Confidence Estimation Using Machine Learning in Immersive Learning Environments

Yudong Tao[*], Erik Coltey[†], Tianyi Wang[†], Miguel Alonso Jr.[†], Mei-Ling Shyu[*],
Shu-Ching Chen[†], Hadi Alhaffar[‡], Albert Elias[‡], Biayna Bogosian[‡], Shahin Vassigh[‡]
[*]Department of Electrical and Computer Engineering
University of Miami, Coral Gables, FL 33124, USA
E-mail: {yxt128, shyu}@miami.edu
[†]School of Computing and Information Sciences
Florida International University, Miami, FL 33199, USA
E-mail: {ecolt003, wtian002, malonsoj, chens}@cs.fiu.edu
[‡]College of Communication, Architecture and the Arts
Florida International University, Miami, FL 33199, USA
E-mail: {halhaffa, aelias, bbogosia, svassigh}@fiu.edu

***Abstract*—As the development of Virtual Reality and Augmented Reality (VR/AR) technology rapidly advances, learning in an artificial immersive environment becomes increasingly feasible. Such emerging technology not only facilitates and promotes an efficient learning process, but also reduces the cost of access to learning materials and environments. Current research mainly focuses on the development of immersive learning environments and the adaptive learning methods based on interactions between trainees and the environment. However, valuable human biometric data available in immersive environments, such as eye gaze and controller pose, have not been explored and utilized to help understand the affective state of the trainees. In this paper, we propose a machine-learning based research framework to estimate trainees' confidence about their decisions in immersive learning environments. Using this framework, we designed an experiment to collect biometric data from a multiple-choice question and answer session in an immersive learning environment. This includes collecting answers from 10 participants on 35 questions and their self-reported confidence in their answers. A Long Short-Term Memory neural network model was used to analyze the data and estimate the confidence with 85.6% accuracy.**

Keywords—immersive learning; confidence estimation; immersive environment; deep neural network; machine learning;

I. INTRODUCTION

Immersive environments built with Virtual Reality or Augmented Reality (VR/AR) technology are in the initial phase of development as a way to provide a scalable and affordable learning environment. Early studies have demonstrated a marked increase in trainees' skills, knowledge, and motivation while completing coursework in such environments [1]. Some examples of related VR/AR applications include: an automotive mechanic VR training system to better study complex parts of automobiles [2], motherboard assembly training using adaptive learning and computer vision techniques in AR [3] and providing a guidance system for catheter-based minimally invasive surgery [4], among many others. These training platforms are very hands-on, making VR/AR an attractive alternative to more traditional approaches that involve expert instructors and example/correction activities, which can be both time-consuming and expensive [5].

The ability to put 3D objects in an immersive space and directly manipulate them can enable the exploration of novel ways to present educational content, such as teaching a computer science course as an interactive escape room [6]. Furthermore, it allows for a more intuitive way to visualize concepts that would otherwise be physically impossible or extremely resource-intensive to actually observe [7]. With this, trainees can engage in more hands-on work, for which a lack thereof has been a common complaint among trainees in education software such as Massive Open Online Courses (MOOCs) [8]. Immersive environments have also been found to aid in the trainees' understanding, enabling rapid skill development [3].

In addition to advanced visualization techniques, a VR/AR environment is a well-suited medium to incorporate an Adaptive Learning System (ALS). These systems can create tailored instruction for different types of trainees or allow access to detailed real-time analytics to an instructor, which has initially shown positive effects on trainee learning [9]. While the problem of tailored instruction is considered non-trivial, VR/AR environments allow for the collection of large amounts of both assessment data and environmental/biometric data, which can be used to model and train an ALS.

Many current approaches to creating an ALS for immersive environments involve using on-board sensors on a VR/AR headset, along with external hardware/software for extra sensory input and to run the algorithms that are the core of the ALS. Some examples include leveraging sensor fusion with an on-body sensor network and computer vision module, which is then fed into a petri-net for accurate workflow recognition [5], using a mixture of on-device input such as eye-gaze along with external objects such as Kinect sensors and gloves used across different interaction layers to provide haptic and auditory output [10], and using AR video/tracking to communicate over TCP/IP with a custom tutoring system using ASPIRE [3].

