Engagement Intensity Prediction with Facial Behavior Features

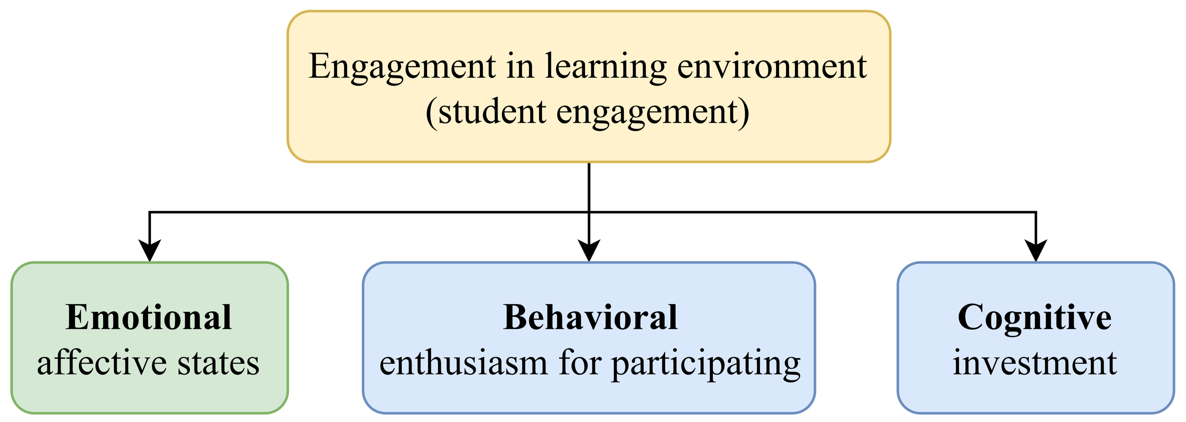
Van Thong Huynh, Hyung-Jeong Yang, Guee-Sang Lee, Soo-Hyung Kim*

