Design of an Online Education Evaluation System Based on Multimodal Data of Learners

Qijia Peng[1], Nan Qie[2], Liang Yuan[3], Yue Chen[2], and Qin Gao[2]

[1]Graduate School of Comprehensive Human Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577 Japan
[2] Institute of Human Factors and Ergonomics, Department of Industrial Engineering, Tsinghua University, Haidian District, Beijing, 100084, P. R. China
[3] Graduate School of System and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577 Japan
pengqj92@hotmail.com

Abstract. Online education breaks the time and space constraints of learning, but it also presents some new challenges for the teachers: less interaction between instructors and learners, and loss of real-time feedback of teaching effects. Our study aims to fill these gaps by designing a tool for instructors that shows how learners' status change along the lecture video timeline. The study uses multimodal data consist of facial expressions and timeline-anchored comments and labels the data with two learning status dimensions (difficulty and interestingness). To acquire training dataset, 20 teaching video clips are selected, and 15 volunteers are invited to watch the videos to collect their facial expressions and subjective learning status ratings. Then we build a fusion model with results from a CNN (Convolutional Neural Network) model and a LSTM (Long Short-Term Memory) model, and design an effective interface to present feedbacks from the model. After evaluation of the model, we put forward some possible improvements and future prospects for this design.

Keywords: Online Education, Multimodal Data, Deep Learning.

1 Introduction

Online courses break the constraints of time and space and share high-quality educational resources through the Internet. Compared with face-to-face learning in classrooms, online learning enriches learners' interactions with learning materials such as lecture videos but decreases social interactions. Especially, a massive open online course (MOOC) usually involves several instructors and thousands of learners. From instructors' perspectives, it is hard to gather and analyze the large scale of learning data from all the students effectively and efficiently. Instructors cannot get feedbacks during lectures as they can in classrooms. They also experience overwhelming information from different sources and feel difficult to fully use it to improve instruction [1]. As a result, learners less interact with instructors and perceive less teaching pres-

ence, i.e., perceived level of effective instruction and guidance from instructors [2, 3] and is found to promote effective learning, performance and satisfaction [4, 5].

To help instructors analyze learning data and further promote learners' teaching presence, educational systems have adopted learning analytics and educational data mining. Most previous research focused on a course level analysis, such as to predict performance, predict dropout, enhance social interactions, and recommend resources [6]. To improve instruction, instructors need more detailed information in a lecture video level, i.e., learners' feedback along with the video timeline, so that they can improve their instructions of different knowledge points. Several studies designed tools to analyze learners' cognitive load and engagement in a lecture video by video clickstream [7–9]. The overall clickstream pattern of a video could well predict the possibility of dropout, but it remained hard to interpret learners' feelings by the click data at a specific video timepoint.

To get learners' feedbacks in fine granularity, researchers could choose some alternative data sources changing along with the video timeline. Physiological signals such as facial expressions and eye gazing, change during lecture learning and can well reflect learners' mental states such as cognitive load and engagement. However, most physiological data are hard to collect in online learning contexts and therefore are adopted by few studies [10]. Some research collected data through web cameras on learners' devices [11–13]. They aimed to design adaptive learning systems for learners but not to show how learners' states change along with the lecture video for instructors.

Besides physiological data, contents generated by learners can also reflect learners' engagement. Most previous research [14, 15] analyzed forum discussion data, which showed a course level of learners' feedbacks. To show a video level of feedbacks, a better data source is timeline-anchored commenting, which has been incorporated in many studies to promote discussions during learning lecture videos [16–21]. Learners can post a comment or an annotation specific to a playback time of the video, and later viewers will see these comments when the video plays to that exact time point. Timeline-anchored commenting can expose learners with more discussions specific to timepoints and further promote learning [17, 18, 20, 22, 23]. We found one study attempted to analyze these timeline-anchored comments and design a visualization tool for instructors [24, 25]. This tool showed how learners' emotion valence, the relevance of discussions, and the topics changed along the video timeline. But it missed learners' perceived difficulty or workload and learners' interest or engagement, which instructors are highly interested in [26, 27].

Therefore, our study aims to fill these gaps and design a system for instructors that shows how learners' perceived learning status change along the lecture video timeline. In addition, most studies only analyzed data from a single data source despite the diversity of sources according to two recent review papers about learning analytic tools [10, 28]. To provide instructors with a more comprehensive view, this study attempts to build a tool to analyze data from different data sources, i.e., facial expressions and timeline-anchored comments, with multimodal methods.