One major building block in an ALS is estimating the trainee's proficiency throughout the provided learning content in order to make decisions. This can be non-trivial, especially in an unstructured immersive environment where a large set of possible inputs need to be accounted for when interacting with the learning content. Paired interaction data have been found to be a good indicator for trainee knowledge in a more traditional environment [11], as well as Item Response Theory (IRT) in the realm of multiple-choice examination [12].

In this paper, we propose a general framework for confidence estimation within an immersive environment. We developed a prototype system using a Magic Leap One headset in the context of a multiple-choice question assessment. Various types of data were collected in the immersive environment, including the layout of the environment, eye-gaze data, controller pose data, and event-driven data from the assessment itself, without any external hardware other than the existing sensor suite on the headset. The collected multimodal data were found to allow for accurate tracking of a trainee's understanding of the content using a Long Short-Term Memory (LSTM) neural network model. The proposed confidence estimation framework has the potential to provide better understanding of the trainees' learning status and accelerate their learning progress.

II. Related Work

A. Learning in Virtual Environment

There is an emerging trend to train professionals in VR/AR immersive environments. This can allow for research work into improving user experience, efficiency, flexibility, and scalability of training for professionals needing new skills in a self-contained package. It can also allow for the democratization of learning, leading to low-cost and effective simulation solutions, which traditionally would have been prohibitively expensive for many educational institutions [1].

Furthermore, emerging AI techniques can enable the integration of adaptive learning (commonly applied to e-learning systems) in an immersive environment. Adaptation can be applied to five core technologies of VR, which are: haptic devices, stereo graphics, adaptive content, assessment, and autonomous agents [10], which would be leveraged to create an ALS. Research into ALSs and Massive Open Online Courses (MOOCs) can help inform an initial framework and highlight what problems can be solved by learning in an immersive environment, along with the difficulties in adapting such a system. Working with labeled knowledge units, learning paths for working through these units, and an AI system for creating these paths are some ideas that could enhance the trainee experience in learning systems like MOOCs [8], which could also be extended to immersive environments.

ALSs typically adjust the training content and difficulty based on a trainee's performance using one or more of the aforementioned techniques. In [2], a performance evaluation module was incorporated in a 3D model assembly training in the VR environment to adjust the difficulty and complexity of the training materials automatically. Similarly, in [3], researchers built an ALS that automatically generated feedback based on trainee's performance during AR motherboard assembly training. The integration of such a training system was shown to improve the efficiency and effectiveness of training in the immersive AR/VR environment.

AI-based computer vision techniques can help recognize the scene and events in the immersive enviornment, allowing for adaptation of training materials to happen in real-time. It can also be used to estimate workflows that trainees are taking for tasks with multiple parts, comparing it to expert instruction and giving feedback accordingly. Westerfield et al. [3] have shown that trainees who learned with context-based materials in a VR/AR environment can achieve better learning performance (25% higher test scores) and efficiency (30% faster speed), compared to video-based materials.

B. Machine Learning on Temporal Data

In immersive environments, various types of data, including biometric data, can be collected continuously as the user interacts with the environment, i.e., data are collected in the form of temporal data [13]. Various machine learning methods can be applied in order to learn patterns and build models based on such data. Hidden Markov Models (HMMs) [14], for example, are one of the most widely applied techniques to model temporal data. For example, human activities can be recognized by an HMM based on the data collected from a set of depth sensors in the context of a smart environment [15].

As high-performance computation accelerators such as Graphical Processing Units (GPUs) become more and more affordable, deep neural network techniques can be explored for analyzing and modeling temporal data. Recurrent neural networks (RNN) [16] have a unique deep neural network architecture designed to learn patterns in a sequence of data. The temporal dependencies among the data are modeled by the parameters in the RNN and are capable of extracting high-level knowledge about the data, which are useful for downstream tasks such as forecasting, temporal data classification, or language modeling. The LSTM [17] is a type of RNN that introduces a three-gate architecture to resolve the vanishing gradient problem typical of most RNN architectures and has shown superior performance in many applications including video classification [18], language modeling [19], and gesture recognition [20].

C. Confidence Estimation

Confidence estimation in the context of trainee understanding, decision-making, and proficiency has been a field of active research for many years, with the goal of producing accurate models of trainee learning upon which an ALS could be built. Bayesian Knowledge Tracing (BKT) has been a mainstay in the field for knowledge modeling, with other algorithms such as the deep learning-based Deep Knowledge Tracing (DKT) being shown to have a substantial performance advantage [21].

