Monitoring Engagement in Online Classes Through WiFi CSI

Vijay Kumar Singh
IIIT Delhi
vijaysi@iiitd.ac.in

Pragma Kar
Jadavpur University
pragyakar11@gmail.com

Ayush Madhan Sohini
IIIT Delhi
ayush19156@iiitd.ac.in

Madhav Rangaiah
IIIT Delhi
madhav19251@iiitd.ac.in

Sandip Chakraborty
IIT Kharagpur
sandipchkraborty@gmail.com

Mukulika Maity
IIIT Delhi
mukulika@iiitd.ac.in

2023 15th International Conference on COMmunication Systems & NETworkS (COMSNETS) | 978-1-6654-7706-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/COMSNETS56262.2023.10041341

***Abstract*—Due to the Covid-19 pandemic, people have been forced to move to online spaces to attend classes or meetings and so on. The effectiveness of online classes depends on the engagement level of students. A straightforward way to monitor the engagement is to observe students' facial expressions, eye gazes, head gesticulations, hand movements, and body movements through their video feed. However, video-based engagement detection has limitations, such as being influenced by video backgrounds, lighting conditions, camera angles, unwillingness to open the camera, etc. In this work, we propose a non-intrusive mechanism of estimating engagement level by monitoring the head gesticulations through channel state information (CSI) of WiFi signals. First, we conduct an anonymous survey to investigate whether the head gesticulation pattern is correlated with engagement. We then develop models to recognize head gesticulations through CSI. Later, we plan to correlate the head gesticulation pattern with the instructor's intent to estimate the students' engagement.**

***Index Terms*—CSI, Engagement detection, Head gesticulations.**

I. Introduction

The Covid-19 outbreak has forced people to move to online spaces either for professional meetings or for attending lectures. The effectiveness of online activities such as online seminars, online courses, etc., depends on the active engagement of users. But due to its online nature, it is challenging to monitor the engagement level during such an activity. Intuitively, during a regular in-person interaction, a user's head gesticulations show an organically evolved similarity with the intent of speech, such as explanation, curiosity, approval, neutrality, etc [1] Active users perform various head gesticulations, such as nodding their heads when they understand a particular concept or agree to a discussion. In this paper, we aim to monitor the engagement level of individuals during an online setup by investigating the correlation between the head gesticulations captured through WiFi CSI and the intent of speech.

Prior work [2] [3] [4] have used visual aspects such as eye gestures, facial expressions, and head gesticulations to estimate the engagement level of students in online classes. However, video-based engagement detection suffers from several limitations, such as poor lighting conditions, wrong camera angles, poor quality of video feed, privacy issues, unwillingness to open the camera, bandwidth constraints, etc. This paper presents the feasibility of estimating engagement levels by recognizing head gesticulations using WiFi signal properties, specifically the Channel State Information (CSI) in the indoor environment. However, the very first research challenge is that engagement level is subjective and varies from student to student. Further, several recent works [5] [6] [7] have used CSI for human activity recognition (HAR), which provides an alternate solution that is less intrusive, ubiquitous, and not bandwidth hungry. However, a significant challenge in a CSI-based HAR system is that since it is trained using CSI data from a controlled environment, it encounters performance degradation when deployed in a natural setting. Therefore, the second research challenge is developing a suitable model that can identify the head gesticulations from CSI data. The third research challenge is to correlate the intent of speech with the head gesticulations to understand the engagement level of individuals. Although there exist works to identify the intent of speech from the acoustic data [8] [9], the models typically return confounding results leading to poor performance of the system. Finally, the fourth challenge is maintaining students' privacy in the proposed system

Considering the above challenges, we propose an automated system to estimate the engagement level. First, we conduct an anonymous survey consisting of several questions related to the correlation between head gesticulations and engagement. A total of 172 people participated in that survey. The survey unveiled that head gesticulations is highly correlated with engagement level. We then create a setup of WiFi based IoT nodes to collect the CSI when users participate in online classes. We recruited 10 student volunteers who regularly participated in online classes while using the setup to collect the CSI data in parallel. We utilize the video feed to annotate the CSI data with different head gesticulations.

This work has the following contributions:

- Utilizing head gesticulations to monitor the engagement level of students in online classes. We conduct an anonymous online survey to unveil that head gesticulations correlate well with engagement.