2 Method

In order to construct the system, we design a data collection platform, collect data from a sample group of students, preprocess the data, and train an artificial neural network model based on those data.

2.1 Data Collection

Before we train the model and build the feedback system, basic work such as choosing evaluation dimensions, collecting training data and data formatting should be done.

Learning Status Dimensions

First, we determine proper dimensions to evaluate learning status. In order to give structural feedback to the teachers, the dimensions should be real-time, relevant to facial expressions and comments, easy to understand and evaluate, and instructive to teaching improvement.

According to the review of Student Rating of Teaching [26, 27], we choose two dimensions: "difficulty/workload" and "level of interest" . We use "Difficult/Easy" and "Interesting/Boring" as the name of these two dimensions in the following study.

Data Collection Platform

We need both facial expressions and comments as training dataset, and watchers' emotion feedback ratings (on difficulty and interestingness) as label dataset. Thus, we build an online experiment platform based on HTML/JavaScript. Participants watch some selected video clips on the platform, and their facial expressions and ratings will be recorded.

First, we choose video clips from documentaries, speeches and public classes on a Chinese video sharing website "Bilibili". This website has both good quantity and quality of educational videos and is famous for the "Danmaku" (a type of timeline-anchored comments) culture in China. The proper video clip should be easy to trigger emotions about difficulty or interestingness and rich in watchers' comments. 20 video clips (in Chinese or English) are finally selected by the researchers after rated separately on both dimensions. The topic of the teaching videos includes physics, math, language and history. Each video clip contains only one specific topic and lasts for around 1 minute, which makes it possible for the emotion feedback ratings to reflect watchers' learning status. Timeline-anchored comments of those video clips are also collected respectively for the study.

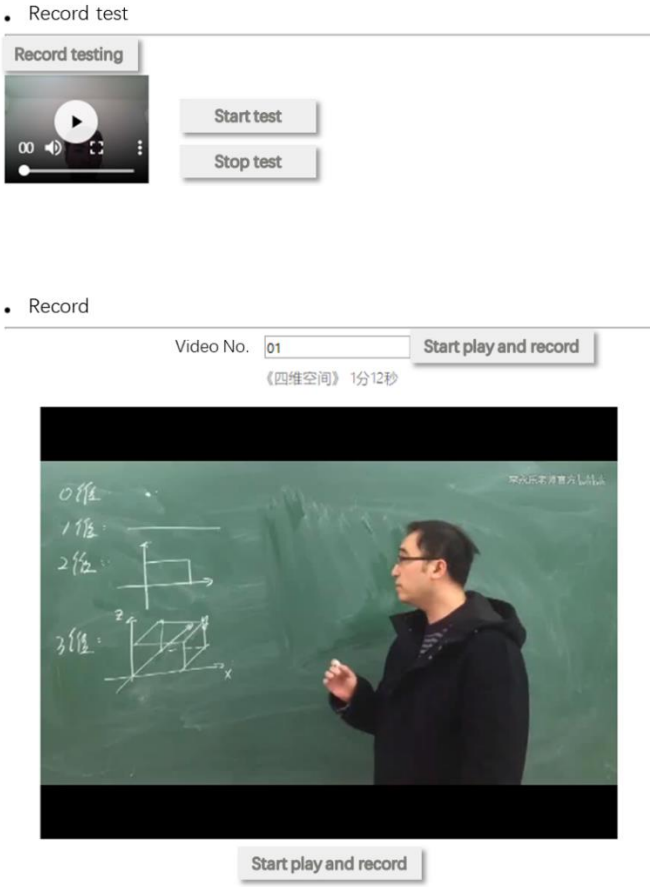

Fig. 1. Facial expressions record interface.

The webpage of the platform collects facial expressions during the participants watching the video by using the camera on the PC (MediaStream interface on Chrome browser), saving as webm format videos (24 fps) (See **Fig. 1**). The webpage starts recording when the video clip starts to play, and end recording when the video is played. The emotion ratings towards the video are collected by two sliders on the webpage (range from -1 to 1, step 0.1), and the sliders will show after finishing watching the video (See **Fig. 2**).

