Egocentric Video Understanding

Sagar Verma
(Junior Research Fellow, IIIT Delhi)
Overview

- **Egocentric:** First-person view
- **Wearable devices**
  - GoPro, Google Glass, Hololens.
- **Applications**
  - Life logging, augmented reality, assistive vision, etc.
- **100 hours’ of videos every minute on Youtube.**
  - Generalized video understanding method is required for easy retrieval, generating metadata, etc.
Overview

Skydive logging using GoPro

Google Glass being used by policemen

Hololens for augmented reality

Mobility assistance using Pivothead

Fig 1: Some example applications of egocentric cameras
First-person action recognition

● One method to recognize all action categories.
  ○ Actions involving hand-object interactions.
  ○ Actions involving no hand-object interactions.
    ■ ‘walk’, ‘run’, etc.
  ○ Short term actions.
    ■ ‘fold’, ‘put’, etc.
  ○ Long term actions.
    ■ ‘spread’, ‘stir’, ‘driving’, etc.
Fig 2: Sample actions from two action categories; top row show actions where some form of hand-object interaction is present, and bottom row shows actions without any hand-object interaction.
Challenges and Proposed Solutions

● Unlike third-person, first person videos do not provide full pose of the actor.
  ○ Important cues are hand motion, handled object attributes and camera ego-motion.

● Two stream CNN-LSTM for complementary information
  ○ Rgb frames for hand and object interactions.
    ■ For better detection, resize object size to match Imagenet’s.
    ■ 8% accuracy increase by resizing alone.
  ○ Optical flow for camera ego-motion.
Challenges and Proposed Solutions

- Actions like ‘open’ and ‘close’ which are “reverse” of each other confuse CNN for rgb frames.
- Use curriculum learning
  - Merge actions which are “reverse” of each other while training CNN on rgb.
  - Use all actions when training LSTM and CNN on flows.
  - 2% increase in CNN on rgb frames by curriculum learning.
Proposed Architecture

Fig 6: Proposed Architecture
## Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Subjects</th>
<th>Frames</th>
<th>Classes</th>
<th>Current</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEA</td>
<td>4</td>
<td>31,253</td>
<td>11</td>
<td>68.50</td>
<td>82.71</td>
</tr>
<tr>
<td>EGTEA+</td>
<td>32</td>
<td>1,055,937</td>
<td>19</td>
<td>NA</td>
<td>66</td>
</tr>
<tr>
<td>Kitchen</td>
<td>7</td>
<td>48,117</td>
<td>29</td>
<td>66.23</td>
<td>71.92</td>
</tr>
<tr>
<td>ADL</td>
<td>5</td>
<td>93,293</td>
<td>21</td>
<td>37.58</td>
<td>44.13</td>
</tr>
<tr>
<td>UTE</td>
<td>2</td>
<td>208,230</td>
<td>21</td>
<td>60.17</td>
<td>65.12</td>
</tr>
<tr>
<td>HUJI</td>
<td>NA</td>
<td>1,338,606</td>
<td>14</td>
<td>86</td>
<td>93.92</td>
</tr>
</tbody>
</table>

Table 1: Accuracy comparison of our method with SOTA
Fig 7: Since most datasets contain only one kind of action category, GTEA and HUJI are mixed to validate our hypothesis. Accuracy obtained is 86.85%.
Recognizing activities from sequence of actions

- An activity comprises of hundreds of actions.
- Can we recognize an activity from the sequence of actions?
  - E.g. Activity ‘Making Tea’ comprises of actions ‘take’ bowl -> ‘pour’ water -> ‘put’ bowl on stove -> ....
Challenges and Proposed Solutions

- An activity can span 100s of actions, 10-50k frames.
  - For sequence length > 1000, vanishing and exploding gradient.
- Linear transformation to Hadamard operation on hidden vector of LSTM/RNN
  - $h_t = \sigma(Wx_t + Uh_{t-1} + b) \Rightarrow h_t = \sigma(Wx_t + u \odot h_{t-1} + b)$
  - Clipping gradients w.r.t. to max seq length.
Challenges and Proposed Solutions

- How to differentiate between two adjacent same action labels.
  - E.g. ‘take’ screwdriver → ‘take’ screw.
- Use CTC beam search to differentiate between two consecutive same action labels.
Challenges and Proposed Solutions

- Current method of late fusion of RGB and flow temporal features is not good enough.
- A temporal-modality attention based fusion method.
  - Attention parameter, \( \alpha = h_k^{rgb}W_{\text{att}} + h_k^{flow}W_{\text{att}} \) where \( h_k^{rgb/flow} \) are hidden outputs of rgb and flow RNN at \( k^{th} \) time step, \( W_{\text{att}} \) is learnable parameter.
  - \( h_k^{\text{combined}} = \alpha h_k^{rgb} + (1-\alpha) h_k^{flow} \)
Proposed Architecture

Fig 8: Proposed Architecture
## Results

Table 2: Accuracy of our method on different egocentric activities datasets. (NA for future experiments)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Subjects</th>
<th>Sequence Length</th>
<th>Actions</th>
<th>Activities</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEA</td>
<td>4</td>
<td>10</td>
<td>11</td>
<td>7</td>
<td>89.28%</td>
</tr>
<tr>
<td>EGTEA+</td>
<td>32</td>
<td>86</td>
<td>19</td>
<td>7</td>
<td>NA</td>
</tr>
<tr>
<td>EPIC</td>
<td>32</td>
<td>329</td>
<td>125</td>
<td>9</td>
<td>NA</td>
</tr>
<tr>
<td>IIITD Plumbing</td>
<td>10</td>
<td>5</td>
<td>12</td>
<td>10</td>
<td>NA</td>
</tr>
<tr>
<td>IIITD PC</td>
<td>6</td>
<td>42</td>
<td>11</td>
<td>2</td>
<td>NA</td>
</tr>
</tbody>
</table>
Conclusion

- We have a generalized method for recognizing different types of action categories.
- From action sequences we can recognize activities.
Future Work

● Can we build a daily life log from actions, activities and social interactions of a person?
  ○ Actions like ‘handshake’, ‘hug’, etc can be used to define social interactions.

● To recognize a person we can use unsupervised way of face matching.
Dissemination of Research Results


*Title might get changed.
References

Questions?

Thank, You!