School of Electronics and Computer Engineering, Chonnam National University

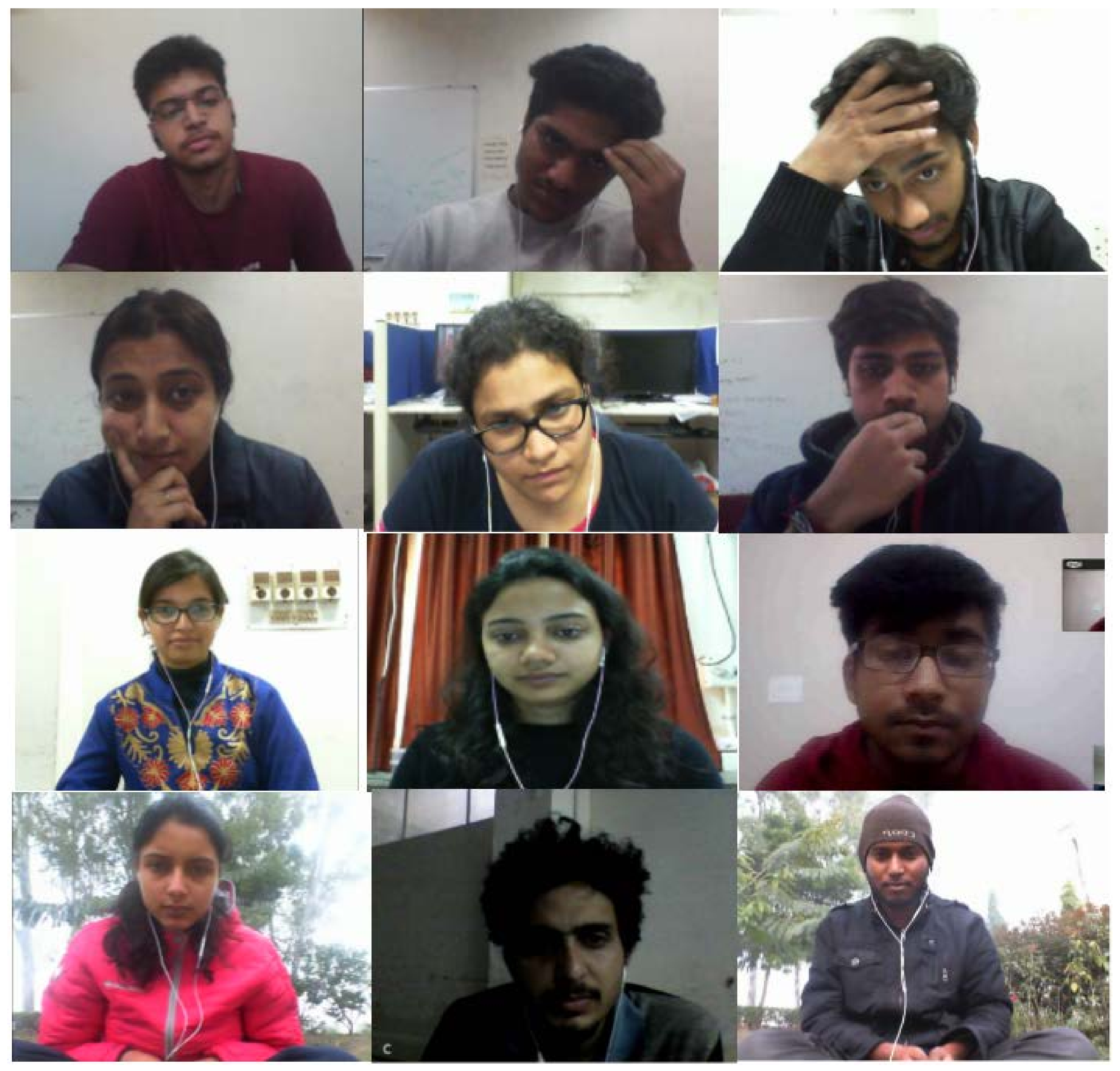
ICMI 2019

SUZHOU 苏州

Introduction



- Engagement prediction in the Wild (EW), a sub challenges in the 7th Emotion Recognition in the Wild 2019 Grand Challenge (EmotiW), *predicts the engagement intensity* of a subject in a video which recorder while the subject is watching and educational video (MOOC) [1].



- Kaur et al. [2] described EW dataset with some examples of frames of the video as the above figure, top to bottom rows show engagement intensity level: [0 (low) - 3 (high)].
- Our method achieved the best result, a mean square error of 0.0597, with three fundamental steps:
  1. Feature Extraction.
  2. Predict the engagement intensity for each type of feature with two different models.
  3. Fusion the results of each type.

