Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction

Jianfei Yang, Kai Wang, Xiaojiang Peng and Yu Qiao
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
yang0478@ntu.edu.sg

Contributions
- We developed a deep multi-instance learning framework that accepts multiple input features, and evaluated how different modality performs using our framework.
- Furthermore, to take full advantage of all available data, we make new data split for model ensemble.
- Experimental results demonstrate the effectiveness of our method, and we eventually win the challenge with MSE of 0.0626.

Introduction of Our Approaches
Multiple Instance Learning Framework
- We formulate the problem as a multi-instance regression. A video sequence $v$ is divided as $k$ segments such as $v = [s_1, s_2, s_3, ..., s_k]$, and each video clip is regarded as an instance. We extract $M$ different modality features $F_i = [f_1, f_2, f_3, ..., f_M]$ from a segment and feed them into our framework.

Multi-modal Features
- LBP-TOP features are extracted for each video clip, which is used as fine-grained facial feature in our approach.
- We use the C3D network pretrained in Sports-1M dataset, and crop the subject body using OpenPose. Then C3D features are extracted by body images in a segment.
- We capture the gaze and head movement features using OpenFace while body posture characteristics via OpenPose.

Dataset Split and Model Ensemble
- We generate new data splits by utilizing all validation data for training. We make the training class balanced as much as possible.

MSE Results on Validation set of official split

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Normalized MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 LSTM + OpenFace</td>
<td>0.0847</td>
<td>0.0821</td>
</tr>
<tr>
<td>1 LSTM + OpenFace</td>
<td>0.0853</td>
<td>0.0830</td>
</tr>
<tr>
<td>1 LSTM + OpenPose</td>
<td>0.0717</td>
<td>0.0739</td>
</tr>
<tr>
<td>2 LSTM + OpenPose</td>
<td>0.0734</td>
<td>0.0732</td>
</tr>
<tr>
<td>1 LSTM + LBP-TOP</td>
<td>0.0969</td>
<td>-</td>
</tr>
<tr>
<td>1 LSTM + C3D</td>
<td>0.0865</td>
<td>-</td>
</tr>
</tbody>
</table>

In the Table 1, the model using facial features gets MSE of 0.085 around and the one using 2 LSTM layers generates a little better result. As for the model using posture features, it outperforms the face-based one by 0.01 generally, which is a powerful proof that our OpenPose features can contribute more to the engagement intensity prediction.

Our Pipeline

Video -> Segments -> Feature Extraction
- OpenFace
- OpenPose
- LBP-TOP
- C3D

Regression Network
- LSTM+FC
- LSTM+FC
- LSTM+FC
- LSTM+FC

Modality Consensus
- Face Regression
- Pose Regression
- LBP-TOP Regression
- C3D Regression

Figure 1: The system pipeline of our approach.

Our Submissions
- OpenPose features only in new split.
- OpenPose features using all data.
- OpenFace and OpenPose features in new split.
- OpenFace and OpenPose features in official split.
- (3) + LBP-TOP + C3D
- (3) + (4)
- OpenFace in new split + LBP + C3D

Conclusions
- We presented our approach in this paper for the engagement intensity prediction in the Emotion Recognition in the Wild Challenge 2018.
- We developed a deep multi-instance learning framework that accepts multiple input features, and evaluated how different modality performs using our framework.
- Statistical feature, local descriptor and deep representation are employed and they compensate for each other. Furthermore, to take full advantage of all available data, we make new data split for model ensemble.