Predicting Group Work Performance from Physical Handwriting Features in a Smart English Classroom

Meishu Song
The University of Augsburg
Germany

Kun Qian
The University of Tokyo
Japan

Bin Chen
Fujitsu Lab
Japan

Keiju Okabayashi
Fujitsu Lab
Japan

Emilia Parada-Cabaleiro
Johannes Kepler Universität Linz
Austria

Zijiang Yang
The University of Augsburg
Germany

Shuo Liu
The University of Augsburg
Germany

Kazumasa Togami
The University of Tokyo
Japan

Ichiro Hidaka
The University of Tokyo
Japan

Yueheng Wang
The University of Tokyo
Japan

Björn Schuller
Imperial College London
United Kingdom

Yoshiharu Yamamoto
The University of Tokyo
Japan

ABSTRACT

Embodied cognition theory states that students thinking in a learning environment is embodied in physical activity. In this regard, recent research has shown that signal-level handwriting dynamics can distinguish learning performance. Although machine learning has been considered to detect how multimodal modalities correlate to specific learning processes, the use of deep learning has received insufficient attention. With this in mind, we build a Group Work Performance Prediction system from analysis of 3D (including strokes frequency) handwriting signals of students in a smart English classroom, with deep convolutional neuronal network (CNN) based regression models. For labelling of their proficiency level, their spoken language performance is being used. The students were working together in groups. A 3D (2D writing coordinates plus frequency) handwriting dataset (3D-Writing-DB) was collected through a collaboration platform known as 'creative digital space'. We extracted the 3D handwriting signal from a table tablet during English discussion sessions. Afterwards, professional English teachers annotated the English speech (values vary from 0 - 5). Our experimental results indicate that group work performance can be successfully predicted from physical handwriting features by using deep learning, as shown by our best result, i. e., 0.32 in regression assessment by applying RMSE for evaluation.

ICDSP 2021, February 26–28, 2021, Chengdu, China

ACM ISBN 978-1-4503-8936-5/21
https://doi.org/10.1145/3458380.3458404

CCS CONCEPTS

• **Human-centered computing** → *Empirical studies in HCI*; • **Applied computing** → *Collaborative learning*.

KEYWORDS

English speaking proficiency, digital classroom, group work, handwriting, deep learning *

ACM Reference Format:
Meishu Song, Kun Qian, Bin Chen, Keiju Okabayashi, Emilia Parada-Cabaleiro, Zijiang Yang, Shuo Liu, Kazumasa Togami, Ichiro Hidaka, Yueheng Wang, Björn Schuller, and Yoshiharu Yamamoto. 2021. Predicting Group Work Performance from Physical Handwriting Features in a Smart English Classroom. In *2021 5th International Conference on Digital Signal Processing (ICDSP 2021), February 26–28, 2021, Chengdu, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3458380.3458404

1 INTRODUCTION

During the past decade, the fast development of information and communication technologies (ICT) has dramatically improved and even changed the paradigms in education evaluation. By leveraging the power of the prevalent internet of things (IoT), sensors and artificial intelligence (AI [12]) have made feasible automatic analysis and monitoring of students' performance [40]. Using technology in classroom has shown to contribute to the development of students' creativity, motivation, and critical thought, encouraging also their capability of solving problems in a more collaborative way [30]. The increasing digital research in today's classrooms has encouraged a recent development of specific computer-based approaches for their application in E-learning environments, such as classroom activity detection [21], hand-rising gesture recognition [23], and classroom motion tracking [15]. Particularly, the proliferation of sensors in classrooms has created an environment in which students' behaviours are continuously monitored and recorded [2, 28].

*Dr. Kun Qian is the corresponding author.