Contact Information

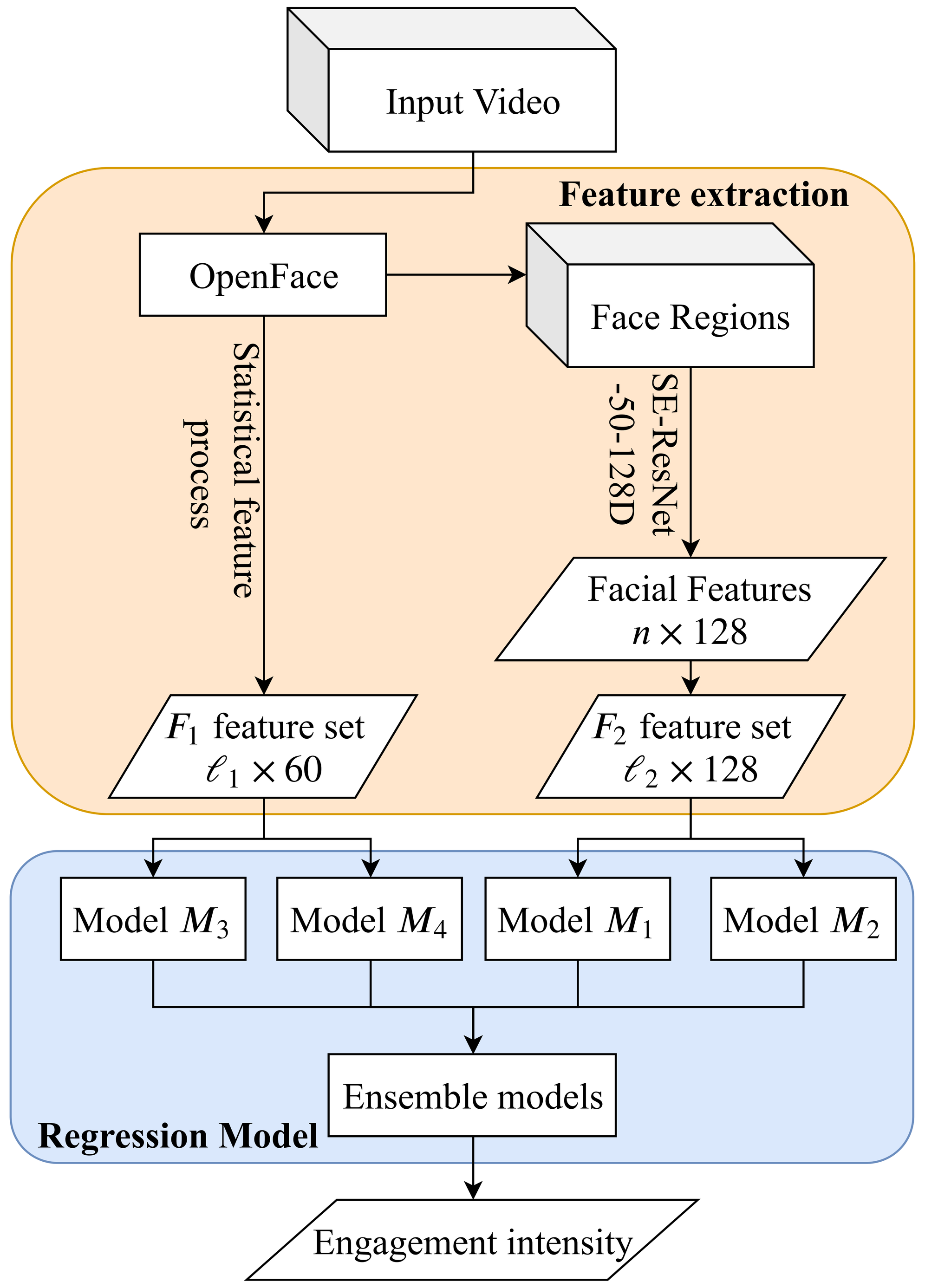
hvthong.298@outlook.com.vn

https://github.com/th2l/SML_EW_EmotiW2019

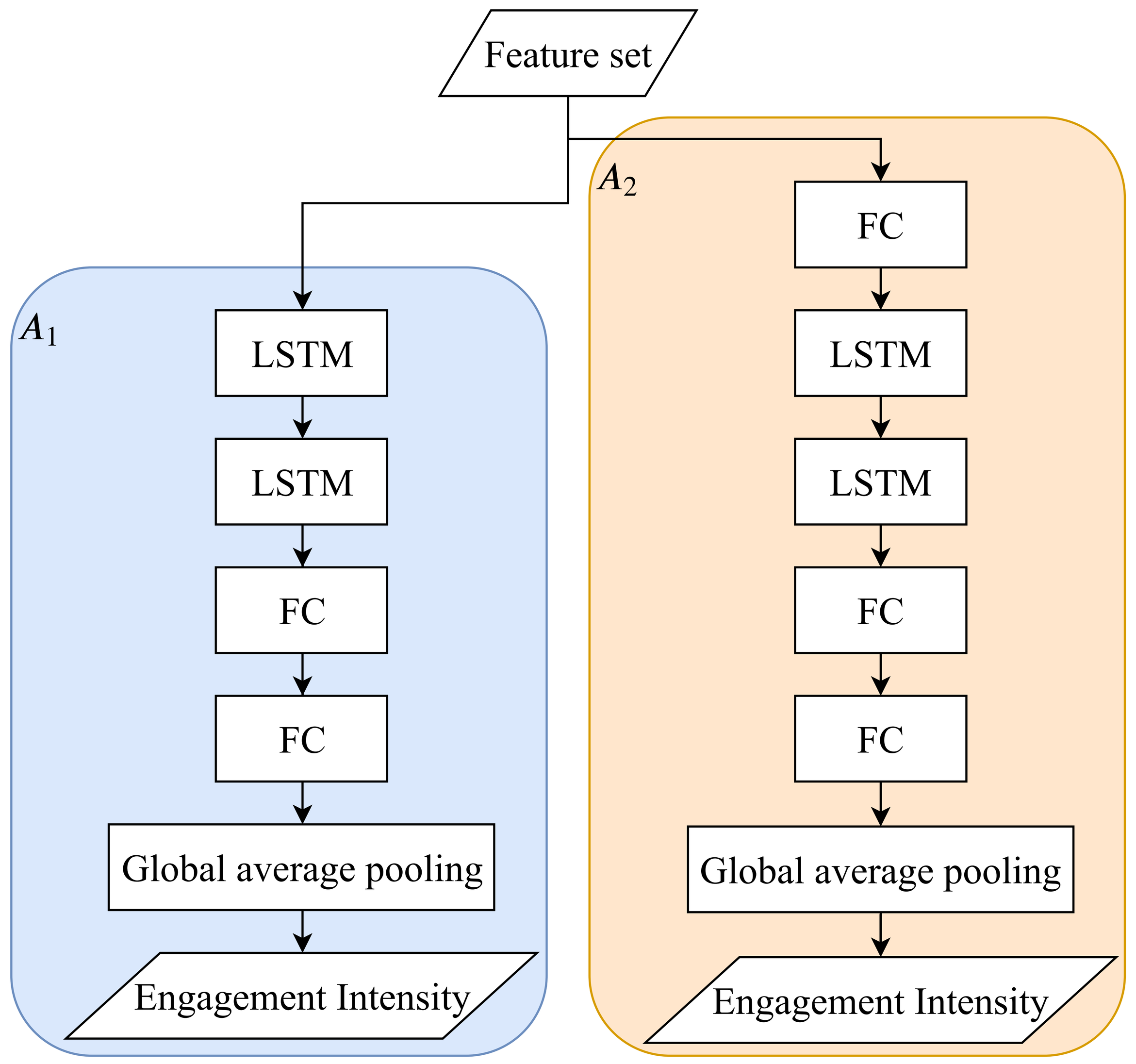
*Corresponding author.

Architecture

Each video goes through OpenFace [3] to extract face region, facial landmark and gaze direction. We divide the video sequence v into ℓ segments $s_1, s_2, \vdots, s_\ell$ with $|s_i \cup s_{i+1}| = 0.5$ and $|s_i| = |s_{i+1}|, i = \overline{1, \ell-1}$.