Item-response theory was demonstrated as an alternative to grading for gauging trainee learning proficiency, by going

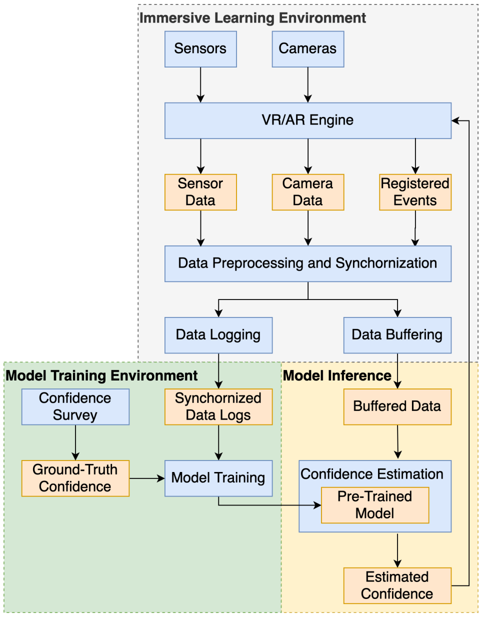


Figure 1. Machine-Learning-Based Confidence Estimation Framework.

down to the question level and using metrics for question difficulty to aid in finding the probability that a trainee will answer it correctly [12]. This technique is well-suited for multiple-choice assessments, wherein the software system could predict how well a trainee will theoretically do on an assessment and can be the basis of an ALS.

Using paired interaction data with tutoring software has also been used to gauge proficiency in a more open-ended environment, such as that of a VR/AR immersive environment. This technique seeks out patterns of behavior that trainees take and determines which ones are indicators of good or bad scores on certain kinds of content using machine learning [11]. This could be very beneficial for more hands-on or laboratory work, which has been the main driver and focus for education within immersive environments.

III. Machine-Learning-Based Confidence Estimation Framework

A. Framework Architecture

Confidence estimation, an essential technique for in-depth understanding of learning status, evaluates whether trainees are confident about their answers or decisions to accomplish an assigned task. It provides cues to better understand why the trainees make decisions and can be used to customize training based on the observations, potentially accelerating the learning progress.

Compared to a conventional learning environment, biometric data, such as eye gaze and hand pose, can be easily collected in an immersive learning environment and can be helpful for confidence estimation. Thus, in this section, we propose a general framework to perform trainee confidence estimation in an immersive learning environment, as well as describe how a machine learning model, the LSTM, can be trained to estimate the trainee confidence for a specific task.

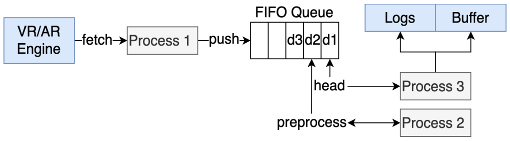


Figure 2. Data Synchonization. Process 1 handles fetching data from VR/AR enginel, Process 2 handles data pre-processing, and Process 3 handles data collection. Process 3 is shared while each type of data has exclusive Process 1 and Process 2.

As shown in Figure 1, the sensor and camera data were collected using a VR/AR engine (Unity) along with a list of registered events in the environment (e.g., a button click). The sensor and camera data are synchronized with the registered event based on the system clock, and after that, the data are preprocessed appropriately. For example, the positions of eye gaze might need to be projected to a canvas in the immersive environment or the registered events might need to be encoded (e.g., one-hot encoding) for dense feature representations.

Once the data have been preprocessed and synchronized, the data can be used for either model training or model inference. To train the trainee confidence estimation model, the synchronized data were logged to disk. A confidence survey was completed right after the users accomplished the test in the immersive environment, where the users were asked how confident they were about their answers. So, the ground-truth confidence was collected from the trainees as a self-reported evaluation and used to train the trainee confidence estimation model using a supervised learning approach. Once a pre-trained model was ready, it was sent to the model inference environment, and the synchronized data was buffered in memory and fed into the pre-trained model, as it was collected. The model inference environment can be the same as the immersive learning environment or a completely different environment. If the trainee confidence estimation model is so large as to be computationally expensive, the model can be deployed as a cloud service and the buffered data can be transmitted to the cloud for trainee confidence estimation.