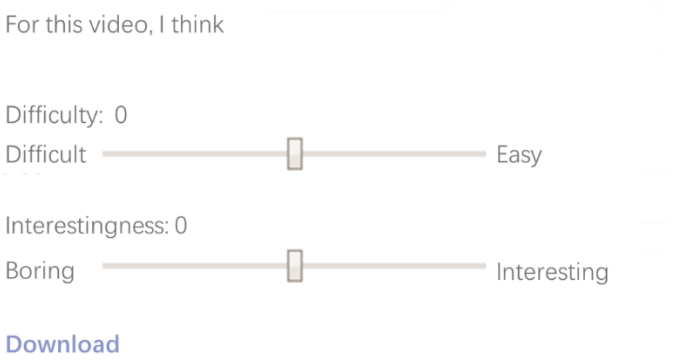

Fig. 2. Learning status rating interface

Participants
Fifteen university students (7 man and 8 woman), aging from 21 to 27, recruited from Tsinghua University and University of Tsukuba participated in the study. All participants are native speakers of Chinese, and have enough English ability to understand videos in English.

2.2 Data Preprocessing

Before the comprehensive fusion training of the multimodal data, we first preprocessed data of facial expressions and comments separately.

For facial expression data, we use interface from opencv in python to detect and extract facial expressions from the recorded videos. We choose one facial expression (screenshot in the video) per second and extract facial expressions by a well-trained universal model (haarcascade_frontalface_default) from opencv. All facial expressions pictures are set in 128px * 128px * 1 channel after resizing and grayscale processing. For each video of each participant we choose 20 pictures to analyze and labeled with his/her emotional ratings in two dimensions (difficulty and interestingness).

For timeline-anchored comments ("Danmaku" in this study), we have to extract those topic-related comments and screen out the irrelevant, meaningless comments such as greetings, internet memes, repeating words and emoticons. Moreover, in "Danmaku" culture, contents tend to be cute and popular, which also makes it not suitable for training in universal emotional analysis model directly. Thus, we set up both stop word list and "word meaning transfer" list. After screening the irrelevant contents, we calculate the emotional tendency classification and its possibility and confidence for each "Danmaku" comment by the interface provided by Baidu NLP. Then we choose those comments with confidence> 0.1 and split the sentence into words by interface from Jieba Chinese text segmentation. Each word is then transferred to a word vector by genism and unified in length. Finally, we get 899 valid

Danmaku comments in total and label them with the average emotional ratings in two dimensions (difficulty and interestingness) of the videos accordingly.

Moreover, the subjective evaluation of "Difficulty" and "Interestingness", which will be used as label in the training process, is convert to a classification factor (-1/1) according to the emotional ratings (positive/negative).

2.3 Model Training

We choose randomly from the preprocessed dataset into the training set, which contains facial expressions of 12 participants and 750 Danmaku comments.

The fusion model is based on Stacked Generalization [29]. We use a 4-level Long Short-Term Memory Network (LSTM) model to classify the data from comments, and a 5-level Convolutional Neural Network (CNN) model with an MSE loss function to train the data of facial expressions. In the fusion level, those two models are integrated by a 2-level Stacked Generalization-based ensemble model, which outputs the final classification of the learning status. Those three models are trained separately.

Fig. 3 summarizes the structure and process of the model. A Danmaku comment labeled by the video's average rating enters the LSTM model and it outputs a predicted classification. A facial expression labeled by the according rating enters the CNN model and it outputs a predicted classification. Then Danmaku comment and facial expressions which are from the same person watching the same video will then enter the fusion level, and the model will output the final predicted classification.

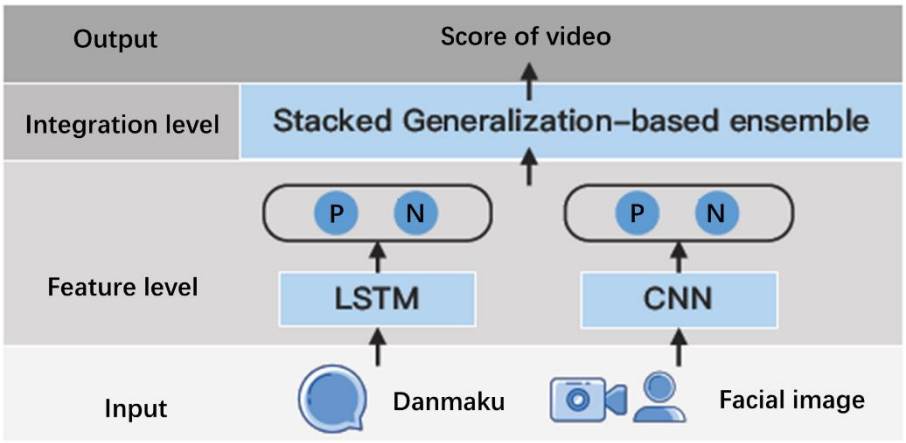

Fig. 3. The fusion model including an LSTM model and an CNN model.