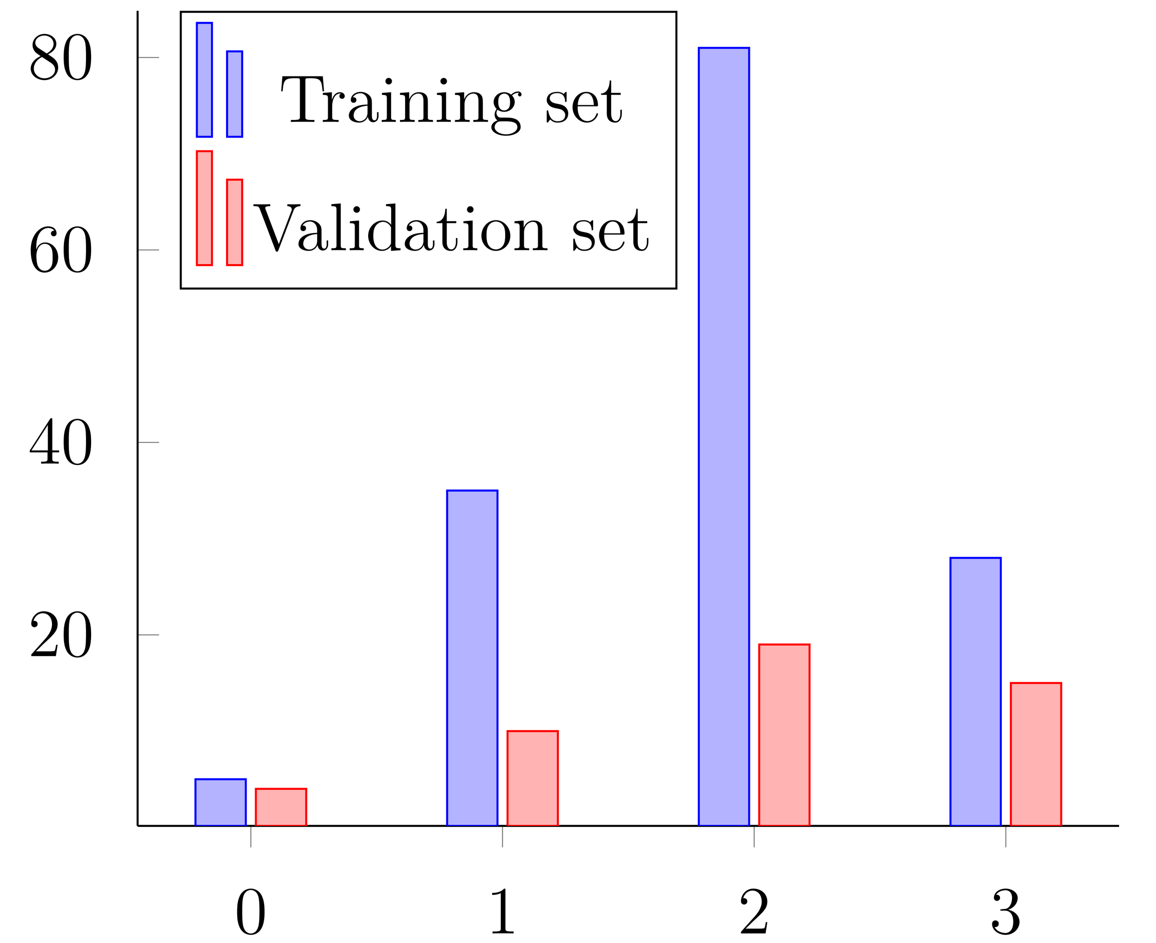


We explore two feature sets: F_1 - eye gaze and head pose features, F_2 - facial features from SE-ResNet [4]. Each feature set is classified by two networks A_1, A_2.



Experiment & Results

EW dataset in EmotiW 2019 contains 4 engagement levels: *disengaged* (DE, 0), *barely engaged* (BE, 1), *engaged* (E,2) and *highly engaged* (HE, 3) with a highly unbalanced.



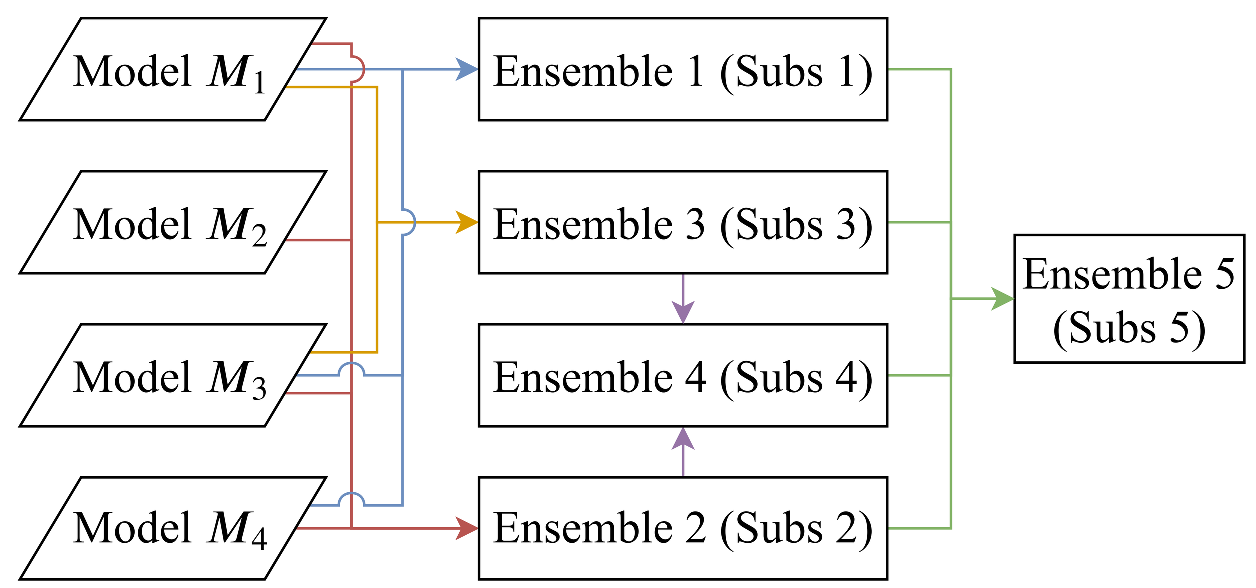
We empirically selected $\ell_1 = 15$ and $\ell_2 = 21$ which are the number of segments for F_1 and F_2, respectively. Dimensions, the output shape of FC, LSTM layers in A_1, A_2 are summarized in the table below.