B. Data Collection

Since the trainee confidence estimation model draws all its input data from the learning environment to train the model and run inference, it is important to make sure that all the collected data are synchronized appropriately. Therefore, each type of data is handled by a separate process in the immersive learning environment to fetch data from the sensor, camera, and monitored events. A first-in-first-out (FIFO) queue is maintained by each process, where the data are pushed into. Once the data are pushed into the queue, the corresponding data preprocessing function is applied to the data and the preprocessing is executed in a separate process. An additional process is used to handle data synchronization and data collection. It pulls data from the queues after the preprocessing and either saves them into the logs for training or puts the processed data into the inference buffer. Figures 2 and 3 show the data synchronization and data collection systems, respectively. The data collection system is primarily meant for temporal sensor data as long as the event logs are triggered by signals in the VR/AR engine.

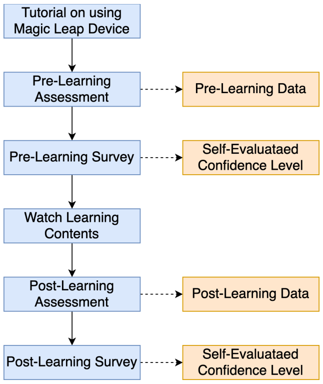


Figure 3. The Overall Framework of Data Collection.

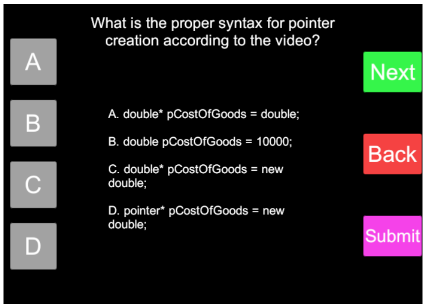
C. *Confidence Estimation in Immersive Environment*

RNNs and their variants, such as LSTMs, are exceptional neural network architectures for modeling sequential data [22]. In this study, we adopt the basic structure of the LSTM network. More specifically, all the input data forms a feature embedding vector that is fed into a neural network with *L* LSTM layers, with each layer composed of *K* hidden units. Then, the intermediate outputs are passed through a fully connected (FC) layer, followed by a dropout layer with 0.5 probability. The activation functions used for the LSTM and FC layers are the Hyperbolic Tangent (Tanh) and Rectified Linear Unit (ReLU), respectively. Then, the second FC (final) layer outputs the probability score generated by a Softmax activation function. Binary cross-entropy is used as the loss function since we are working on a binary classification problem, with the target label being either confident ($Y=1$) or not confident ($Y=0$).

D. *Adaptive Learning based on Confidence Estimation*

Once the LSTM model is trained, the confidence can be estimated while trainees are learning in the immersive environment in real-time. Therefore, the immersive learning system will be able to know not only whether the decisions made by the trainees are correct or not but also how confident they are about their decisions. Based on the estimated confidence, the training can be better customized to accelerate the learning progress [23].

By understanding the confidence of the user for a particular task or skill, the system could optimize the length of time needed for the associated task. For example, if the user has high confidence for a specific task or skill, that task could be sped up or potentially switched out for another task that the user has less confidence on. Conversely, if the user has low confidence for a specific task or skill, that task could be slowed down, repeated multiple times throughout a lesson, or be presented in alternative formats to optimize learning and increase user confidence. Confidence in the skills needed by professionals in Architecture, Engineering, and Construction is crucial as the building industry has a tremendous responsibility for the safety and lives of everyone who engages with a built structure. Additionally, when working with heavy and potentially dangerous machinery, including industrial robotics, confidence in the skills and tasks involving these technologies is crucial for the safety of the operator and those around them.



What is the proper syntax for pointer creation according to the video?

A. double* pCostOfGoods = double;

B. double pCostOfGoods = 10000;

C. double* pCostOfGoods = new double;

D. pointer* pCostOfGoods = new double;

Figure 4. The UI Design for Multiple-Choice Question Answering.

IV. EXPERIMENT SETUP AND RESULTS

A. *Experiment Environment*

To validate the effectiveness of the proposed trainee confidence estimation framework, a multiple-choice question and answering assessment application and data collection protocol were implemented for the immersive environment. In this experiment, the immersive environment was created in Unity (version 2019.2.18) and the Magic Leap One Creator Edition with Magic Leap SDK (version 0.22.0). As shown in Figure 4, the immersive environment uses a flat UI plane for the experiment, which consists of a screen showing all of your options (A, B, C, D) on the left, and control buttons to navigate to the adjacent questions or submit the answer on the right. The questions and possible responses were loaded into the environment using a JSON file.