- Non-intrusive way of determining head gesticulations through WiFi CSI. A proof of concept to recognize head gesticulations through WiFi CSI.

II. Motivation and Survey Assumptions

The development of this work has been motivated by and facilitated through an anonymous survey that aimed at (1) understanding the correlation of different head gestures with the overall engagement of a learner in online courses and (2) proving the primary hypotheses. A total of 172 people from India and Kuwait responded to the survey until now. We intended to involve participants who were students (90.1%) and Scholars/Faculty (7.6%). The rest were IT Professionals. 77.3% of the participants were male, 18% were female, and the rest preferred not to mention their gender. Almost all the participants (99.4%) were highly experienced with online classes, with 52.9% attending online classes more than once a day.

The survey recommended a laptop/ desktop-based system to be developed as 98.3% of the participants preferred using laptops. The necessity of an automated engagement detection system has also been recommended by the fact that 94.2% of the participants mentioned that they perform other activities during online classes like checking phones, playing games, watching videos, sleeping, etc., along with positive activities like taking notes and solving problems.

The following correlation between head gesticulations and engagement has been observed from the survey that also aligns with our initial hypotheses:

- **Visual Context switching is standard in online classes**: The survey revealed that 48.8% of the participants would periodically look away from the screen. Moreover, 17.4% mentioned that visual Context switching could only be avoided for short videos. This observation recommended developing a system independent of visual focus in estimating learners' engagement.
- **Visual attention is not a good indicator of engagement**: While for short online lectures of $5 - 15$ minutes of duration, visual gaze can indicate attention (90.1%), for longer sessions of > 15 minutes, more than 47.7% of the participants mentioned that continuous gazing at the screen could indicate boredom and inattentiveness (inattention blindness [10]).
- **Head gesticulation is an indicator of engagement**: 84.3% of the participants indicates head nodding as active listening and understanding. Further analysis of nodding and its implications revealed that 64% of the participants would nod their heads to the lecturer in agreement/disagreement when engaged in an online lecture. Even if a lecturer can't see the learner's reaction, 46.5% of the learners would certainly, and 29.7% would sometimes nod their head in agreement/disagreement. 54.1% would not nod if the instructor can view the learners, and 69.2% would not nod if the instructor can not see them while they are disengaged.

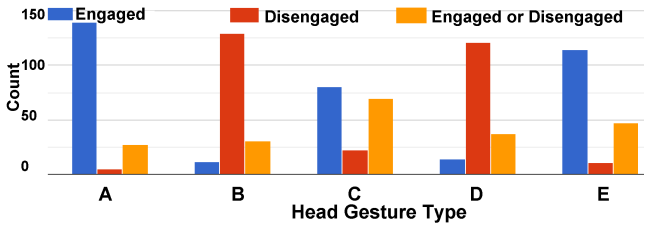

Fig. 1. The correlation between engagement and types of head-gestures– A: Head nodding, B: Downward orientation of head for prolonged time, C: Downward orientation of head for a short time, D: Looking to the left/right frequently, E: Looking to the left/right rarely

For other types of head gestures, the frequent horizontal head movement would indicate distraction (indicated by 77.3% of the participants). However, gazing down can mean engagement or disengagement but will always be associated with multitasking (positive or negative) [11]. We also find that for certain head gestures like vertical movements (gaze down), the correlation with engagement depends on time. For example, a shorter span of gaze down indicates attention while the learner takes notes (48.8%), while prolonged gaze down can indicate disengagement (59.4%). Figure 1 provides a graphical representation of the correlation between head gestures and overall engagement, as reported by the survey participants.

- **Head gesticulation also assist in the manual assessment of attention**: 54.7% and 27.9% of the participants mentioned that head gestures can aid in the manual assessment of learner's engagement always and sometimes, respectively. When asked to rate (1-5) the importance of head gesticulations in understanding the learner's engagement in a non-verbal mode, 41.9% rated it as 4 (very important) and 20.3% rated it as 5 (extremely important).

