Robust Physician Gaze Prediction Using a Deep Learning Approach

Tianyi Tan
College of Computing and Digital Media
DePaul University
Chicago, USA
ttan6@mail.depaul.edu

Enid Montague
College of Computing and Digital Media
DePaul University
Chicago, USA
emontag1@cdm.depaul.edu

Jacob Furst
College of Computing and Digital Media
DePaul University
Chicago, USA
jfurst@cdm.depaul.edu

Daniela Raicu
College of Computing and Digital Media
DePaul University
Chicago, USA
draicu@cdm.depaul.edu

***Abstract*—The patient-physician relationship is an integral part of primary care visits. To build a better relationship, understanding the communication between patient and physician is the key. This study focused on analyzing the gaze, one of the most important non-verbal behaviors found to influence patient outcomes. Gaze analysis often needs a manual rating process which might be time-consuming, costly, and unreliable. This research aimed to support automated analysis of physician-patient interaction using a deep convolutional neural network with transfer learning to a build robust model for physician gaze prediction. Utilizing only 3 minutes of 15 videos capturing 3 physicians interacting with different patients in a clinical setting, the model achieved over 98% accuracy for train, test, and validation sets. By visualizing the convolutional layers and comparing sample frames from different interactions, results highlighted several patterns shared across frames predicted correctly from both seen and unseen video sequences. The proposed work has the potential to informed the future design of technologies used to capture the clinical interaction and provide real-time feedback for physicians, which will contribute to the improvement of care quality.**

***Index Terms*—primary care visits, deep learning, automatic labeling, physician gaze**

I. Introduction

Primary care is an essential part of an effective health care system that emphasizes continuous and preventive care. Effective patient-centered communication is integral to the patient-provider relationship and has been identified as a dimension of physician competency. This research will aim to support automated analysis of primary care visits and can facilitate feedback and reflection systems that help support effective communication, reduce stress and improve the quality of care. This study utilized video/image processing and built convolutional neural network models to recognize patterns of human interactions during primary care visits. By leveraging a large existing dataset of recorded interactions, the research focused on developing methods to help build an automatic annotation tool to extract more information about eye-gaze. A significant body of research of physician behavior analysis depended on an intensive manual annotation process of human coders who rated or annotated live, videotaped or audio clips of interactions based on prescribed methodologies [1]. The practice of manual rating systems is often time-consuming, labor intensive, context dependent and highly subjective to the biases of human annotations [2], [3].The lack of consensus of what to measure and conflicting findings for non-verbal behavior increased the difficulties in quantifying how physician behaviors enhance patient outcomes such as satisfaction and adherence [1]. An effective automated annotation system which provides interaction feedback quickly and reliably may provide more consistent and instructive measurement of physician-patient interactions.

II. Related Work

It is widely accepted that effective interpersonal communication between the physician and patient is essential for patient outcome, such as understanding recommendations for treatment, adherence to therapy, and health outcomes [4], [5]. Most tools evaluating clinician-patient communication were based on verbal cues such as the process analysis system, the verbal response mode, or the Roter Interaction Analysis System (RIAS) but the evaluation of nonverbal interaction has been comparatively less frequent in the literature [6]. However, nonverbal behavior, consisting of three components: the face (e.g. gazing, and smiling), the body (e.g. posture and body orientation), and gesturing (e.g. thumbs up, scratching and clenching) [3], plays an important role in interpersonal judgment which is crucial in physician-patient communication [6]. Eye gaze has been one of the most important cues among all non-verbal behaviors. One study found that the absence of smiling and lack of eye contact were associated with a decrease in physical and cognitive functioning of the

