WIP: Toward Automated Evaluation of Student Presentation Body Language

Yassine Belkhouche
deptpartment of computer science
Missouri State University
Springfield, USA
yassinebelkhouche@missouristate.edu

***Abstract*—This research WIP paper introduces a framework for automating the evaluation of students' class presentations. During oral class presentations, a human judge (mostly a faculty) will evaluate several factors in a short presentation time. This makes the judging task difficult, inconsistent, and highly subjective. Certain evaluation metrics can be assessed using an automated system rather than a human judge. The evaluation system provides objective, consistent, and fair judging of students' presentations. The presenter's body language is among the factors that can be evaluated by such a system. Body language is crucial for communicating messages, emotions, and confidence. Misused body language can convey wrong messages, and be a source of anxiety, distraction, and misunderstanding. The aim is to establish an automated, continuous, computer vision-based system for objective body language/nonverbal presentation skills evaluation. We divide the non-verbal communication skills into four categories: head pose, eye contact, facial expression and emotion, and body pose and gesture. In this paper we focus on head pose as an indicator to evaluate the presenter's eye contact with the audience. We labeled an existing head-pose image dataset into three classes: eye contact (EC), weak eye contact (WEC), and no eye contact(NEC). We designed and trained a neural network model to learn and predict the speaker's eye contact using the extracted features. Finally, we established a scoring system that combines the output of the model to generate a final presentation score. The scoring system is a weighted scheme system, where the user can specify which of the three categories may be weighted higher than the other categories. The proposed system can be used by engineering students to obtain feedback and improve non-verbal presentation and communication skills.**

***Index Terms*—Automatic presentation evaluation, automatic, eye contact assessment, Feature representation, Neural network.**

I. Introduction

Students must acquire and develop numerous crucial skills throughout their college education to master the art of delivering effective presentations. These skills hold significant value for students not only during their college years but also in their future professional endeavors. Eye contact is one of the essential skills needed to deliver successful presentations. Establishing and maintaining continuous eye contact is of significant importance when delivering oral presentations. A recent study shows students' difficulties in using eye contact. Figure 1 shows the percentages of students having difficulties using eye contact when giving presentations [13]. In college, faculty members assess this skill during term or project presentations. The evaluations and feedback given are brief and subjective. In addition to that, this assessment could be tedious and time-consuming. To address these issues, researchers have started investigating automating this process using various multimodal sensing devices. In recent years, advancements in computer vision and machine learning have made it possible to automate the assessment of eye contact skills. The automated system will also offer objective and unbiased evaluations. The rest of this paper is organized as follows. Section II will introduce the literature review associated with face pose detection. Section IV will introduce the methodology, this section will cover the design and training of a neural network model for face pose characterization, face features extraction and model training. The results are discussed in section V.

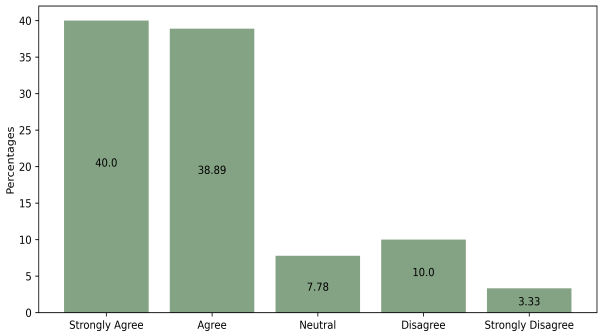


Fig. 1: Students having difficulties using eye contact [13]

II. Literature Review

In recent years, many researchers have been working to establish comprehensive automated systems for oral presentations and social skills evaluation and training. The authors of [2] proposed an automatic evaluation system for social skills training. This system was built to help rehabilitate social skills for people with certain types of disabilities. This system provides the user with an evaluation as well as feedback. The system uses experts labeled multimodal (video, voice, and range data) datasets to train several random forest classifiers. These models learn the most important features associated with social skills. After learning these features, the system provides an evaluation and feedback to the user for future improvement. Chen et al. [3] introduced an approach leveraging multi-modal data for assessing oral presentation delivery