These findings motivated the development of an engagement detection system that solely depends on head gesticulations, without other additional behavioral traits of the learners

III. Proposed System

In this section, we discuss the details of the proposed system and its workflow (Fig. 2)

A. RAW CSI Data Collection and Denoising

We use two WiFi-enabled ESP32 microcontrollers as shown in Fig. 3 to collect the CSI data. The transmitter (Tx) is analogous to any common WiFi router/modem. The receiver (Rx) is connected to a system running the CSI Tool Kit [12] and an online meeting application. The WiFi standard is IEEE802.11n with $2.4GHz$ channels, and the packet sending rate is 100 packet/sec

We collect raw CSI values of 108 data sub-carriers of $40MHz$ channel. The CSI is given as $X \times N$ matrix, where X is the total number of instances and N is the total number of columns. For each sub-carrier, there are two components in the matrix – the real component (h_{re}) and the imaginary

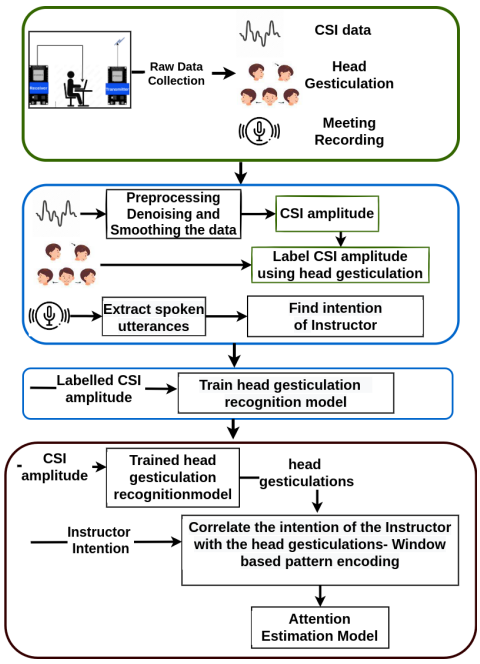

Fig. 2. Proposed workflow diagram.

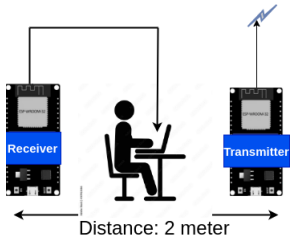

Fig. 3. End-to-end setup to collect the video, meeting Recording and CSI data.

component (h_{img}). We compute the amplitude (A) for each sub-carrier using the following equations.

$$A = \sqrt{(h_{img})^2 + (h_{re})^2}; \quad (1)$$

WiFi CSI is very noisy due to obstacles like stationary objects and people moving around in the environment. Multi-path channels, hardware/software errors, and transmit/receive processing impact the collected CSI data. To remove the high-frequency noise that is an outlier, we use Hampel filter [13]. It replaces the detected outlier data point with the median (m^i). The $1-D$ wavelet transform filter with Daubechies 4 (db4) wavelet [13] is used to reduce the noise due to obstacles. It preserves the head gesticulation information in different scenarios while retaining distinct peaks. Next, we use Savitzky-Golay smoothing filter [13] to smooth the data. It removes the meaningless fluctuations and keeps the essential features.

B. Ground Truth Annotation

We have simultaneously recorded the video feed of the participants for ground truth generation. We implement a model using MediaPipe [14] and Perspective-n-Point (PnP) [15] algorithm to recognize head gesticulations using the video feed. The ground truth head gesticulations are a) **Forward (FR)**, b) **Looking Left (LL)**, c) **Looking Right (LR)**, d) **Looking Up (LU)**, e) **Looking Down (LD)**, f) **Nodding (Nod)**, and g) **Shaking (SK)**. MediaPipe Face Mesh provides a solution to estimate 468 3D face landmarks in real time. First, the face is detected in the video using a face detector, and then a neural network is fed with the cropped image of the face, which determines the location of landmarks. Further, we use the Perspective-n-Point (PnP)algorithm to determine the head's orientation. The model provides *head gesticulations* using these head orientations. For example, nodding involves looking up, looking forward, and looking down, and shaking involves looking right, forward, and left orientations.