Input	Model	Network	Dimension
F_1	M_3	A_1	$\ell_1 \times [128, 128, 128, 128, 1]$
F_1	M_4	A_2	$\ell_1 \times [100, 128, 128, 48, 128, 1]$
F_2	M_1	A_1	$\ell_2 \times [64, 128, 64, 128, 1]$
F_2	M_2	A_2	$\ell_2 \times [64, 64, 128, 48, 64, 1]$

Our ensemble models based on two techniques: Support Vector Regression (SVR) with RBF kernel and the following equation

$$V_{fused} = \sum_{k=1}^{m} \alpha_k V_k, \quad \text{w.r.t} \quad \sum_{k=1}^{m} \alpha_k = 1 \qquad (1)$$

where V_k denotes the output of model k. The following figure describe the progress of our fusion to achieve final model.



Experiment & Results

Subs	Test$_{\text{MSE}}$ DE	BE	E	HE	Overall
Yang et. al. [5]	-	-	-	-	0.0626
Niu et. al. [6]	-	-	-	-	0.0724
Thomas et. al. [7]	-	-	-	-	0.0792
Chang et. al. [8]	-	-	-	-	0.0813
Ensemble 1	0.3342	0.0834	**0.0133**	0.0660	0.0787
Ensemble 2	0.3289	0.1087	0.0270	**0.0353**	0.0911
Ensemble 3	0.2686	0.0644	0.0231	0.0640	0.0696
Ensemble 4	**0.2204**	0.0405	0.0320	0.1022	0.0628
Ensemble 5	0.2461	**0.0297**	0.0224	0.1378	**0.0597**

Reference

[1] A. Dhall, R. Goecke, S. Ghosh, and T. Gedeon, "EmotiW 2019: Automatic emotion, engagement and cohesion prediction tasks," in *Proceedings of the 2019 on International Conference on Multimodal Interaction*, p. in press, ACM, 2019.

[2] A. Kaur, A. Mustafa, L. Mehta, and A. Dhall, "Prediction and localization of student engagement in the wild," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8, IEEE, 2018.

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 59–66, IEEE, 2018.

[4] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[5] J. Yang, K. Wang, X. Peng, and Y. Qiao, "Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 594–598, ACM, 2018.

[6] X. Niu, H. Han, J. Zeng, X. Sun, S. Shan, Y. Huang, S. Yang, and X. Chen, "Automatic engagement prediction with gap feature," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 599–603, ACM, 2018.

[7] C. Thomas, N. Nair, and D. B. Jayagopi, "Predicting engagement intensity in the wild using temporal convolutional network," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 604–610, ACM, 2018.

[8] C. Chang, C. Zhang, L. Chen, and Y. Liu, "An ensemble model using face and body tracking for engagement detection," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*, pp. 616–622, ACM, 2018.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2017R1A4A 1015559, NRF-2018R1D1A3A03000947).