and slide quality. Their study utilized audio, video, and depth modalities to assess six important criteria of presentation performance: speech organization, volume/voice, language, slides, body language, and confidence. This method combines these performance metrics to deliver a comprehensive final score. The authors of the paper [4] introduced a methodology aimed at assessing nonverbal behavioral cues in presentations. They utilized primary sources such as speech, facial expressions (including head and eye gaze movements), and body posture to evaluate presentation competence. Through a thorough examination of nonverbal features extracted from these sources, they proposed a fusion technique to aggregate these features and generate a comprehensive evaluation of presentations. The method proposed by Hincks [5] assesses the student's ability to use their voices in a lively manner. The speaker's voice pitch was analyzed and used as an indicator of the liveliness of the presentation. The method introduced by Echeverria [6] combines both video and Kinect data to evaluate presentation eye contact and body poster. The presenter's performance level is evaluated as good or bad using machine learning and features extracted from both modalities. A deep learning-based framework for presentation assessment was proposed by [8]. Specifically, the authors used a bidirectional long-short term memory model. This framework used two modalities: video and body skeleton generated using depth sensors to provide presentation assessment. Video data representation plays an important role in the accurate assessment of presentation delivery skills. This was emphasized in the work introduced by [9]. The authors used unsupervised learning to generate low-level audiovisual descriptors and self-organizing maps for video classification. Each of these descriptors is used to assess different presentation aspects. The authors introduced a technique to combine all performance measures to provide a final presentation performance score. A method for automatic assessment of public speaking skills using multimodal cues was introduced in [10]. The proposed method used audio, video, and 3D motion capture sensors. The authors established a scoring model using features extracted from speech content, delivery, hand, body, and head movements. A more recent work by [11] used multimodal transfer learning for oral presentation assessment. The proposed method used three types of features: linguistic, acoustic, and visual features. The linguistic features were extracted from the recorded text and represented using word embeddings. The acoustic features were extracted by segmenting the audio files into 5-second tracks and receiving the prosodic, voice quality, and spectrum information from each track. The visual features are 2D face landmarks extracted from face regions such as the eye, mouth, and eyebrows.

III. Problem Statement

Consider the classroom setup depicted in Figure 2, where a student is delivering a class presentation. We aim to explore whether we can establish an automated system to address the following research questions:

- Can the system determine if the presenter is facing the audience and maintaining effective eye contact during his speech?
- Can the system generate an objective assessment score for the presenter's eye contact?
- Can this score be utilized to offer feedback and compare the performance of different presenters?

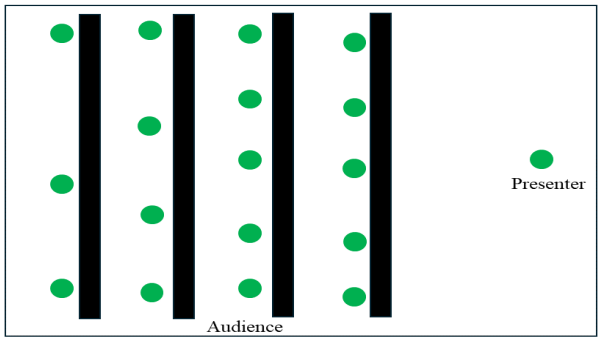


Fig. 2: Problem statement

IV. Proposed Framework

Assessing the presenter's eye contact with the audience is directly linked to the presenter's head pose. Existing methods for eye contact evaluation rely on human evaluators [12]. We established a system capable of identifying three head poses direct eye contact (e.g. facing the audience), weak eye contact (e.g. looking sideways), and no eye contact (e.g., looking up or down). The proposed framework is shown in figures 3, 4, 5 bellow:

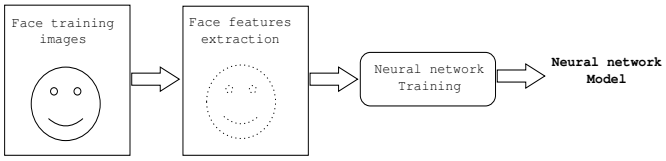


Fig. 3: Neural network training

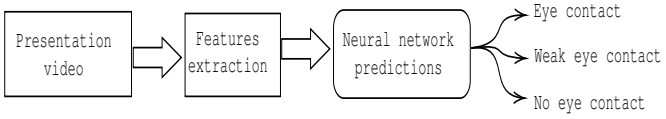


Fig. 4: Neural network prediction

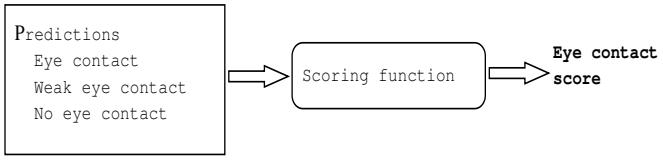


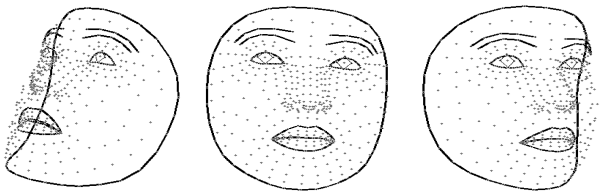
Fig. 5: Eye contact score

The proposed method consists of three steps: recognizing the presenter's head pose, predicting the presenter's head pose,