2471-7819/20/$31.00 ©2020 IEEE
DOI 10.1109/BIBE50027.2020.00168

patients [4]. Other research revealed that there was a positive correlation between the eye contact, length of the visit and the patient's perception on clinician empathy [7]. More eye contact with the patient and less gaze at the chart, close proximity, and forward leaning of the physician improved positive patient outcomes such as patient self-disclosure [8]. Therefore, it is important to provide tools to evaluate the non-verbal characteristics of physician-patient communication. The existing tools [9], [10] are relying mainly on human coding to evaluate nonverbal communication in a medical encounter which has been considered as a time-consuming process [2]. From human coding studies that produced evaluation of the reliability of the human coders, the inter-rater correlations ranged between 0.53 and 0.96 for non-verbal activities [2], [7], [10]–[12]. A study also showed that expert annotators on average had higher inter-annotator agreement compared with non-expert annotators evaluated with kappa statistics between the annotations on certain concepts that might not have a clearly defined annotation rule such as aesthetic, quality [13], which was similar to the case in evaluating non-verbal interactions. Also, the study pointed out that annotations based on a majority vote of repeated annotations was able to smooth the noise of human judgement. Thus, it is important to provide automatic annotation tools that can learn from the expert annotation and provide faster, cheaper and more reliable annotations. An increasing number of studies have been using application of machine learning to reduce the burden and reliance on human raters. However, only 8.7% of the studies were related to evaluation of interpersonal and communication skills of the physician, according to a systematic review [14]. A logistic regression classifier was used to predict mutual followership based on the synchrony calculated by optical flow in a simulated medical setting [2]. Built on the methodology of Hart et al. [2], previous work of Gustein, Montague, Furst and Raicu provided physician gaze predictions based upon engineered numerical features obtained from optical flow and body position with an Adaboost algorithm [15]. In previous studies, the classification accuracy on six interactions across three physicians ranged from 80% to 93% on randomly selected test sets. The study proposed utilized the same dataset as the previous study [15] and expanded the dataset by analyzing more interactions for each doctor. There were some limitations of the previous work coming from the dataset and the algorithm. The patient-center videos and doctor-center videos analyzed might provide missing values for body positioning and optical flow measurements due to camera angle issues. The most common issue would be the physician might not be captured by the camera in the patient-center video. The proportion of frames with missing values can range from 0% to 76.24%, which might indicate the model cannot predict over 70% of the frames in extreme cases. This might affect the robustness of the model [15] for unseen videos. In this study, to enhance the robustness and analyze the interactions that had missing values for the previous methodology, only doctor-centered videos which focused upon the physician's face and had reliable camera settings were analyzed with a convolutional neural network. This study also aimed to predict across different interactions and different doctors using only 50% of the video length and providing reliable predictions for unseen sequences. Human activity recognition in video has been explored by various studies using two different methods: handcrafted features (i.e., feature extraction from body and motions) and deep learning learned feature representation [16]. The review pointed out that the deep learning-based solution provided more robust feature extraction and classification in video and benefited real-life applications. Most of the literature used spatiotemporal features to train the network while few used raw video frames as input. A large number of techniques and neural network architectures have been designed and tested on action recognition and human interaction, but they mainly focused on a particular application domain with different activities (e.g. boxing, brushing teeth in UCF 101) per video. Preliminary experiments for the proposed study using complex two-stream models [17] built on one of the largest action recognition dataset UCF-101 [18] showed models designed for categorizing video of human daily tasks were not applicable to the unique primary care visits dataset and prediction of eye gaze. Thus, the study proposed a model using simple architecture with a small dataset obtained from real clinical setting to provide insights in this particular domain.

III. Methodology

A. Data

The research analyzed videos from a dataset that contained raw versions of the videos of clinical interactions of 10 physicians and 101 patients. Fig. 1 shows an example of an interaction captured from the different views (physician-centered, patient-centered, wide-frame, and multi-channel). The 101 videos of clinical interactions had variations in lighting, camera placement and number of people presented in a certain camera view. For each interaction, three cameras with different focuses were set: one lens with patient's chair at the centered (Patient-centered), one lens capturing the face of physician (Physician-centered) and one wide-view lens (Wide-frame). The Multi-channel video was a collection of three videos. Manual annotations encoding physician and patient gaze were obtained using the Noldus Observer XT software [19].

For the visit recordings, human annotations providing information regarding the relative start and stop time for the physician gaze were used to generate ground truth. The start time and stop time were transformed into frame-by-frame label representations for the classification algorithm. Videos were chosen based on following principles and findings. First, the physician was required to be present in front of the camera throughout the entire chosen sequence to provide valid frames for the algorithm. Also, preliminary analysis showed that videos with doctors close to the camera and showing most of the face at the center of the camera were optimal for the analysis. So interactions with sequences satisfying the optimal settings were chosen from three doctors and each contained five interactions. Only Physician-centered videos of the chosen

Fig. 1. Sample Frames of Patient-centered videos, Physician-centered videos, Wide-frame videos, and Multi-channel videos.