In this regard, research in the realm of multimodal learning analysis has assessed how students communicate, collaborate, and solve problems within a smart classroom [37]—an area of research recently stimulated by the evaluation of neuro-physiological data from these technology-based learning environments [42]. Indeed, it has been shown that users' tactile manipulation of tangible interfaces enables the development of cognitive and physical connections [35]—a research outcome supported by the higher students' ability to identify objects' parts and position when using touch-based smart-pad interfaces w. r. t. traditional mouse-based technology [20]. In this regard, to evaluate how digital classrooms might encourage students' learning, multimodal learning analytics—an emerging field of research that combines the study of different natural communication modalities, such as speech, writing, or gesturing—has been successfully taken into account [27, 29]. Yet, this endeavour is challenging: Merely providing students with multimodal learning resources does not necessarily lead to the use of such resources in assessment practices [36].

Previous research has shown that classroom discussion is a fertile ground to develop higher-order thinking, i. e., the acquisition of the critical skills which enable the ability of solving problems in new situations [5]. Indeed, during discussions, students make comments and build on each other's ideas, which encourages the debate [14], a situation specially tailored to collect cross communication modalities. Meaning-making, i. e., the process that enables students to make knowledge and experiences meaningful [16], is encouraged by implementing group work in the classroom w. r. t. traditional accommodating environments [10]. Indeed, assessing the learning potential of student group discussions has become an important feature of classroom research. The predominant methodological form underpinning such research efforts falls under the broad umbrella of discourse analysis—a research method in which language is analysed beyond the sentence, i. e., considering how the context of the discourse affects the interaction between sentences [19]. To this end, group discussions should be analysed by taking into account both the content and the communicative meaning-related aspects [1]. These aspects are displayed, for instance, through users' physiological behaviours, such as handwriting signals, which can be evaluated as indicators of specific cognitive processes [6].

From the early 80s, research aimed to provide user-friendly input solutions for handwriting and drawing recognition, which do not rely on a sensor-based tablet, has been developed [17]. In this regard, previous work on the estimation of students' attention has shown that Kinect features, such as handwriting signals or the RightEyeClosed indicator, clearly correlate to the attention level [41]. Indeed, pen based writing, one of the most ubiquitous aspects in classrooms to encourage students' learning [22], is characterised by signal-level handwriting dynamics, e. g., average duration, distance, or pressure, aspects that might encode a particular communicative meaning. Nevertheless, despite recording handwritten signals through technological devices is a common practise in today's classrooms; the investigation of how these could correlate to attributes from other communication modalities, such as speech proficiency, has received little attention at a group learning level [26].

In this regard, we evaluate how handwritten signals, collected during group discussion work in a smart English classroom, might be used as indicators of learning performance. Although it has been shown that handwriting signals correlate to learning aspects [41] such as students' attention, previous research on the automatic prediction of English speaking proficiency is mostly performed through the evaluation of speech-based features [43], while handwriting-based features have not been yet considered. Meanwhile, machine learning has been already successfully applied to automatically recognise teaching-learning related concepts, such as the classification of teacher's questions into different cognitive levels [39]; yet, the extent to which deep learning might contribute to the development of this research question is still under investigated [24]. To this end, in the presented work, the automatic prediction of group work performance from students' 3D handwriting features through a Convolutional Neural Network (CNN) model in a smart English classroom is considered for the first time.

Figure 1: Students' interaction with a table tablet by using digital cards and electronic pens during group work.

2 DATA ACQUISITION: 3D-WRITING-DB

2.1 Group Work Discussion Setting

Since students' participation in cooperative exercises within group work especially increases their motivation in linguistic activities, we performed the data collection in the Communication English course. In each session, teachers provided 5 topics per group work of students, e. g., 'Cashless Society' or 'Paperless Classroom', as well as related texts to be read. After reading the provided materials, the students of each group work had 30 min to discuss and write notes about each topic; thus, each session lasted 150 min (30 min × 5 topics). The students wrote their notes in digital cards by using infrared electronic pens. The digital cards are 'virtual' cards projected onto each group's table—from now on we will refer to each table as 'table tablet' (cf. Figure 1).