and assigning a score to the entire presentation. To recognize the user head pose, we designed and trained a neural network using a labeled face pose dataset. After training this model it will be deployed to predict the head pose. Given the prediction results, we used a scoring function to assess user eye contact during the entire presentation. In the remaining of this section, we will discuss each step in details.

A. Face feature extraction

In order to recognize the head/face pose, we extracted 478 facial landmarks from face images using the Mediapipe tool [1]. These face landmarks are represented as 3D points, they are extracted from different face areas such as the iris, eyebrow, mouth, and nose. Each point is represented using the point coordinates (x, y, z), which will create a feature vector of size 1434. Figure 6 shows the landmarks extracted for three faces (no eye contact, good eye contact and weak eye contact).



(a) No eye contact (b) Good eye contact (c) Weak eye contact

Fig. 6: Face feature extraction

B. Neural network model

Figure 7 shows the neural network architecture proposed in this study. This architecture consists of 13 layers with varying numbers of neurons. ReLu activation function shown in equation (1) is applied to all layers except the final layer, which employs SoftMax activation. The total parameter count for the model is 6,520,561. To train the network, the dataset is divided into three folds: 70% for training, 10% for validation, and 20% for testing. The training phase consists of 1500 epochs. The best model is determined by validation accuracy and saved accordingly. Evaluation using the testing dataset reveals the best model achieving an accuracy of 88.42%.

$$ReLu(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \tag{1}$$

$$SoftMax(z) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_i}} \tag{2}$$

C. Training dataset

In this paper, we used the head pose image dataset provided by [7]. This dataset consists of 2,790 face images of 15 individuals, with each person having 93 images of various head poses. The images were captured with pan and tilt angles ranging from -90 to +90 degrees. We categorized the dataset as follows: images with pan and tilt angles between -15 and 15 degrees are labeled as "eye contact"; images with angles between -30 and -15 or +15 and +30 degrees are labeled as "weak eye contact"; and images with angles between -90 and -30 or +30 and +90 degrees are labeled as "no eye contact". Note that this dataset labeling depends on the classroom size, the camera location, and the audience. This labeling could be biased depending on the person performing the labeling. To minimize this bias, several experts could perform the labeling of the dataset, and consider only agreements between these experts.

D. Eye contact score

Given a recorded video presentation containing T frames, we calculate the eye contact score as follows:

$$EYE_{score} = w_1 * D_{EC} + w_2 * W_{EC} + w_3 * N_{EC} \tag{3}$$

Where:

- D_{EC} is the number of presentation frames classified as "Eye Contact" divided by the total number frames in the presentation.

$$D_{EC} = \frac{\#frames = EC}{T} \tag{4}$$

- W_{EC} is the number of presentation frames classified as "Weak Eye Contact" divided by the total number frames in the presentation.

$$D_{EC} = \frac{\#frames = WEC}{T} \tag{5}$$

- N_{EC} is the number of presentation frames classified as "No Eye Contact" divided by the total number frames in the presentation.

$$N_{EC} = \frac{\#frames = NEC}{T} \tag{6}$$

- w_i are weights associated with each class. These weigh could be assigned manually to emphasis the significance/ insignificance of each class.

V. Results

To validate our method, we used the recording of four video presentations. We deployed the learned neural network model to classify each frame into one of three classes: eye contact, weak eye contact, or no eye contact. Figures 8, 9, 10, and 11 show quantitative assessment of these four presentations. They show the number of frames classified as "eye contact", "weak eye contact", and "no eye contact". They also show the percentages of each class in each video. Table I below summarizes the video characteristics and the eye contact score computed using equation 3. In this demonstration the weights w_1, w_2, w_3 are given the values 1, 0.5 and 0.25 respectively.

It is clear from Table 1 and figures 8, 9, 10, 11 that the eye contact scores provide a good reflection of the presenters's eye contact with the audience.

