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

Jianfei Yang
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore

Kai Wang
SIAT, Shenzhen Key Laboratory of Virtual Reality and
Human Interaction Technology
P.R. China

Xiaojiang Peng*
SIAT, Shenzhen Key Laboratory of Virtual Reality and
Human Interaction Technology
P.R. China

Yu Qiao†
SIAT, Shenzhen Key Laboratory of Virtual Reality and
Human Interaction Technology
P.R. China

ABSTRACT

This paper elaborates the winner approach for engagement intensity prediction in the EmotiW Challenge 2018. The task is to predict the engagement level of a subject when he or she is watching an educational video in diverse conditions and different environments. Our approach formulates the prediction task as a multi-instance regression problem. We divide an input video sequence into segments and calculate the temporal and spatial features of each segment for regressing the intensity. Subject engagement, that is intuitively related with body and face changes in time domain, can be characterized by long short-term memory (LSTM) network. Hence, we build a multi-modal regression model based on multi-instance mechanism as well as LSTM. To make full use of training and validation data, we train different models for different data split and conduct model ensemble finally. Experimental results show that our method achieves mean squared error (MSE) of 0.0717 in the validation set, which improves the baseline results by 28%. Our methods finally win the challenge with MSE of 0.0626 on the testing set.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**;

KEYWORDS

Engagement intensity prediction; multiple instance learning

ACM Reference Format:
Jianfei Yang, Kai Wang, Xiaojiang Peng, and Yu Qiao. 2018. Deep Recurrent Multi-instance Learning with Spatio-temporal Features for Engagement Intensity Prediction. In *2018 International Conference on Multimodal Interaction (ICMI '18), October 16–20, 2018, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3242969.3264981

1 INTRODUCTION

Online education is becoming more and more popular due to abundant education resources and convenience. In the e-learning environment such as Massive Open Online Courses (MOOCs), people can easily access the knowledge they require and even get certificate in specific fields. Different from the normal learning phase that can be assessed directly by teachers around, student engagement in the online education scenario can only be described by camera recording during learning and course feedback. How to model the student engagement using limited data is still a challenge.

In this task, we aim to model the engagement intensity by merely videos in which subjects are watching online educational courses. The task formulates the engagement into four levels $(0, 1, 2, 3)$ indicating *completely disengaged, barely engaged, engaged* and *highly engaged* respectively [12]. For each video provided, we are able to find out the face and body of each subject clearly, though sometimes the environment is pretty dark or there exist other people in the background. These situations compound the difficulties of engagement prediction. Our approach presented in this paper is designed to tackle the problems from the perspective of multiple instance learning (MIL), multi-modal learning and deep representation learning.

Related work. Student engagement can be analyzed in multi-dimensions and components. In definition, it consists of *Behavioral Engagement, Emotional Engagement, Cognitive Engagement* and *Agentic Engagement* [7]. In MOOCs, automatic engagement prediction can be based on kinds of data modalities such as student response [10], facial [14, 16] or body movements in learning videos [6, 18], behavior in test quizzes [9] and even advanced physiological and neurological measures [3, 8]. Among them, video data is a good trade-off between capturing convenience and granularity. Using videos, Whitehill et al. analyze facial features and builds a SVM classifier for engagement prediction. Apart from faces of subjects, their postures also make a important role in engagement, which is utilized in [2]. Another interesting work employs both facial features and test logs to analyze their learning levels[5]. In summary, existing methods focus on extracting various features from human body and faces, and adopt different regression or classification model

*corresponding author
†common corresponding author
‡SIAT: Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences

ICMI '18, October 16–20, 2018, Boulder, CO, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5692-3/18/10…$15.00
https://doi.org/10.1145/3242969.3264981

