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 di□erent 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 ne□ectiveness 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 di□erent modality features $F_k = [f_k^1, f_k^2, f_k^3, ..., f_k^m]$ 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.

Our Pipeline

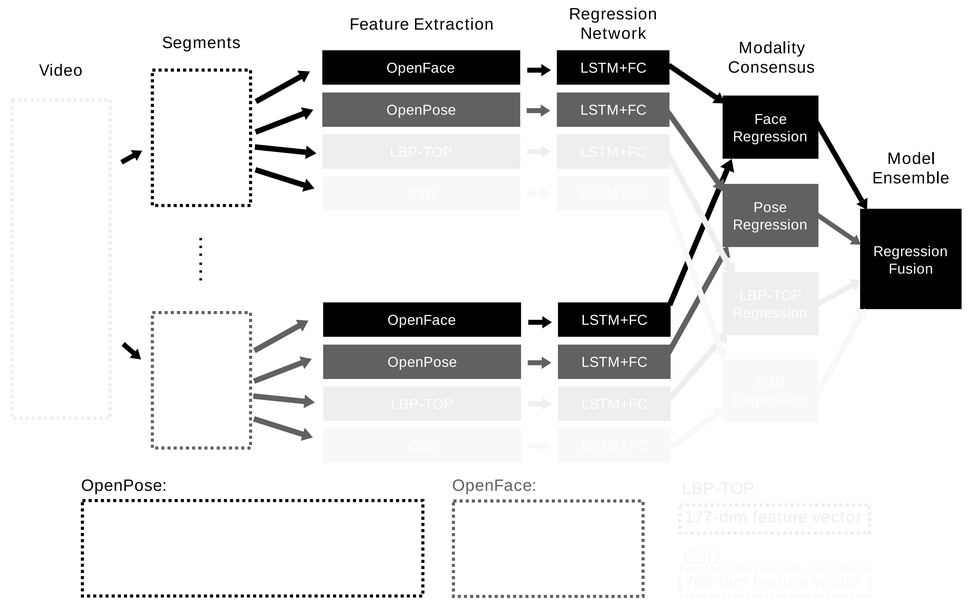

Figure 1: The system pipeline of our approach.

MSE Results on Validation set of o□cial split

Table 1: MSE Results on Validation set of o□cial split

Method	MSE	Normalized MSE
2 LSTM + OpenFace	0.0847	0.0821
1 LSTM + OpenFace	0.0853	0.0830
1 LSTM + OpenPose	**0.0717**	0.0739
2 LSTM + OpenPose	0.0734	0.0732
1 LSTM + LBP-TOP	0.0909	-
1 LSTM + C3D	0.0865	-

- 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.

MSE Results on Validation set of new split

Table 2: MSE Results on Validation set of new split

Method	MSE	Normalized MSE
2 LSTM + OpenFace	**0.0398**	0.0410
2 LSTM + OpenPose	0.0853	0.0830
1 LSTM + OpenPose	0.0671	0.0717

- In Table 2, it is amazing that our new split reduces the MSE. Our face-based approach on new split achieves 0.0398 MSE and pose-based one attains best MSE of 0.0671. Our adjustment of quantitative balance leads to better convergence, and moreover, the supplement of o□cial validation data enhances the generalization of our approach, which blossoms a lot in the ensemble with the models optimized by o□cial data.

Our Submissions

1. OpenPose features only in new split.
2. OpenPose features using all data.
3. OpenFace and OpenPose features in new split.
4. OpenFace and OpenPose features in o□cial split.
5. (3) + LBP-TOP + C3D
6. (3) + (4)
7. OpenFace in new split + LBP + C3D

Table 3: MSE Results of all models of submissions.

Runs	Validation ($\times 10^{-4}$)	Test ($\times 10^{-4}$)				
	Overall	NE	BE	E	SE	Overall
1	**853**	**2180**	917	540	4407	1040
2	123	3854	1169	753	2210	1353
3	364	2505	**473**	**154**	1628	**626**
4	734	3185	1076	508	1477	1072
5	541	2781	541	198	1443	698
6	782	2987	725	174	**988**	745
7	431	2275	589	257	2676	730

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 di□erent 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.

中国科学院深圳先进技术研究院
SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY
CHINESE ACADEMY OF SCIENCES