In order to enable the teachers to quickly monitor the overall classroom situation, the 'creative digital space' system [4], i. e., a digital environment based on the real-time analysis of group activities aimed to support active learning classes, was considered.

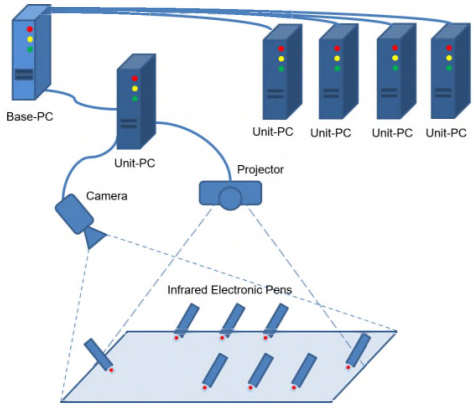


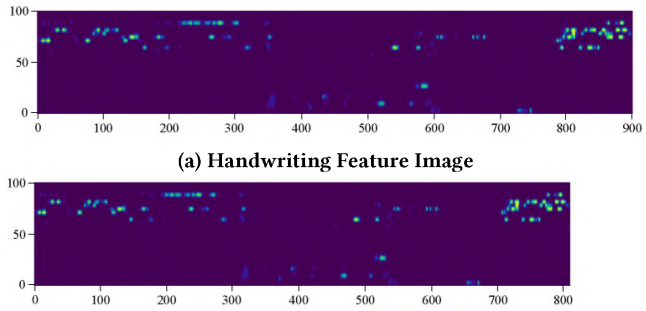
Figure 2: Digital Classroom equipment. A standard table tablet, equipped with a Unit-PC, a Projector, a Camera, and eight Infrared Electronic Pens (one per student), is shown. The Digital Classroom, equipped with five table tablets—each with its own Unit-PC (all of them connected to the Base-PC)—enables to collect data from five work groups.

The study was developed in a High School from Tokyo (Japan). A total of 80 students with ages from 15 to 17 years old and spoken English proficiency from low to high, participated in the study. The students were randomly split into 2 classroom-groups with 40 students each (A and B). In each classroom-group, they were subsequently divided into 5 group work teams of 8 students. Each classroom-group performed 5 sessions, i. e., 10 sessions were carried out in total. After each session, each student presented the arguments in a short oral presentation (1-3 min); presentations were recorded by other students with mobile-phones and tablets.

2.2 Data Collection and Annotation

To collect the handwriting data, the Digital Classroom used alternatively by the classroom-groups A and B, was equipped with 5 table tablets (one per each work group), i. e., 5 Unit-PCs with projector and camera, all connected to the Base-PC (cf. Figure 2). In their interaction with the table tablet, the students could create digital cards to write handwritten content with infrared electronic pens. Handwriting signals were produced by the students and extracted from the 'creative digital space' through the Base-PC, i. e., without the need of other card operation signals. Indicators based on operation frequency and handwriting signals were obtained according to the respective weighted accumulations of the number of strokes written in the digital cards that occurred within one time window of 10 seconds length [4]—note that the weights for the accumulating values decreased exponentially over time. The handwriting signals were collected in image format, generating one image for every table tablet. Each image was processed to 100×900 pixels—900 pixels relates to the session length, i. e., 150 minutes. During the class, students could write notes at any time on the table tablet, and all of them were required to take notes.

A total of 50 images: 1 image $\times$ 5 table tablets $\times$ 5 sessions $\times$ 2 classroom-groups, were collected. However, since unbalanced student distribution might bias the experimental results, only data from table tablets of discussion sessions with eight participating students were considered. Due to some students' absence, likewise, a



(a) Handwriting Feature Image

(b) Handwriting Feature Image with Data Augmentation

Figure 3: Table tablet images representing handwriting signals. Strokes in cards and time are indicated in the x and y axis respectively: a) gives the original image (with a size of 100×900 pixels); b) gives a new image obtained after applying data augmentation (with a size of 100×810 pixels).