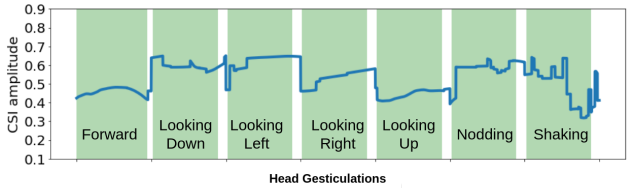

Fig. 4. Change in CSI data due to different head gesticulations

C. Instructor's Speech Intent Detection

We collect the voice recording from the meetings. We implement the window-based text recognition from recording. We timestamped each window with starting and ending timestamps to synchronize with the CSI data. Next, we use a pre-trained intent classification model [16] to recognize the intention of the instructor's sentences/utterances. We consider the intentions that are primarily expressed during a class interaction, such as 'explanation' when the instructor describe any topic and 'asking question' when the instructor asks any question.

D. Head Gesticulation Recognition Model

We use the most fundamental feature extraction method and estimate amplitude from the CSI data using equation (1). We synchronize the head label with the CSI data using timestamps. In this way, we obtain a labeled amplitude dataset. We train the multiclass classification model to classify the head gesticulations using CSI data.

E. Engagement Level Estimation

We will use the window-based head gesticulation pattern matching with the intention to quantify the engagement level. We will deploy the trained classification model to recognize head gesticulation. We will recognize the intention of the instructor, as mentioned in section III-C. The window-based pattern-matching algorithm will quantify the engagement level aligned with the assumption.

IV. Proof of Concept: Head Gesticulation Recognition using CSI Data

A. Data Collection Setup

We recruit 10 participants for data collection. We set up an end-to-end system to conduct online classes, as shown in Fig. 3. In this setup, each volunteer sits between Tx and Rx (Line-of-Sight) and joins the online class in three sessions

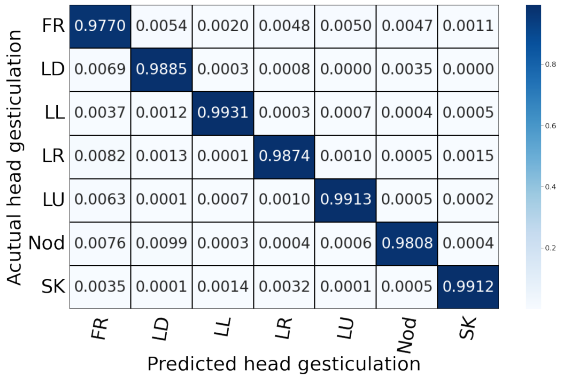

Fig. 5. Classification accuracy of XGB model

TABLE I
PERFORMANCE OF CLASSIFICATION MODEL

Model name	Accuracy	Precision	Recall	F1-score
XGB	**98%**	**97.14%**	**98.71%**	**98%**
SVM	86%	82.14%	90.17%	85.71%
GB	60%	43.25%	50.24%	52.25%
RF	53%	27.85%	26.00%	25.14%
LR	57%	29.57%	50.85%	32.14%

scheduled on different days. The duration of the online class is 1 hour. The lecture content types are categorized into different categories such as length (5 to 45 min), difficulty (easy to advance), and prerequisite (known to unknown). While attending the online classes, students appear for the in-lecture quizzes questions related to the content being covered. After each class, the student appears for an online quiz. The performance of the poll questions reflects visual engagement, and the performance in the post-class quiz reflects cognitive engagement. We collect the raw CSI data, head gesticulations, and video feed of each session. We manually annotate the video feed in terms of engagement level with 4 annotators, which aligns with the observation mentioned in section II.

B. Performance of Recognition Model

We convert the CSI data into amplitude using the equation (1). Fig. 4 shows the changes in the physical property of the CSI data due to head gesticulations. Different head gesticulation has different patterns in CSI data. The classification model needs to learn the different patterns of head gesticulation. The dataset is imbalanced, and "Forward" head gesticulation has the highest number of instances, and "Nodding" and "Shaking" have the lowest number of instances. We implement and compare the performance of Multiclass Logistic Regression (LR), Support Vector Machine (SVM), and the ensemble model includes Random Forest(RF), XGBoost (XGB), and Gradient Boosting (GB). We find the best hyperparameter using grid-search for these models and train all these models. The performance of LR, SVM, RF, GB, and XGB trained with CSI amplitude is listed in Table I. The XGB model outperforms the other classification models. XGBoost performs well with imbalanced data because it performs fine-tuning of extensive hyperparameters and scales imbalanced data. Fig. 5 shows that the XGB model can recognize each head gesticulation accurately.