Considering that information in Danmaku comments cannot easily reflect feelings about "Difficulty", we use only facial expression data with CNN model in "Difficulty" dimension.

3 Model evaluation

We have facial expression data from 3 participants and 149 Danmaku comments in the test set. Accuracy, precision, recall and F-measure are calculated separately according to the results from the test set (see **Table 1**).

Table 1. Accuracy, Precision, Recall and F-measure of the model.

		Interestingness	Difficulty
Accuracy		0.624	0.584
Precision	Positive	0.652	0.744
	Negative	0.612	0.365
	Average	0.632	0.555
Recall	Positive	0.429	0.615
	Negative	0.797	0.511
	Average	0.613	0.563
F-measure	Positive	0.517	0.674
	Negative	0.692	0.426
	Average	0.605	0.550

The accuracy of this model for interestingness is 62.4%, and for difficulty is 58.4%, which is higher than a random classification (50%). Considering the sample size of training dataset, the results of accuracy is acceptable and bigger size of training set is necessary for the improvement of the model's accuracy.

The precision for interestingness is 65.2% for positive feedback and 61.2% for negative feedback, and for difficulty is 74.4% for positive and 36.5% for negative. The precision is better when detecting "interesting", "boring" and "difficult", but not good in "easy".

The recall is only 42.9% for "interesting" and 51.1% for "difficult", but better performance on "boring" (79.7%) and "easy" (61.5%). The reason might lie in the habits of facial expression from the watchers. Comparing with "boring" and "easy", "interesting" and "difficult" are more likely to provoke larger facial expression changes, and thus the model tends to relate rich expressions to interesting and difficult learning status. But considering that the sample of facial expressions in this study are chosen randomly within around 60 seconds, rich expressions in limit time could possibly diluted by large number of plain expressions, which leads to a relatively worse results on recall. A possible improvement method is to find more effective emotion trigger and improve time accuracy for expression detection.

F-measure is more comprehensive and considers both precision and recall. For results of F-measure, the model of interestingness (60.5% on average) is better than the model of difficulty (55.0% on average), showing the advantage of using related multimodal data and fusion of models.

4 Discussion

This study put forward a comprehensive evaluation method for the multi-dimensional feedbacks from students to the instructors in online education situations. In this section, we discuss about the contribution of this model based on facial expressions and comments in theoretical aspects, and the future prospects for implementation of the system in practical aspects.

4.1 Theoretical Contribution

In theoretical aspect, the methodology of this study can be used to establish links of more learning status with facial expressions and comments. Recent researches and products are already able to recognize general emotions (such as happy, angry, etc.) with facial expressions, but in online education situations, learning status may not be always associated with general emotions. In the context of online learning, this study presents a possible method to detect more learning related status, such as concentration and distraction, by analyzing facial expressions labeled with learners' status. Also, the timeliness provided by timeline-anchored comments may also help detecting real-time learning status by emotional recognition of the comments.

Moreover, this study provides another way to recognize emotions from comments like "Danmaku" in the future. Comments in "Danmaku" are more like memes on the internet, and thus have special language style and are quite different from the language in everyday life. The vague range of the meaning and the lack of context make it difficult for traditional method (such as TF-IDF and word2vec) to handle. However, in online education situations, this study provides a platform where learning status related emotions can be detected, and thus make it easier to label those timeline-anchored comments even if we don't know their exact meanings.

4.2 Practical Contribution

In practical aspect, this model and system can help reduce teaching pressure of instructors facing hundreds of students at the same time when teaching on the internet. An experienced teacher can handle the students' learning status so that he/she may improve or adjust teaching strategy, but in online real-time teaching situations the instructors require more feedbacks from the students. This study provides a practical method to show comprehensive and real-time feedbacks, with data obtained from common channels: facial expressions from live camera, and real-time comments from Danmaku or chatting room.

To illustrate the feasibility of this idea, we design an interface to clearly present all the data generated from the model (see **Fig. 4**). The upper part shows the education video, and the lower part shows the trends of emotions on a line chart with an axis of timeline.