Once the assessment was completed, the users were asked to take a survey and self-evaluate whether they were confident about their answers in the environment. After that, the users watched several learning videos which contain all of the answers to the questions in the assessment. In the end, a post-learning assessment using the same questions was presented in the immersive environment, followed by another survey to self-evaluate their confidence level about their answers in the post-study assessment.

B. *Data Collection Protocol and CEIE Dataset*

To estimate the confidence in the context of multiple-choice question answering, 10 participants were exposed to the experiment environment and answered 35 questions using the Magic Leap One Creator Headset in both pre- and post-learning sections.

The raw data collected from the AR device include fixation data, pointer data, and assessment data.

- **Fixation Data**: includes the normalized world position of eye gaze fixation, eye gaze confidence, and time stamps.
- **Pointer Data:** includes the normalized world plane position of the pointer, whether the button is pressed, and time stamps.
- **Assessment Data**: time-stamped data that is saved whenever a trainee answers a question or clicks any buttons in the assessment, along with the correct answers.

TABLE I. STATISTICS OF CEIE DATASET.

Dataset		Training	Test
Before Augmentation	***#Samples***	461	127
	#Confident	253	50
	#Not Confident	208	77
After Augmentation	***#Samples***	20003	3854
	#Confident	8756	1117
	#Not Confident	11247	2737

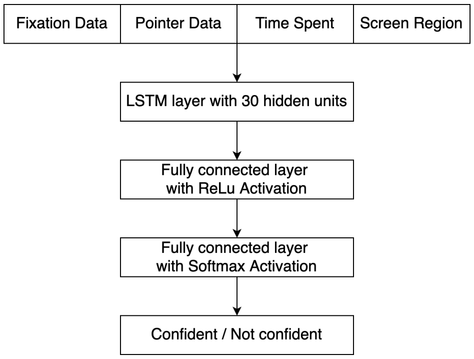

Fixation Data
Pointer Data
Time Spent
Screen Region
LSTM layer with 30 hidden units
Fully connected layer with ReLu Activation
Fully connected layer with Softmax Activation
Confident / Not confident

Figure 5. Confidence Estimation Model for Multiple-Choice Question Assessment.

Both fixation and pointer data are collected at the same time intervals as the assessment data are recorded, whenever a button is pressed and at the end of the assessment.

After that, the world position of fixation data is converted to the gaze fixation relative to the flat plane where the assessment is being shown (as shown in Figure 4), along with data on which part of the assessment your eyes are focused on. The screenshots of the user interface are divided into 11 regions, which include the question title, each of the 4 answer buttons on the left part of the screen, each of the 4 options in the center of the screen, a single area that covers the 3 buttons on the right part of the screen and the background. Furthermore, the part of the gaze is coded using one-hot encoding.

Both types of data are standardized into the coordinates defined by the 700x500 resolution screen. The world position of pointer data is converted in the same manner as well. The length of time it takes a trainee to answer the questions is computed, based on the timestamps of the assessment data. Data is collected once the pre-learning assessment begins and is stopped once it ends. The data is collected again after the user has seen the learning materials and decides to start the post-learning assessment. Right after the pre- and post-assessment, the trainee was asked to take a survey on whether they are confident about their answer to each question, respectively.

Following this data collection protocol, data on 700 questions (10 testers, two tests per tester, and 35 questions per test) were collected, which is called CEIE (Confidence Estimation in the Immersive Environment) dataset. We used 8 trainees' samples for training and 2 trainees' data for testing. To avoid overfitting and improve the model performance, we augment the dataset by truncating each sample based on the shortest sequence length among all the questions. Those samples with shorter sequence lengths are padded with 0. Overall, we collected 23,857 samples. Table I summarizes the statistics of the training and test datasets. Please note that the augmentation procedure is performed on the test dataset in order to maintain consistency with the training dataset. During the testing phase, a majority vote is applied to the results of all truncated samples that belong to the same question to produce the reported question-level result.

TABLE II. PERFORMANCE COMPARISON AMONG ALL COMBINATIONS OF FOUR INPUT MODALITIES ON THE CEIE DATASET. F INCIDATES FIXATION FEATURE, P INDICATES POINTER FEATURE, T INDICATES TIME SPENT FEATURE, AND S INDICATES SCREEN AREA FEATURE.