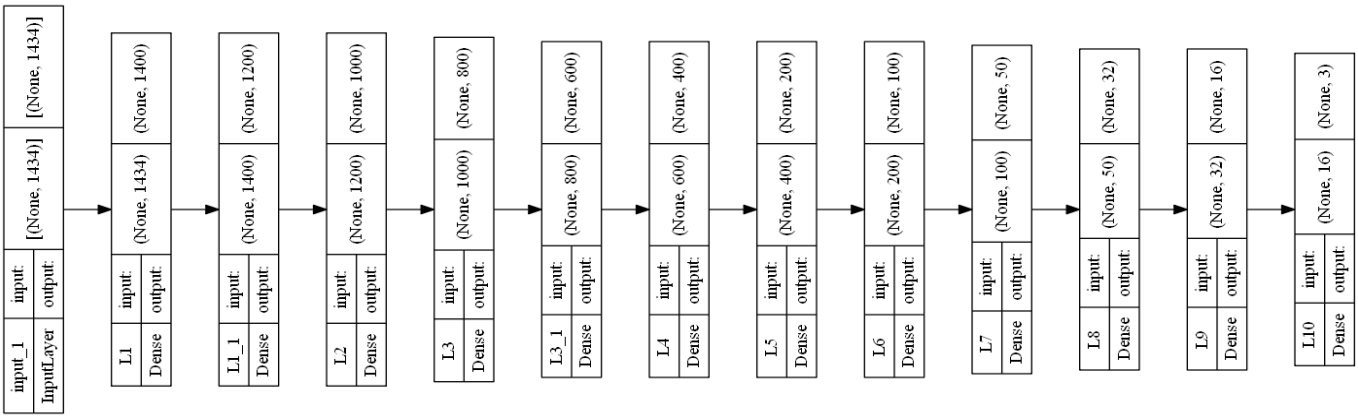
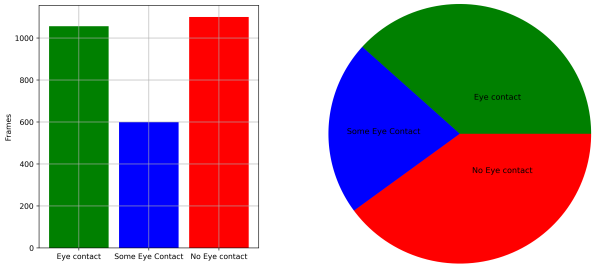


Fig. 7: Neural network architecture

TABLE I: Eye contact score for the four presentations

video	number of frames	eye contact score
1	2756	0.59
2	3024	0.93
3	1422	0.65
4	2745	0.52

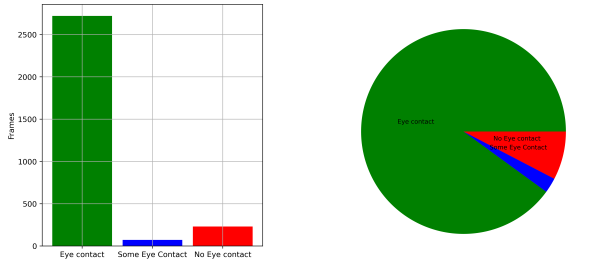


(a) (b)

Fig. 8: Presentation 1: (a) number of predicted frames for each class (b) percentages of frame predictions

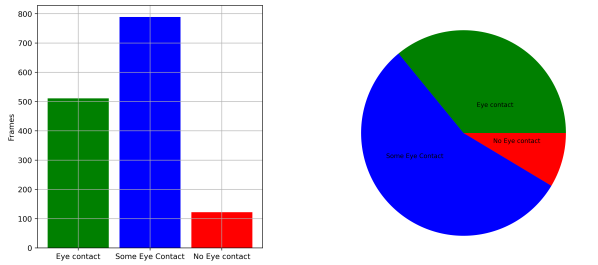
VI. Conclusion and Future Work

In this paper, we designed and trained a neural network model to learn and predict the speaker's eye contact using face-extracted features. The proposed model achieved 88.4 accuracy. This model is used to predict the class of a new frame in a video presentation. We established a scoring system that combines the output of the model prediction to calculate a final presentation score. The scoring system is a weighted scheme system, where the user can specify which of the three categories may be weighted higher than the other categories. In future work, we will consider adding more modalities and evaluate other parameters such as body pose and gestures.



