Egocentric Height Estimation

Jessie Finocchiaro
Aisha Urooj Khan
Dr. Ali Borji
Compact 3D CNN

- Used Shmuel Poleg's CNN and made it regression based*
- Made last layer have one output and ReLU activation

Network input and output

- **Input:** Sparse Optical Flow Vectors
  - 60 frames concatenated for x and y directions
  - 32x32x120
  - Typically 10-20 of these inputs per video

- **Output**
  - A continuous measurement for each 60 frame segment of the video.
  - Take the average of results to get estimate for entire video
Compact 3D CNN

- Normalized data between 0 and 1
- Tried training network with learning rates of 0.01, 0.001, 0.0001

<table>
<thead>
<tr>
<th></th>
<th>Actual</th>
<th>Network</th>
<th>Retrained; decreased LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chest</td>
<td>121</td>
<td>121.087</td>
<td>164.48</td>
</tr>
<tr>
<td>Waist</td>
<td>99</td>
<td>92.96</td>
<td>90.7</td>
</tr>
<tr>
<td>Head</td>
<td>166</td>
<td>174.56</td>
<td>166.457</td>
</tr>
<tr>
<td>Loss</td>
<td>----</td>
<td>----</td>
<td>.1058</td>
</tr>
</tbody>
</table>
Leave One Out training-Temporal network

- 10 subjects
- Train on 9, test on 1
- $R^2 = .4557$
- MAE = 18.03 cm
Spatial CNN

- Use same CNN, but instead of inputting optical flow vectors, input is pixel intensities
- Condense input from 120 frames to 60 frames
Spatial CNN Results- LOO

- \( R^2 = 0.4267 \)
- \( MAE = 19.51 \text{ cm} \)
Two-stream network

- Concatenate outputs of two networks
- Connect them
- Give one “mega output”

\[
R^2 = 0.5773 \quad \text{MAE} = 14.53 \text{ cm}
\]

\[
R^2 = 0.6505 \quad \text{MAE} = 14.04 \text{ cm}
\]
Current work

- Person identification from a small dataset
  - Try to improve the results of Peleg's paper *An Egocentric Look at Video Photographer Identity*