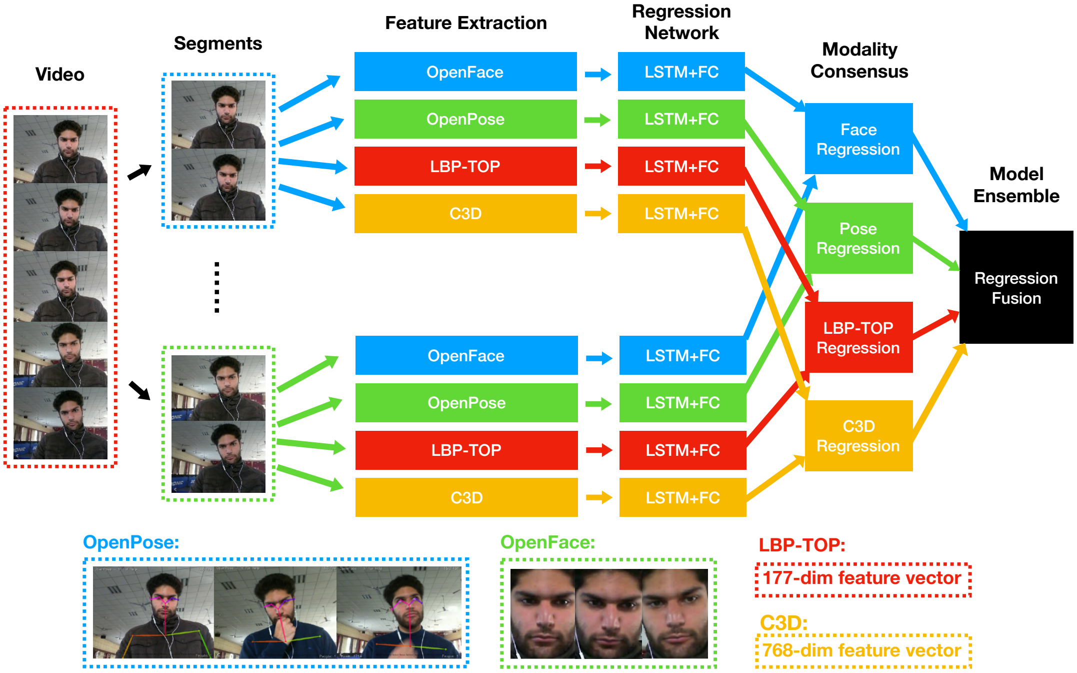

Figure 1: The system pipeline of our approach.

for prediction. We can see that they extract features either from a whole video sequence or sampling images, which brings about redundancy or information loss, and they mainly employ statistical features and traditional image descriptors, which is rather limited. Considering these limitations, we propose our multi-instance learning framework that not only preserves segment-level information, but also combines both traditional and novel deep features.

2 APPROACH

2.1 System Pipeline

Our system pipeline is shown in Figure 1. We design a multi-instance regression framework that allows various features such as Local Binary Patterns (LBP), Convolutional 3D (C3D) and statistical temporal features. Then we ensemble these multi-modal results and enhance our approach by several dataset splits with more balanced distribution.

2.2 Multiple Instance Learning framework

Existing methods mostly extract features from each frame or a whole video sequence, but this usually leads to redundancy or loss of local information. Considering the limitations above, 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 $F_k = [f_k^1, f_k^2, f_k^3, ..., f_k^m]$ from a segment and feed them into our framework. Section 2.3 will introduce our features in detail. These features are then processed by LSTM based network and three fully connected layers. Note that the regression parameters for different feature modality are different, while those of the same modality can adopt shared weights. Then we can obtain a regression value r_i for each segment s_i. Eventually we employ a mean pooling layer for segmental consensus, and optimize our network by MSE loss between annotated engagement level and regression result. Our loss can be summarized as follows:

$$L_m = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \frac{1}{k}\sum_{j=1}^{k} G_m(f_j^m))^2, \tag{1}$$

where G_m denotes the regression network transformation for feature f_m. This loss can optimize our framework for one modality m and we can train the network for multiple modalities, obtaining M regression parameters for M modalities.

Specifically, the videos in the training data are captured at 30fps and we preserve the video frame to a resolution of 640 × 480 at 6 fps. The size is adjusted according to different feature extraction scheme. Then we divide the sequence by average length and these video clips are the inputs of our framework. With this multi-instance network, the next problem is to select discriminative modal features.

2.3 Multi-modal Features

Our feature selection should be related to the factors of engagement. Imagine that in the video where a subject is watching an online educational clip, the movements of the subject should indicate their degree of concentration. The statistical features such as standard deviation (std) reflects the degree of movement and thus concentration. In addition, their specific actions such as thinking, writing or scratching head are also useful hints for engagement prediction. Inspired by these possible factors, we design our features from three perspectives including four modalities.

2.3.1 *Statistical Characteristics: Gaze, Head and Body Posture.*

It is obvious that engagement intensity is negatively correlated to the degree of human motions [7]. If the subject conducts little movements of gaze and pose, it may focus on the lecture content. If the subject is barely engaged, he should be absent-minded with drifting gazes and substantial body movements. To this end, we capture the gaze and head movement features using OpenFace [1] while body posture characteristics via OpenPose [13].

Eye Gaze: Gaze coordinates of both eyes are extracted via OpenFace for each frame. We then calculate the average variance of the points compared to the mean location in each segment, and 6 gaze-related features are obtained for one segment.

Head Pose: Movements of head are sometimes aligned with gaze drifting but sometimes not. We consider that when a subject moves his head then their gaze also make some changes. However, conversely we cannot get same inference. This insight reminds us that head pose features should be complementary with gaze. We utilize standard deviation of 6 head pose features in OpenFace, and concatenate both head and gaze characteristics for each segment.