total of 42 images were collected. In Figure 3a, an example of a table tablet image is given. We applied the 'thermal image' concept [18] to the images, i. e., every image's spot indicates the handwriting frequency in the digital cards: high temperature represents high frequency, low temperature indicates low frequency. The English proficiency ground truth was retrieved from the oral presentations. For doing this, the students' speaking performance was evaluated by five English teachers from 0 (poorest) to 5 (best): we computed the arithmetic mean across all raters to infer the gold standard [7] from individual annotations. Since one unique image was collected to represent the handwriting features of the eight students participating in each work group, the averaged English proficiency score across the eight students was also computed to be associated to each image.

3 EXPERIMENTAL SET-UP

3.1 Data Augmentation

Since the use of bigger datasets usually increases machine learning algorithm performance [38], we applied data augmentation to the 42 image samples. By erasing a single rectangular patch with a random size from a given image, Random Erasing [44] has shown to be a method performing well on image data augmentation. Following this method, we randomly erased 10 % of the data, which shows minimal effect w. r. t. the original images: in Figure 3b, an image after data augmentation, presenting a size of 100×810, is displayed. For every image, through the application of the Random Erasing data augmentation method, 20 new samples were generated. In total, 840 samples (20×42) were considered to carry out the experiments.

3.2 Data Partition Set-up

In this study, Leave-One-Subject-Out (LOSO) cross-validation evaluation was performed to satisfy the group independence evaluation constraint [11]. In this context, the 840 samples were divided into 42 group-independent folds, with each fold containing only the handwriting signals from one work group. With the LOSO evaluation scheme, one of the 42 folds was used as test set while the other

Table 1: Regression results (RMSE) obtained from CNN architectures: Shallow (M)odels, with four different (Conv)olutional Layers; Deep (M)odels, with four different Residual Network(ResNet).

Shallow M	Conv2	Conv4	Conv6	Conv8
RMSE Value	0.41	**0.32**	0.36	0.38
Deep M	ResNet18	ResNet34	ResNet50	ResNet101
RMSE Value	0.56	0.45	0.48	0.39

41 folds were considered together as training set. This process was repeated 42 times until all the folds were utilised as test set.

3.3 CNN models

The algorithms on visual recognition currently available are mostly based on deep Convolutional Neural Networks (CNNs) [8, 34], which achieved outstanding performance with small datasets in recent years [33]. When increasing the complexity of CNNs, the training process based on stochastic gradient descent—the multi-layer backpropogation—can easily lead to the gradient 'dispersion' or vanishing gradient. Moreover, there is a phenomenon that the training error increases as the depth increases [13]. The principal of a Residual Network (ResNet) introduces a novel architecture that helps to ease the degradation problem—higher training error when using more layers—and hence allows for training of a very deep network [13]. Thus, in this study, we considered eight CNN models from shallow to deep: for shallow topologies, we consider traditional CNNs with increasing convolutional layer number; for deep architectures, we consider typical ResNet layouts, such as ResNet 18, ResNet 34, and so on. For better comparison, we group our eight CNN models into Shallow Models and Deep Models according to complexity. Following previous work [25], the Root Mean Squared Error (RMSE) is considered as loss function and evaluation metric in all CNN models. The computation of RMSE is defined as:

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}},$$

where n indicates the total number of samples, and $\hat{y}_i$ and y_i represent the prediction and the ground truth for the i-th sample, respectively.