Modality	Accuracy	F1 Score
F	0.6725	0.5411
P	0.6657	0.5634
T	0.7975	0.7979
S	0.6729	0.5413
F+T	0.8427	0.8459
P+T	0.8108	0.8143
S+T	**0.8560**	**0.8584**
P+S	0.6704	0.5440
F+S	0.6725	0.5411
F+P	0.6661	0.6003

C. Model Training

Based on the CEIE dataset, the proposed LSTM-based confidence estimation model for multiple-choice question assessment was implemented, as shown in Figure 5, where one LSTM layer with 30 hidden units were deployed, followed by two fully connected layers. The input embedding vector concatenates four groups of data: fixation data, pointer data, time spent, and screen region of the fixation. The fixation and the pointer data refer to the positions of the gaze and the pointer on the plane, the time spent data refers to the time taken to answer the question, and the screen region is the one-hot-encoded part of the UI that the gaze focuses at. The model was trained using the Adam optimizer [24] and the learning rate was set to 0.001.

D. Model Performance

Table II illustrates the results of all possible combinations of the four data modalities. It can be noted that among models trained on single modality, the one that used time as the input produced the best score. This is not surprising because, in general, the longer a trainee spends on a question, the less confidence they have. Therefore, time spent is a crucial feature. There are no noticeable performance differences between models that only utilize one of the other three modalities. Generally speaking, models that used time spent as an input feature tended to have better performance than the ones that did not. Among all the models, the one that used both screenshot area and time spent features achieved the best performance with an accuracy of 85.60% and a 0.8584 F1 score. The model which used fixation data and time spent features had slightly worse performance (84.27% accuracy value and 0.8459 F1 scores). The next best performer was the model that used fixation, screen area, and time spent features. It indicates that the combination of time event-based features and context-based features can significantly improve the performance of just using a single modality. All other models demonstrate marginal improvement or even inferior performance than the model which uses only the time spent feature. Therefore, using the time spent model can provide a suitable baseline upon which to base future development.

V. Conclusion and Future Work

In this paper, we introduced a general confidence estimation framework for immersive learning environments, which includes four main components: data collection, data synchronization, confidence estimation, and model training and inference. A prototype system was implemented in an AR environment with the Magic Leap headset for a multiple-choice question assessment. The CEIE dataset was collected from 10 participants and an LSTM model using multi-modality data as the input was trained to estimate the confidence. The model reached and accuracy of 85.6% and F1 score of 0.8584 on the test dataset.

The multiple-choice question assessment was relatively simple, and the current environment only implements a 2D plane for simplicity. In the future, the proposed confidence estimation framework will be evaluated for more complicated tasks in a 3D environment, with a much larger variance in the type of content shown. Also, additional types of biometric data will be explored to improve trainee confidence estimation performance.

Acknowledgment

This research was partially supported by NSF OIA-1937019 and USDA NIFA 2017-67018-26229.

References

[1] S. Hauze and J. Frazee, "Virtual Immersive Teaching and Learning: How Immersive Technology is Shaping the Way Students Learn," in Proceedings of EdMedia + Innovate Learning, 2019, pp. 1445-1450.

[2] W. X. Quevedo, J. S. Sánchez, O. Arteaga, M. Á. V., V. D. Zambrano, C. R. Sánchez, and V. H. Andaluz, "Virtual reality system for training in automotive mechanics," in 4th International Conference on Augmented Reality, Virtual Reality, and Computer Graphics, 2017, pp. 185–198.

[3] G. Westerfield, A. Mitrovic, and M. Billinghurst, "Intelligent augmented reality training for motherboard assembly," International Journal of Artificial Intelligence in Education, vol. 25, no. 1, pp. 157–172, 2015.

[4] G. Turini, S. Condino, U. Fontana, R. Piazza, J. E. Howard, S. Celi, V. Positano, M. Ferrari, and V. Ferrari, "Software Framework for VR-Enabled Transcatheter Valve Implantation in Unity," in 6th International Conference on Augmented Reality, Virtual Reality, and Computer Graphics, 2019, pp. 376-384.

[5] D. Gorecky, S. F. Worgan, and G. Meixner, "COGNITO: a cognitive assistance and training system for manual tasks in industry," in European Conference on Cognitive Ergonomics, 2011, pp. 53–56.