Body Posture: For further considerations, human body movement is also an indicator of engagement intensity. Their actions reflected by body movement should include more information of specific purposes such as contemplation or writing notes. These actions can be captured by detection of body key points via OpenPose. We select 14 frequent-detected keypoints that indicating the upper body movement and also use their standard deviations as features. Since OpenFace and OpenPose work in different scheme, the information of body and face cannot be synchronized. We regard the body feature as an independent modality.

The statistical method we adopt in face, head and body features is rather limited. They can only reflect the degree of movement of different component, but the concrete actions and gaze changing patterns are neglected. More severely, high-level extraction using open library losses much spatial information especially for faces in the video.

2.3.2 *Facial Descriptor: LBP-TOP.*

We believe that apart from gaze, emotion is also a significant indicator of engagement intensity. Faces in all frames are detected by MTCNN [19] which is a cascade CNN-based face detection method. Using the detected faces, we are able to compute some features that may infer the emotions. To this end, we think about Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [20] that describes spatio-temporal texture. Specifically, the output of MTCNN is a face image with size of 112×96 and we can obtain a collection of faces for one segment. Then LBP-TOP features are extracted for each video clip, which is used as fine-grained facial feature in our approach. We do not employ the LBP-TOP features of a full video sequence due to redundancy.

2.3.3 *Action Features: C3D.*

Recently, deep learning improves many computer vision tasks significantly. C3D[11, 15] is a novel deep feature that achieves good performance in activity recognition. As we have modeled facial characteristics by LBP-TOP, C3D can be a robust representation for the body action. 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. The body image is resized to 228×228 for C3D input. We finally obtain 768 dimensions C3D feature as a modality for one segment.

2.4 Dataset Split and Model Ensemble

To fully make use of training and validation data, we not only train our model in the official training-validation split, but also make our own data split. We generate new data splits by utilizing all validation data for training. The new training split consists of all official validation data and some training data. We make the training class balanced as much as possible while we preserve the ratio of training and validation as same as the official one (150:48). We also preserve the subject independence in the new split. Our different splits mainly address some intractable problems if compared with only using official splits:

Better generalization: The public validation dataset can also help our model to generalize well. Also, if we just use all data for training, it is hard to tell whether the model performs good as training dataset.

Data balance: Original training data have severely imbalanced samples of each class, which hinders the training. When we split the dataset, we reconcile the balanced class distribution with our training. Using the new split, we always get better result for each modalities in the new validation dataset. In other words, our model converges more thoroughly and faster.

Once we get results from various modalities and diverse splits, the next step is to ensemble these models. Weighted summation is regarded as a useful fusion method for regression problem. Although we try kinds of weights in the experiment, the same weight always leads to the best result. Hence, assuming we have K models to ensemble, we adopt the same weight $\frac{1}{K}$ for each model.

3 EXPERIMENTS

In this section, we firstly present the *Engagement in the wild* dataset in EmotiW 2018 and the implementation details. Then we show the evaluations of our models trained on different modality feature, as well as the ensemble result.

3.1 Dataset

The dataset [12] used for engagement intensity prediction of EmotiW 2018 [4] is collected from 78 subjects (25 female and 53 male) in total. Their ages vary from 19 to 27 years. In the first phase, we access to 150 training and 48 validation videos and each video only include one subject for approximate 5 minutes long. In the second phase, 67 test videos are provided and our model should predict the intensity for each one. The dataset is collected in unconstrained environments including office, hotels, open ground etc, and in different time including day and night. Some videos suffer from extremely dark light and some contain several moving onlookers.

3.2 Implementation Details

For the model configurations, we use two LSTM layers for each modality feature and the LSTM layer has 64 hidden states. We also investigate the effects of using single LSTM layer. Then three dense layers are connected with size of 1024, 512 and 128 respectively. We implement the whole framework by *Pytorch*. The initial learning rate is set to 0.01 and multiplied by 0.1 every 20 epochs.

The total epoch number of training is 60. During training, we also utilize a momentum of 0.9 and a weight decay of $5e^{-4}$ for better convergence.

We preprocess all frames by the OpenPose[13] and OpenFace toolbox[1]. For LBP-TOP and C3D, we also adopt open source codes to extract them. We use MATLAB codes for LBP-TOP feature extraction and Caffe codes for C3D. For data splits, we evaluate our approach on both *official split* and *new split.*

Table 1: MSE Results on Validation set of official 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	-

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

3.3 Experiment Results

Evaluation on official split We firstly conduct experiment on official split, and the results are shown in Table 1. Normalized result in the table means that we approximate the regression value that is larger than 0.8 and less than 0.2 to 1.0 and 0.0 respectively. Our normalization method is able to improve some model slightly because all the models often generate regression result from 0.2 to 1.0. This is probably because the training samples belonging to 0.0 and 1.0 are too puny.