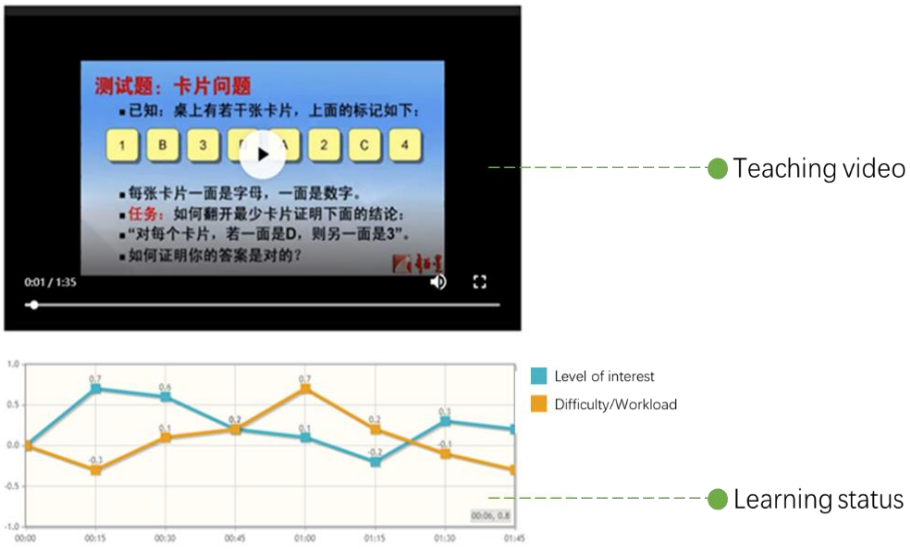

Fig. 4. Feedback interface of learning status

Based on the multimodal data collected from the video watchers, the chart shows the confidence of the emotional classification from the fusion model at a frequency of once per 15 seconds. When the number is positive, the more it closes to 1, the more likely the watcher feels interesting/easy about the video content; when the number is negative, the more it closes to -1, the more likely the watcher feels boring/difficult. This feedback might help the instructors to judge whether their students are in a good learning status.

Moreover, the feedbacks are also anchored with time. Click the data point on the chart, and the upper video will skip to the time accordingly. This would help the instructors to focus on the exact time that really matters on students' learning status.

5 Conclusion

In this study we collect the multimodal data including facial expressions and comment text when the students watch teaching videos online and label them with two dimensions (interestingness, difficulty) in the subjective learning status. Then after pre-processing such as face recognition from video screenshots and normalization of comment text, we build a fusion model with artificial neural network methods to calculate the real-time learning status in the two dimensions. The study also designs a result display interface, showing the results from the model and interactive functions that are easy for instructors to check the real-time teaching effect.

Our system can give objective, specific timeline-based feedback to instructors, which overcomes the shortages of previous feedback methods. Moreover, we put

forward a possible approach to link multiple learning status with facial expressions and recognition of learning status related comments based on the method of our study.

References

1. Bill & Melinda Gates Foundation: Teachers Know Best: Making Data Work for Teachers and Students. Bill & Melinda Gates Foundation (2015).
2. Garrison, D.R., Arbaugh, J.B.: Researching the community of inquiry framework: Review, issues, and future directions. The Internet and Higher Education. 10, 157–172 (2007).
3. Picciano, A.G.: Beyond student perceptions: Issues of interaction, presence, and performance in an online course. Journal of Asynchronous learning networks. 6, 21–40 (2002).
4. Akyol, Z., Garrison, D.R.: The development of a community of inquiry over time in an online course: Understanding the progression and integration of social, cognitive and teaching presence. Journal of Asynchronous Learning Networks. 12, 3–22 (2008).
5. Ke, F., Kwak, D.: Online learning across ethnicity and age: A study on learning interaction participation, perception, and learning satisfaction. Computers & education. 61, 43–51 (2013).
6. Papamitsiou, Z., Economides, A.A.: Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society. 17, (2014).
7. Kim, J., Guo, P.J., Cai, C.J., Li, S.-W. (Daniel), Gajos, K.Z., Miller, R.C.: Data-driven Interaction Techniques for Improving Navigation of Educational Videos. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology. pp. 563–572. ACM, New York, NY, USA (2014).
8. Shi, C., Fu, S., Chen, Q., Qu, H.: VisMOOC: Visualizing video clickstream data from massive open online courses. In: 2014 IEEE Conference on Visual Analytics Science and Technology (VAST). pp. 277–278 (2014).
9. Sinha, T., Jermann, P., Li, N., Dillenbourg, P.: Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. arXiv:1407.7131 [cs]. (2014).
10. Vieira, C., Parsons, P., Byrd, V.: Visual learning analytics of educational data: A systematic literature review and research agenda. Computers & Education. 122, 119–135 (2018).
11. Pham, P., Wang, J.: Predicting Learners' Emotions in Mobile MOOC Learning via a Multimodal Intelligent Tutor. In: International Conference on Intelligent Tutoring Systems. pp. 150–159. Springer (2018).
12. Pham, P., Wang, J.: Adaptive Review for Mobile MOOC Learning via Multimodal Physiological Signal Sensing - A Longitudinal Study. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. pp. 63–72. ACM, New York, NY, USA (2018).
13. Soltani, M., Zarzour, H., Babahenini, M.C.: Facial Emotion Detection in Massive Open Online Courses. In: World Conference on Information Systems and Technologies. pp. 277–286. Springer (2018).
14. Chen, B., Chang, Y.-H., Ouyang, F., Zhou, W.: Fostering student engagement in online discussion through social learning analytics. The Internet and Higher Education. 31, 21–30 (2018).
15. Gillani, N., Eynon, R.: Communication patterns in massively open online courses. The Internet and Higher Education. 23, 18–26 (2014).

