.6</td>
<td>70.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Mode</th>
<th>IDS ↓</th>
<th>MOTA ↑</th>
<th>MOTP ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>mTLD [1]</td>
<td>Offline</td>
<td>35</td>
<td>-8.7</td>
<td>65.3</td>
</tr>
<tr>
<td>ADMM [2]</td>
<td>Offline</td>
<td>428</td>
<td>50.6</td>
<td>85.7</td>
</tr>
<tr>
<td>IHTLS [3]</td>
<td>Offline</td>
<td>375</td>
<td>52.7</td>
<td>85.8</td>
</tr>
<tr>
<td>Pre-Trained [4]</td>
<td>Offline</td>
<td>151</td>
<td>48.3</td>
<td>88.0</td>
</tr>
<tr>
<td>mTLD2 [1]</td>
<td>Offline</td>
<td>278</td>
<td>52.6</td>
<td>87.9</td>
</tr>
<tr>
<td>Siamese [4]</td>
<td>Offline</td>
<td>126</td>
<td>56.7</td>
<td>87.8</td>
</tr>
<tr>
<td>Triplet [4]</td>
<td>Offline</td>
<td>126</td>
<td>56.6</td>
<td>87.8</td>
</tr>
<tr>
<td>SymTriplet [4]</td>
<td>Offline</td>
<td>105</td>
<td>56.8</td>
<td>87.8</td>
</tr>
<tr>
<td>MuFTiR-dpm</td>
<td>Online</td>
<td>78</td>
<td>4.5</td>
<td>61</td>
</tr>
<tr>
<td>MuFTiR-tiny</td>
<td>Online</td>
<td>420</td>
<td>48.8</td>
<td>65.5</td>
</tr>
</tbody>
</table>

---

**Not too far, but online**


Online Open World Face Recognition From Video Streams

Link: https://youtu.be/6S7D6Dgmt3Y
Qualitative results
Conclusion

• Online Open World Face Recognition From Video Streams
  • Fully implemented on a GPU
  • Wide applicability: Enables face recognition with auto enrollment of subjects

• Applicability in other contexts:
  • Person Detector – Person Descriptor
  • Car detector – Car Descriptor
  • Traffic Signal Detector – Traffic Signal Descriptor
  • …

• Future developments:
  • Exploit the data diversity in the memory to train online a Deep CNN.