4 RESULTS AND DISCUSSION

The RMSE results indicate the absolute fit of the CNN model to the data, i. e., how close the observed data points are to the model's predicted values [3]. From Table 1, all our results show that using the 3D handwriting dataset can well predict the group work performance in English classroom. In general, Shallow Models achieve better results, only the performance of a CNN model with 2 Convolutional Layers is slightly outperformed by the ResNet-101, (cf. RMSE of 0.41 and 0.39 for Conv2 and ResNet-101, respectively, in Table 1); which might be due to the 'residual' architecture performing well in this case. The best result (RMSE = 0.32) is achieved by a CNN with 4 Convolutional layers, which might be due to the fact that more complex and deep architectures cause severe overfitting during training, which is also observed in previous work [9]. This

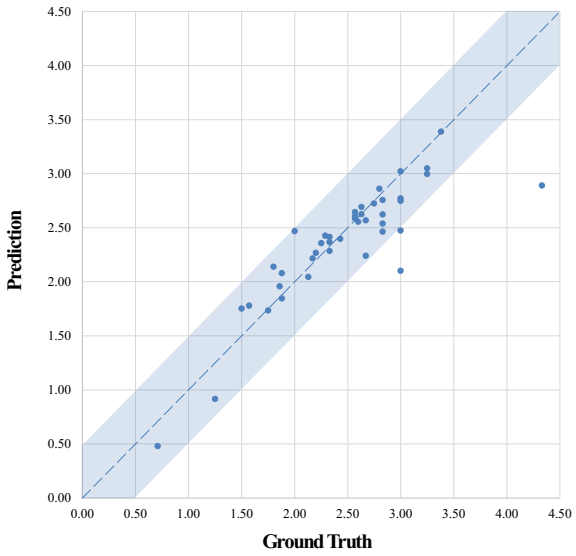


Figure 4: Best RMSE values: ground truth and predictions achieved from 4 Convolutional Layers CNN. Shallow blue area indicates error value less than 0.5.

assumption is indeed supported by the continuous decrease of the training loss, which contrasts with the rapid rise of the test loss. Furthermore, the ResNet Models' training phase is time-consuming and high in computational cost, while Shallow Models, presenting a lightweight architecture, can be more easily optimised, which enables Shallow Models to achieve a better 'rendering' of the 3D handwriting dataset structure.

In the CNN model with 4 Convolutional Layers, i. e., the one which achieved the best results, a max pooling layer after each convolutional layer was considered. To enhance the computing performance of the network, the numbers of channels on four convolutional layers were set to 16, 32, 64, and 64 respectively. Afterwards, the output vectors were flattened and fed into a dense layer, whose output size was set to 1. In Figure 4, the best predictions w. r. t. the ground truth achieved with this CNN model (i. e., Conv4) are indicated. Our experimental results show that the predictions fit the ground truth values in 93 % of the cases: an error score lower than 0.5 is represented with a blue shaded area.

The best experimental result achieved by the CNN with 4 Convolutional Layers, with a RMSE of 0.32 (cf. Table 1), confirms that students' English speaking proficiency can be predicted by using handwriting signal images. Large prediction errors were also observed, especially for ground truth with high scores. For instance, the system predicted a high (proficiency) target score of 4.3 as 2.8, which means that a group of students with high proficiency was identified as having poor speaking skills. This might be due to the limited training data available for high scores—a reason why the model would not have sufficient capability to predict such scores to the best performance of the network architecture. Besides, it might also be due to an exceptional behaviour of this particular group of students, i. e., showing a tendency to write less, despite their high speaking proficiency.

5 CONCLUSION

In this study, we evaluated the relationship between handwriting signals and group work performance. Our experimental results indicate that 3D-handwriting signals are an effective feature to automatically identify group work performance in English classroom. To the best of our knowledge, this is the first study that applies cross-modality analytics to classroom discussion at a group work level. In future work, we plan also to explore internal relationships among other communication modalities, such as gestures, facial expressions, speech, and handwriting signals. Moreover, other kinds of wearables, e. g., watch-type devices equipped with an activity monitor—which might be suitable in describing the evaluated behaviour—will also be taken into account [31, 32]. In addition, we will also assess the performance of other machine learning methods, such as long short term memory models. Given the promising results shown by the use of physical handwriting features in the recognition of students' group work performance, we also plan to collect further work group data from students' interactions within the 'creative digital space'.