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. However, when we conduct experiments of LBP-TOP and C3D features, we find that the results are below our expectations. We think that it may stem from overfitting problem on account of inadequate training data. The model based on LBP-TOP and C3D achieves MSE of 0.0909 and 0.0865 respectively. If the dataset can be more abundant, we infer that high-level features should play a more momentous role in prediction.

Since statistical features seem to work better, we also use new split to train our models based on them. Shown in Table 2, it is amazing that our new split effectively 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 official validation data enhances the generalization of our approach, which blossoms a lot in the ensemble with the models optimized by official data.

Eventually, we summarize our 7 submissions as follows. Table 3 shows the corresponding result in the validation and test set. MSE across each intensity level is also provided including *not engaged (NE), barely engaged (BE), engaged (E)* and *strongly engaged (SE).* The 1st, 2nd and 4th runs are overfitting in the validation set. The 3rd run got the best result in the test dataset. All of the last three runs had reasonable results in the test set.

(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 official 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

4 CONCLUSION

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. Experimental results demonstrate the neffectiveness of our method, and we eventually win the challenge with MSE of 0.0626.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (U1613211, 61502152), Shenzhen Basic Research Program (JCYJ20150925163005055, JCYJ20170818164704758), and International Partnership Program of Chinese Academy of Sciences (172644KYSB20160033,172644KYSB20150019).

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, Mahadev Satyanarayanan, et al. 2016. Openface: A general-purpose face recognition library with mobile applications. *CMU School of Computer Science* (2016).

[2] Nigel Bosch, Sidney K D'Mello, Ryan S Baker, Jaclyn Ocumpaugh, Valerie Shute, Matthew Ventura, Lubin Wang, and Weinan Zhao. 2016. Detecting Student Emotions in Computer-Enabled Classrooms.. In *IJCAI*. 4125–4129.

[3] Maher Chaouachi, Pierre Chalfoun, Imène Jraidi, and Claude Frasson. 2010. Affect and mental engagement: towards adaptability for intelligent systems. In *23rd International FLAIRS Conference*.

[4] Abhinav Dhall, Amanjot Kaur, Roland Goecke, and Tom Gedeon. 2018. EmotiW 2018: Audio-Video, Student Engagement and Group-Level Affect Prediction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[5] Sidney K D'Mello, Scotty D Craig, and Art C Graesser. 2009. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology* 4, 3-4 (2009), 165–187.

[6] Sidney K DâĂŹMello and Arthur Graesser. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20, 2 (2010), 147–187.

[7] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.

[8] Benjamin S Goldberg, Robert A Sottilare, Keith W Brawner, and Heather K Holden. 2011. Predicting learner engagement during well-defined and ill-defined computer-based intercultural interactions. In *International Conference on Affective Computing and Intelligent Interaction*. Springer, 538–547.

[9] E Joseph. 2005. Engagement tracing: using response times to model student disengagement. *Artificial intelligence in education: Supporting learning through intelligent and socially informed technology* 125 (2005), 88.

[10] Kenneth R Koedinger, John R Anderson, William H Hadley, and Mary A Mark. 1997. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education (IJAIED)* 8 (1997), 30–43.

[11] Zheng Li, Jianfei Yang, Juan Zha, Chang-Dong Wang, and Weishi Zheng. 2016. Online visual tracking via correlation filter with convolutional networks. In *Visual Communications and Image Processing (VCIP), 2016*. IEEE, 1–4.

[12] Aamir Mustafa, Amanjot Kaur, Love Mehta, and Abhinav Dhall. 2018. Prediction and Localization of Student Engagement in the Wild. *arXiv preprint arXiv:1804.00858* (2018).

[13] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2.

[14] Lianzhi Tan, Kaipeng Zhang, Kai Wang, Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2017. Group emotion recognition with individual facial emotion CNNs and global image based CNNs. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 549–552.

[15] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*. 4489–4497.

[16] Kai Wang, , Xiaoxing Zeng, Jianfei Yang, Debin Meng, Kaipeng Zhang, Xiaojiang Peng, and Yu Qiao. 2018. Cascade Attention Networks For Group Emotion Recognition with Face, Body and Image Cues. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (in press)*. ACM.

[17] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing* 5, 1 (2014), 86–98.

[18] Xiang Xiao, Phuong Pham, and Jingtao Wang. 2017. Dynamics of affective states during mooc learning. In *International Conference on Artificial Intelligence in Education*. Springer, 586–589.

[19] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters* 23, 10 (2016), 1499–1503.

[20] Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence* 29, 6 (2007), 915–928.