V. FUTURE WORK

This paper discusses our initial observations towards developing a robust passive sensing platform for student engagement monitoring during an online class. Next, we plan to develop an end-to-end system to estimate the engagement level of students in the online class. A thorough usability study needs to be done to understand the proposed approach's practicality. Nevertheless, the proposed idea can boost researchers' interest in exploring such pervasive sensing techniques to develop a system for promoting active participation in an online class.

REFERENCES

[1] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis," IEEE Trans. Audio Speech Lang. Process., vol. 15, no. 3, pp. 1075–1086, Mar. 2007.

[2] P. Kar, S. Chattopadhyay, and S. Chakraborty, "Gestatten: Estimation of User's Attention in Mobile MOOCs From Eye Gaze and Gaze Gesture Tracking," Proc. ACM Hum.-Comput. Interact., vol. 4, no. EICS, pp. 1–32, Jun. 2020.

[3] M. U. Uçar and E. Özdemir, "Recognizing Students and Detecting Student Engagement with Real-Time Image Processing," Electronics, vol. 11, no. 9, p. 1500, May 2022.

[4] A. R. Basinillo, B. M. Oracion, R. Magno, and L. Vea, "Development of a Model that Detects Student's Disengagement during an Online Lecture Presentation through Eye Tracking and/or Head Movement," in Proceedings of the 2019 International Conference on Information Technology and Computer Communications - ITCC 2019, Singapore, Singapore,pp. 59–65, 2019.

[5] Z. Shi, Q. Cheng, J. A. Zhang, and R. Y. Xu, "Environment-Robust WiFi-based Human Activity Recognition using Enhanced CSI and Deep Learning," IEEE Internet Things J., pp. 1–1, 2022.

[6] J. Su, Z. Liao, Z. Sheng, A. X. Liu, D. Singh, and H.-N. Lee, "Human Activity Recognition Using Self-powered Sensors Based on Multilayer Bi-directional Long Short-Term Memory Networks," IEEE Sensors J., pp. 1–1, 2022.

[7] Z. Hao, D. Zhang, X. Dang, G. Liu, and Y. Bai, "Wi-CAS: A Contactless Method for Continuous Indoor Human Activity Sensing Using Wi-Fi Devices," Sensors, vol. 21, no. 24, p. 8404, Dec. 2021.

[8] B. Sharma, M. Madhavi, and H. Li, "Leveraging Acoustic and Linguistic Embeddings from Pretrained speech and language Models for Intent Classification." arXiv, Feb. 15, 2021.

[9] P. G. Shivakumar, M. Yang, and P. Georgiou, "Spoken Language Intent Detection using Confusion2Vec," in Interspeech 2019, Sep. 2019, pp. 819–823.

[10] A. Mack, "Inattentional blindness: Looking without seeing", Current directions in psychological science, 2003 Oct;12(5):180-4.

[11] P. Kar, S. Chattopadhyay, and S. Chakraborty, "Bifurcating Cognitive Attention from Visual Concentration: Utilizing Cooperative Audiovisual Sensing for Demarcating Inattentive Online Meeting Participants", Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 498, 34 pages, November 2022.

[12] S. M. Hernandez and E. Bulut, "Lightweight and Standalone IoT Based WiFi Sensing for Active Repositioning and Mobility," in 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM), Cork, Ireland, Aug. 2020, pp. 277–286.

[13] Y. Ma, G. Zhou, and S. Wang, "WiFi Sensing with Channel State Information: A Survey," ACM Comput. Surv., vol. 52, no. 3, pp. 1–36, May 2020.

[14] A. M. Al-Nuimi and G. J. Mohammed, "Face Direction Estimation based on Mediapipe Landmarks," in 2021 7th International Conference on Contemporary Information Technology and Mathematics (ICCITM), Aug. 2021, pp. 185–190.

[15] Y. Wu and Z. Hu, "PnP Problem Revisited," J Math Imaging Vis, vol. 24, no. 1, pp. 131–141, Jan. 2006.

[16] "arpanghoshal/EmoRoBERTa · Hugging Face." https://huggingface.co/arpanghoshal/EmoRoBERTa (accessed Nov. 15, 2022).