ACKNOWLEDGMENTS

This work is funded by the European Union's Horizon 2020 programmes under grant agreement No.,826506 (sustAGE), the Zhejiang Lab's International Talent Fund for Young Professionals (Project HANAMI), P.,R.,China, and the JSPS Postdoctoral Fellowship for Research in Japan (ID No.,P19081). Also from the Japan Society for the Promotion of Science (JSPS), Japan and the Grants-in-Aid for Scientific Research (No.,19F19081 and No.,17H00878) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

REFERENCES

[1] Charles Antaki, Michael Billig, Derek Edwards, and Jonathan Potter. 2003. Discourse analysis means doing analysis: A critique of six analytic shortcomings. *academia.edu* (2003), 1–12.

[2] Eugene Bagdasaryan, Griffin Berlstein, Jason Waterman, Eleanor Birrell, Nate Foster, Fred B Schneider, and Deborah Estrin. 2019. Ancile: Enhancing Privacy for Ubiquitous Computing with Use-Based Privacy. In *Proc. Privacy in the Electronic Society*. London, UK, 111–124.

[3] Tianfeng Chai and Roland R Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* 7, 3 (2014), 1247–1250.

[4] Bin Chen, Koki Hatada, Keiju Okabayashi, Hiroyuki Kuromiya, Ichiro Hidaka, Yoshiharu Yamamoto, and Kazumasa Togami. 2019. Group Activity Recognition to Support Collaboration in Creative Digital Space. In *Proc. Computer Supported Cooperative Work and Social Computing*. Austin, USA, 175–179.

[5] Paul Cobb, Terry Wood, Erna Yackel, and Betsy McNeal. 1992. Characteristics of classroom mathematics traditions: An interactional analysis. *American educational research journal* 29, 3 (1992), 573–604.

[6] Philip R Cohen and Sharon Oviatt. 2017. Multimodal Speech and Pen Interfaces. In *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations*. Vol. 1. 403–447.

[7] Eduardo Coutinho, Florian Hönig, Yue Zhang, Simone Hantke, Anton Batliner, Elmar Nöth, and Björn Schuller. 2016. Assessing the prosody of non-native speakers of english: Measures and feature sets. In *Proc. Language Resources and Evaluation (LREC'16)*. 1328–1332.

[8] Jun Deng, Nicholas Cummins, Jing Han, Xinzhou Xu, Zhao Ren, Vedhas Pandit, Zixing Zhang, and Björn Schuller. 2016. The university of Passau open emotion recognition system for the multimodal emotion challenge. In *Proc. Chinese Conference on Pattern Recognition*. Chengdu, China, 652–666.

[9] Mohammad Sadegh Ebrahimi and Hossein Karkeh Abadi. 2018. Study of residual networks for image recognition. *arXiv preprint arXiv:1805.00325* (2018).

[10] Tobias Fredlund, John Airey, and Cedric Linder. 2012. Exploring the role of physics representations: An Illustrative example from students sharing knowledge about refraction. *European Journal of Physics* 33, 3 (2012), 657.

[11] Jing Han, Kun Qian, Meishu Song, Zijiang Yang, Zhao Ren, Shuo Liu, Juan Liu, Huaiyuan Zheng, Wei Ji, Tomoya Koike, Xiao Li, Zixing Zhang, Yoshiharu Yamamoto, and Björn W Schuller. 2020. An early study on intelligent analysis of speech under COVID-19: severity, sleep quality, fatigue, and anxiety. *arXiv preprint arXiv:2005.00096* (2020), 1–5.

[12] Jing Han, Zixing Zhang, and Bjorn Schuller. 2019. Adversarial training in affective computing and sentiment analysis: Recent advances and perspectives. *IEEE Computational Intelligence Magazine* 14, 2 (2019), 68–81.

[13] Tien Ho-Phuoc. 2018. CIFAR10 to compare visual recognition performance between deep neural networks and humans. *arXiv preprint arXiv:1811.07270* (2018).