[6] S. Bolivar, D. Perez, A. Carrasquillo, A.S. Williams, N.D. Rishe, F.R. Ortega, "3D Interaction for Computer Science Educational VR Game," in International Conference on Human-Computer Interaction, 2019, pp. 408-419.

[7] S. Greenwald, W. Corning, M. Funk and P. Maes, "Comparing Learning in Virtual Reality with Learning on a 2D Screen Using Electrostatics Activities," Journal of Universal Computer Science, vol. 24, no.2, pp. 220-245, 2018.

[8] H. Yu, C. Miao, C. Leung, and T.J. White. "Towards AI-powered personalization in MOOC learning," NPJ Science of Learning, vol. 2, no. 15, pp. 1-5, 2017.

[9] K. Holstein, B.M. McLaren, V. Aleven, "Student Learning Benefits of a Mixed-Reality Teacher Awareness Tool in AI-Enhanced Classrooms," in International Conference on Artificial Intelligence in Education, 2018, vol. 10947, pp. 154-168.

[10] N. Vaughan, B. Gabrys, and V. N. Dubey, "An overview of self-adaptive technologies within virtual reality training," Computer Science Review, vol. 22, pp. 65–87, 2016.

[11] A. Rafferty, J. Davenport, and E. Brunskill. "Estimating student knowledge from paired interaction data," In Educational Data Mining, 2013, pp. 260-263.

[12] J. Johns, S. Mahadevan, and B. Woolf. "Estimating student proficiency using an item response theory model," In International Conference on Intelligent Tutoring Systems, 2006, pp. 473-480.

[13] S. Pouyanfar, Y. Yang, S.C. Chen, M.L. Shyu, and S.S. Iyengar, "Multimedia Big Data Analytics: A Survey," ACM Computing Surveys, vol. 51, no. 1, pp. 10:1-10:34, 2018.

[14] M.J. Beal, Z. Ghahramani, and C.E. Rasmussen. "The infinite hidden Markov model," in Advances in Neural Information Processing systems, 2001, pp. 577-584.

[15] A. Jalal, S. Kamal, and D. Kim. "Human depth sensors-based activity recognition using spatiotemporal features and hidden markov model for smart environments," Journal of Computer Networks and Communications, vol.2016, pp. 1-11, 2016.

[16] S. Pouyanfar, S. Sadiq, Y. Yan, H. Tian, Y. Tao, M.P. Reyes, M.L. Shyu, S.C. Chen, and S.S. Iyengar. "A survey on deep learning: Algorithms, techniques, and applications." ACM Computing Surveys, vol. 51, no. 5, pp. 1-36, 2018.

[17] S. Hochreiter, and J. Schmidhuber. "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[18] A. Graves, S. Fernández, and J. Schmidhuber. "Bidirectional LSTM networks for improved phoneme classification and recognition," in International Conference on Artificial Neural Networks, 2005, vol.2, pp. 799-804.

[19] H. Tian, Y. Tao, S. Pouyanfar, S.C. Chen, M.L. Shyu. "Multimodal Deep Representation Learning for Video Classification," World Wide Web, vol. 22, no. 3, pp 1325-1341, 2019.

[20] G. Zhu, L. Zhang, P. Shen, and J. Song. "Multimodal gesture recognition using 3-D convolution and convolutional LSTM," IEEE Access, vol. 5, pp.4517-4524, 2017.

[21] K.H. Wilson, X. Xiong, M. Khajah, R.V. Lindsey, S. Zhao, Y. Karklin, E.G.V. Inwegen, B. Han, C. Ekanadham, J.E. Beck, N. Heffernan, and M.C. Mozer. "Estimating student proficiency: Deep learning is not the panacea," In Workshop on Machine Learning for Education in Neural Information Processing Systems, 2016, pp. 1-8.

[22] F.A. Gers, J. Schmidhuber, and F. Cummins. "Learning to forget: Continual prediction with LSTM," Neural Computing, vol. 12, no. 10, pp. 2451-2471, 2000.

[23] M. Heilbron and F. Meyniel. "Confidence resets reveal hierarchical adaptive learning in humans," PLoS Computational Biology, vol. 15, no. 4, pp. e1006972, 2019.

[24] D.P. Kingma and J. Ba. "ADAM: A method for stochastic optimization," in International Conference on Learning Representations, 2015, pp. 1-13.