DNN for action recognition in videos

Omar Seddati, Stéphane Dupont, Said Mahmoudi
TCTS Lab, University of Mons, Mons, Belgium

Context
- action recognition in videos
- Moving pictures contain two types of information, the first relating to appearance and the second to the dynamics of motion.

Database
Different types of actions: Human Object Interaction, Body Motion Only, Human-Human Interaction, etc.
- HMDB51: 6849 clips divided into 51 action categories.
- UCF101: 13320 videos from 101 action categories.

Approach
- Our system is based on the use of ConvNets, a state-of-the-art technique in the field of object recognition within images.
- ConvNets applied to 2D images enable capture spatial configurations.
- Dense optical flow maps represent the displacement of each pixel between two successive frames. Hence, local temporal information is projected to a space similar to pixel space, and ConvNets can be effectively used for dynamic information.

Spatial DNN
Using Convolutional neural networks allow us to extract low-level features (shared among categories) and high-level features (more global and more invariant).

Temporal DNN
- 32M parameters.
- 1.4M neurons.

Results
Results achieved using two optical flow techniques (previous state-of-the-art 46.6%):

<table>
<thead>
<tr>
<th>Method</th>
<th>Maximin</th>
<th>Mean</th>
<th>Maxmin</th>
<th>M.V.</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV-L1</td>
<td>51.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brox</td>
<td>46.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Accuracy using different decision fusion techniques:
- Maximin: 44.38
- Mean: 49.73
- Maxmin: 43.40
- M.V.: 48.36
- Product: 51.24