16. Chen, Y., Gao, Q., Yuan, Q.: DanMOOC: Enhancing Content and Social Interaction in MOOCs with Synchronized Commenting. In: Rau, P.-L.P. (ed.) Cross-Cultural Design: 9th International Conference, CCD 2017, Held as Part of HCI International 2017, Vancouver, BC, Canada, July 9-14, 2017, Proceedings. pp. 509–520. Springer International Publishing, Cham (2017).
17. Chen, Y., Gao, Q., Yuan, Q., Tang, Y.: Facilitating students' interaction in MOOCs through timeline-anchored discussion. International Journal of Human–Computer Interaction. Accepted, (2019).
18. Lee, Y.-C., Lin, W.-C., Cherng, F.-Y., Wang, H.-C., Sung, C.-Y., King, J.-T.: Using Time-Anchored Peer Comments to Enhance Social Interaction in Online Educational Videos. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 689–698. ACM, New York, NY, USA (2015).
19. Leng, J., Zhu, J., Wang, X., Gu, X.: Identifying the Potential of Danmaku Video from Eye Gaze Data. In: Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference on. pp. 288–292. IEEE (2016).
20. Yao, Y., Bort, J., Huang, Y.: Understanding Danmaku's Potential in Online Video Learning. In: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems. pp. 3034–3040. ACM, New York, NY, USA (2017).
21. Yousef, A.M.F., Chatti, M.A., Schroeder, U., Wosnitza, M.: A usability evaluation of a blended MOOC environment: An experimental case study. The International Review of Research in Open and Distributed Learning. 16, (2015).
22. Chen, Y., Gao, Q., Rau, P.-L.P.: Understanding Gratifications of Watching Danmaku Videos–Videos with Overlaid Comments. In: Cross-Cultural Design Methods, Practice and Impact. pp. 153–163. Springer (2015).
23. Chen, Y., Gao, Q., Rau, P.-L.P.: Watching a Movie Alone yet Together: Understanding Reasons for Watching Danmaku Videos. International Journal of Human–Computer Interaction. 33, 731–743 (2017).
24. Sung, C.-Y., Huang, X.-Y., Shen, Y., Cherng, F.-Y., Lin, W.-C., Wang, H.-C.: ToPIN: A Visual Analysis Tool for Time-anchored Comments in Online Educational Videos. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems. pp. 2185–2191. ACM, New York, NY, USA (2016).
25. Sung, C.-Y., Huang, X.-Y., Shen, Y., Cherng, F.-Y., Lin, W.-C., Wang, H.-C.: Exploring Online Learners' Interactive Dynamics by Visually Analyzing Their Time-anchored Comments. In: Computer Graphics Forum. pp. 145–155. Wiley Online Library (2017).
26. Spooren, P., Brockx, B., Mortelmans, D.: On the validity of student evaluation of teaching: The state of the art. Review of Educational Research. 83, 598–642 (2013).
27. Zabaleta, F.: The use and misuse of student evaluations of teaching. Teaching in Higher Education. 12, 55–76 (2007).
28. Mangaroska, K., Giannakos, M.N.: Learning analytics for learning design: A systematic literature review of analytics-driven design to enhance learning. IEEE Transactions on Learning Technologies. 1–1 (2018).
29. Wolpert, D.H.: Stacked generalization. Neural networks. 5, 241–259 (1992).