Fisher Vector Faces applied to face tracks:

- Robust to face misalignments
- A single descriptor per track
- Compact: low dimensional & binarised

Compact face track descriptor: a novel Video Fisher Vector Faces:

- Contributions
  - Video Fisher Vector Faces: a novel, discriminative, efficient, and very compact face track descriptor
  - Robust to face misalignments
  - A single descriptor per track
  - Compact: low dimensional & binarised

Video Fisher Vector Faces (VFV)

Fisher Vector Faces applied to face tracks:
- **Video pooling**: one easy-to-use descriptor per track
- **Jittered pooling**: efficient data augmentation
- **Binarisation**: extreme compression
- **Hard-assignment fisher vector**: 6 times faster

A powerful single-frame face descriptor:
- Dense sampling of local descriptors (SIFT)
- Fisher Vector encoding
  - Gaussian Mixture Model codebook
  - First and second order statistics
- Discriminative low-rank Mahalanobis metric

Learn to Compare & Compress

Objective Function and Learning

\[
\min_{V, W} \sum_{i,j} \max \left[ 1 - y_{ij}(b - d_{i,j}^W(\phi_i, \phi_j)) \right] \quad \text{label} \uparrow \text{bias}
\]

Non-convex functions optimized using SGD. Large reduction in dimensionality without performance loss (64K → 128).

Binarisation

Goal: further reduce memory footprint.

Method: Parseval Tight Frame Expansion

1. Start with \( m \)-dimensional descriptors \( \phi \)
2. Sample a random \( n \times n \) matrix \( M \) with \( n > m \)
3. Decompose \( M = QR \)
4. \( U \leftarrow \) first \( m \) columns of \( Q \)
5. Binarisation \( \text{sign}(U \phi) \) has \( q \) bits only

Typical use case: compress 128-D float descriptors (4096 bit) down to 1024 bits without accuracy loss (4x reduction).

Experiments

- Excellent performance with small training sets
- Cross-task and cross-dataset transfer

Face Verification on YouTube Faces

- Restricted: train on only pre-specified pairs
- Unrestricted: use any pair

Parameter Tuning

Comparison with the State of the Art

<table>
<thead>
<tr>
<th>Method</th>
<th>Feat. Dim.</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinbis et al.</td>
<td>-</td>
<td>42.50</td>
</tr>
<tr>
<td>Our VF</td>
<td></td>
<td>30.11</td>
</tr>
<tr>
<td>Cinbis et al. (trained on LFW)</td>
<td>-</td>
<td>36.20</td>
</tr>
<tr>
<td>Our VF &amp; Flip</td>
<td>128 x 2</td>
<td>25.77</td>
</tr>
<tr>
<td>Our VF binar. 2048 bit + flip</td>
<td>128</td>
<td>21.90</td>
</tr>
</tbody>
</table>

Face Classification on Oxford Buffy

- 7 episodes from season 5 of “Buffy the Vampire Slayer”
  - Training data obtained from alignment of transcripts and subtitles

<table>
<thead>
<tr>
<th>Method</th>
<th>Proj. Dim.</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM &amp; Proj-n train set.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cinvic</td>
<td>1024</td>
<td>none</td>
</tr>
<tr>
<td>Sivic</td>
<td>1024</td>
<td>none</td>
</tr>
<tr>
<td>Sivic</td>
<td>1024</td>
<td>none</td>
</tr>
<tr>
<td>Buffy</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Youtube Faces</td>
<td>none</td>
<td>0.86</td>
</tr>
<tr>
<td>Youtube Faces + jitt.</td>
<td>1024</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Acknowledgement: This work was supported by ERC grant VisRec no. 228180 and EU Project AXES ICT-269980