(a) (b)

Fig. 9: Presentation 2: (a) number of predicted frames for each class (b) percentages of frame predictions



(a) (b)

Fig. 10: Presentation 3: (a) number of predicted frames for each class (b) percentages of frame predictions

References

[1] Mediapipe https://ai.google.dev/edge/mediapipe/solutions/guide

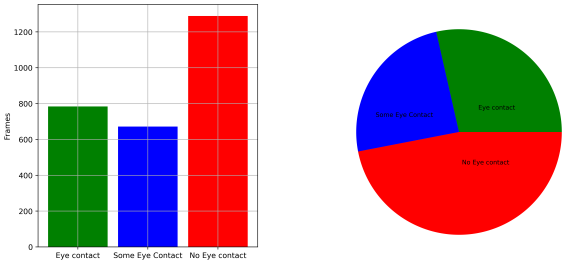
[2] T. Saga and H. Tanaka and Y. Matsuda and T. Morimoto and M. Uratani and K. Okazaki and Y. Fujimoto and S. Nakamura, "Automatic evaluation-feedback system for automated social skills training", Scientific Reports 13, 6856 (2023). https://doi.org/10.1038/s41598-023-33703-0

[3] L. Chen and C. W. Leong and G. Feng and C. M. Lee, "Using Multimodal Cues to Analyze MLA'14 Oral Presentation Quality Corpus: Presentation Delivery and Slides Quality". In Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge (MLA '14). Association for Computing Machinery, New York, NY, USA, 45–52. https://doi.org/10.1145/2666633.2666640

[4] O. Sumer and C. Beyan and F. Ruth and O. Kramer and U. Trautwein and E. Kasneci "Estimating Presentation Competence using Multimodal Nonverbal Behavioral Cues." ArXiv abs/2105.02636 (2021)

[5] R. Hincks, "Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism", System, Volume 33, Issue 4, 2005, Pages 575-591, https://doi.org/10.1016/j.system.2005.04.002.

[6] V. Echeverria, Allan Avendano, Katherine Chiluiza, Anibal Vasquez, and Xavier Ochoa, "Presentation Skills Estimation Based on Video and Kinect Data Analysis", In Proceedings of the 2014 ACM workshop on



(a) (b)

Fig. 11: Presentation 4: (a) number of predicted frames for each class (b) percentages of frame predictions

Multimodal Learning Analytics Workshop and Grand Challenge (MLA '14). Association for Computing Machinery, New York, NY, USA, 53–60. https://doi.org/10.1145/2666633.2666641

[7] N. Gourier and D. Hall and J. Crowley, "Estimating Face orientation from Robust Detection of Salient Facial Structures". FG Net Workshop on Visual Observation of Deictic Gestures, 2004.

[8] J. Li, Y. Wong and M. S. Kankanhalli, "Multi-stream Deep Learning Framework for Automated Presentation Assessment," 2016 IEEE International Symposium on Multimedia (ISM), San Jose, CA, USA, 2016, pp. 222-225, doi: 10.1109/ISM.2016.0051.

[9] F. Haider and M. Koutsombogera and O. Conlan and C. Vogel and N. Campbell and S. Luz, "An Active Data Representation of Videos for Automatic Scoring of Oral Presentation Delivery Skills and Feedback Generation", Frontiers in Computer Science, 2, 2020, doi: 10.3389/fcomp.2020.00001

[10] L. Chen and G. Feng and C. Leong and J. Joe and C. Kitchen and C. M. Lee. "Designing An Automated Assessment of Public Speaking Skills Using Multimodal Cues". Journal of Learning Analytics, 3(2), 261-281, 2026, https://doi.org/10.18608/jla.2016.32.13

[11] S. S. Y. Tun and S. Okada and H. Huang and C. W. Leong, "Multimodal Transfer Learning for Oral Presentation Assessment", in IEEE Access, vol. 11, pp. 84013-84026, 2023, doi: 10.1109/ACCESS.2023.3295832

[12] J. Chiara and H. Roy and R. Johannes and S. Ellen and H. Marij, "The Measurement of Eye Contact in Human Interactions: A Scoping Review", Journal of Nonverbal Behavior, vol 44, doi: 10.1007/s10919-020-00333-3.

[13] Wa Thai Nhu Phuong, Phan Vinh Khang, "Using body language in giving presentations", International Journal Of All Research Writings, vol 4, issue 9, 2023.