[14] Jeremy Hodgen. 2007. Formative assessment. Tools for transforming school mathematics towards a dialogic practice. In *Proc. European Society for Research in Mathematics Education*. Larnaca, Cyprus, 1886–1895.

[15] Yu-Liang Hsu, Cheng-Ling Chu, Yi-Ju Tsai, and Jeen-Shing Wang. 2014. An inertial pen with dynamic time warping recognizer for handwriting and gesture recognition. *IEEE Sensors Journal* 15, 1 (2014), 154–163.

[16] Michael Ignelzi. 2000. Meaning-making in the learning and teaching process. *New directions for teaching and learning* 82 (2000), 5–14.

[17] Fadi Imad, Sharifah Mumtazah Syed Ahmad, Shaiful Hashim, Khairulmizam Samsudin, and Marwan Ali. 2018. Real-Time Pen Input System for Writing Utilizing Stereo Vision. *System* 2 (2018), 1000–1009.

[18] Olivier Janssens, Rik Van de Walle, Mia Loccufier, and Sofie Van Hoecke. 2017. Deep learning for infrared thermal image based machine health monitoring. *IEEE/ASME Transactions on Mechatronics* 23, 1 (2017), 151–159.

[19] Barbara Johnstone. 2018. *Discourse analysis*. John Wiley & Sons.

[20] Hyeon Woo Lee. 2015. Does Touch-based Interaction in Learning with Interactive Images Improve Students' Learning? *The Asia-Pacific Education Researcher* 24, 4 (2015), 731–735.

[21] Hang Li, Yu Kang, Wenbiao Ding, Song Yang, Songfan Yang, Gale Yan Huang, and Zitao Liu. 2020. Multimodal learning for classroom activity detection. In *Proc. International Conference on Acoustics, Speech and Signal Processing*. Onlinestream, 9234–9238.

[22] Zedong Li, Hao Liu, Cheng Ouyang, Wei Hong Wee, Xingye Cui, Tian Jian Lu, Belinda Pingguan-Murphy, Fei Li, and Feng Xu. 2016. Recent advances in pen-based writing electronics and their emerging applications. *Advanced Functional Materials* 26, 2 (2016), 165–180.

[23] Wang Liao, Wei Xu, SiCong Kong, Fowad Ahmad, and Wei Liu. 2019. A two-stage method for hand-raising gesture recognition in classroom. In *Proc. Educational and Information Technology*. Cambridge, UK, 38–44.

[24] Jionghao Lin, Shirui Pan, Cheng Siong Lee, and Sharon Oviatt. 2019. An explainable deep fusion network for affect recognition using physiological signals. In *Proc. International Conference on Information and Knowledge Management*. Suzhou, China, 2069–2072.

[25] Shu Liu, Bo Li, Yang-Yu Fan, Zhe Guo, and Ashok Samal. 2017. Facial attractiveness computation by label distribution learning with deep CNN and geometric features. In *Proc. International Conference on Multimedia and Expo*. Hong Kong, China, 1344–1349.

[26] Jessica M Nolan, Bridget G Hanley, Timothy P DiVietri, and Nailah A Harvey. 2018. She who teaches learns: Performance benefits of a jigsaw activity in a college classroom. *Scholarship of Teaching and Learning in Psychology* 4, 2 (2018), 93.

[27] Sharon Oviatt and Adrienne Cohen. 2013. Written and multimodal representations as predictors of expertise and problem-solving success in mathematics. In *Proc. International conference on multimodal interaction*. Sydney, Australia, 599–606.

[28] Emilia Parada-Cabaleiro, Alice Baird, Nicholas Cummins, and Björn W Schuller. 2017. Stimulation of psychological listener experiences by semi-automatically composed electroacoustic environments. In *Proc. International Conference on Multimedia and Expo*. Hong Kong, China, 1051–1056.

[29] Emilia Parada-Cabaleiro, Anton Batliner, Alice Baird, and Björn Schuller. 2020. The perception of emotional cues by children in artificial background noise. *International Journal of Speech Technology* 23, 1 (2020), 169–182.

[30] Stephen Petrina. 2006. *Advanced teaching methods for the technology classroom*. IGI Global, Vancouver, Canada.

[31] Kun Qian, Hiroyuki Kuromiya, Zhao Ren, Maximilian Schmitt, Zixing Zhang, Toru Nakamura, Kazuhiro Yoshiuchi, Björn W Schuller, and Yoshiharu Yamamoto. 2019. Automatic detection of major depressive disorder via a bag-of-behaviour-words approach. In *Proc. Image Computing and Digital Medicine*. Xi'an, P. R. China, 71–75.

[32] Kun Qian, Hiroyuki Kuromiya, Zixing Zhang, Jinhyuk Kim, Toru Nakamura, Kazuhiro Yoshiuchi, Björn W Schuller, and Yoshiharu Yamamoto. 2019. Teaching machines to know your depressive state: on physical activity in health and major depressive disorder. In *Proc.Engineering in Medicine and Biology Society*. Berlin, Germany, 3592–3595.

[33] Zhao Ren, Qiuqiang Kong, Jing Han, Mark D Plumbley, and Björn W Schuller. 2019. Attention-based Atrous Convolutional Neural Networks: Visualisation and Understanding Perspectives of Acoustic Scenes. In *Proc. International Conference*

on Acoustics, Speech and Signal Processing. Brighton, UK, 56–60.

[34] Maximilian Schmitt and Björn Schuller. 2019. End-to-end audio classification with small datasets–making it work. In *Proc. European Signal Processing Conference*. A Coruña, Spain, 1–5.

[35] Orit Shaer and Eva Hornecker. 2010. Tangible user interfaces: past, present, and future directions. *Foundations and Trends in Human-computer Interaction* 3, 1–2 (2010), 1–137.

[36] Kenneth Silseth and Øystein Gilje. 2019. Multimodal composition and assessment: A sociocultural perspective. *Assessment in Education: Principles, Policy & Practice* 26, 1 (2019), 26–42.

[37] Meishu Song, Zijiang Yang, Alice Baird, Emilia Parada-Cabaleiro, Zixing Zhang, Ziping Zhao, and Björn Schuller. 2019. Audiovisual Analysis for Recognising Frustration during Game-Play: Introducing the Multimodal Game Frustration Database. In *Proc. International Conference on Affective Computing and Intelligent Interaction*. Cambridge, UK, 517–523.

[38] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proc. International Conference on Multimodal Interaction*. Glasgow, UK, 216–220.

[39] Anwar Ali Yahya, Addin Osman, Ahmad Taleb, and Ahmed Abdu Alattab. 2013. Analyzing the cognitive level of classroom questions using machine learning techniques. *Procedia-Social and Behavioral Sciences* 97 (2013), 587–595.

[40] Zijiang Yang, Kun Qian, Zhao Ren, Alice Baird, Zixing Zhang, and Björn Schuller. 2020. Learning multi-resolution representations for acoustic scene classification via neural networks. In *Proc. Sound and Music Technology*. Haerbin, China, 133–143.

[41] Janez Zaletelj. 2017. Estimation of students' attention in the classroom from kinect features. In *Proc. International Symposium on Image and Signal Processing and Analysis*. Ljubljana, Slovenia, 220–224.

[42] Yu Zhang, Fei Qin, Bo Liu, Xuan Qi, Yingying Zhao, and Dan Zhang. 2018. Wearable neurophysiological recordings in middle-school classroom correlate with students' academic performance. *Frontiers in Human Neuroscience* 12 (2018), 457.

[43] Yue Zhang, Felix Weninger, Anton Batliner, Florian Hönig, and Björn Schuller. 2016. Language proficiency assessment of English L2 speakers based on joint analysis of prosody and native language. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 274–278.

[44] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. 2017. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896* (2017).