interactions were analyzed. Due to the fact that different interactions might have variation in the total visit length and different duration of the video available for analysis, the duration analyzed for each interaction was set to be six minutes. Six minutes was considered to be representative of the non-verbal behaviors during the longer length of the clinical visits based on research findings [20]. Each of the six minutes of the fifteen videos was analyzed from the moment when the physician was present in the Physician-centered video and having interactions with the patient and up to the moment before the start of the physical exam or reaching the desired duration of time. Equal number of frames were extracted from the video at a rate of 29.97 frames per second and aligned with the transformed frame by frame annotation using Avid Media Composer and a preprocessing pipeline built using Python. After mapping the annotations to the frame level, a human annotator confirmed the frame labels for physician gaze for the sequence of interest. The frame label was Patient if the physician was looking at the patient confirmed by both annotation decision made by annotators observing Multi-channel videos and an additional human annotator. Similarly, the frame label was Other if the physician was not considered to be looking at the patient.

B. Deep Learning for Gaze Classification

As defined by [21], the concept of transfer learning consists of two components: domain (a feature space X and a marginal probability distribution P(x)) and task (a label space Y and an objective predictive function $f(\cdot)$). Given a source domain D_S and task T_S, transfer learning helps to improve the learning of $f(\cdot)$ in target domain D_T using the knowledge learned from D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$. With limited sample size of the dataset in this study, transfer learning using a pre-trained model built on much larger datasets with different but related domains and tasks was applied. From preliminary experimental results, VGG-16 proposed by Visual Geometry Group (VGG) [22] performed better and extracted more representative feature vectors for this particular task in this special domain compared to other popular pre-trained models for computer vision tasks such as AlexNet [23], Inception V3 [24] and ResNet50 [25]. VGG-16 utilized 13 convolutional layers and 3 fully connected layers. In this study, the model was trained using extracted features of VGG-16 pre-trained on ImageNet with all blocks frozen and fine-tuned 3,149,825 parameters of an added Global Max Pooling layer, a Dropout layer, and 5 fully connected layers as shown in Fig. 2.

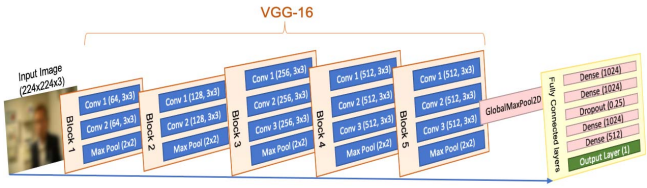

Fig. 2. Model Architecture (first five blocks are frozen layers from VGG-16 architecture (Blue Color), last two blocks are added layers).

The network weights are optimized using the Adam [26] algorithm which is a stochastic gradient descent method with adaptive estimator of lower-order moments with adaptive learning rate. The model was trained to predict binary class label: Patient (the physician was gazing patient) and Other (the physician was gazing elsewhere) at the frame level. Individual Frame extracted from the videos provided information and descriptions about scenes as well as objects. The first three minutes were used for building the model and the remaining three minutes were held out to test the validity of the model for unseen video sequences as shown in Fig. 3. The performance of the model was validated against human annotations.

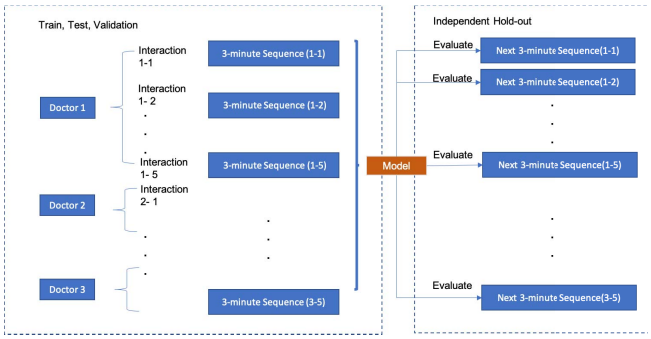

Fig. 3. Experiment Design.

C. Model Building

The first 3 minutes of duration of interests (3-minute Sequence) were used for model building processing. By stratified sampling, 70% of the frames were utilized for training and testing with 67%-33% split while 30% of the frames were used as a validation set for model evaluation. Selected frames were resized to 224x224x3 to be the same as the default input shape for the VGG-16 architecture.

D. Independent Hold-out Set

The remaining 3 minutes of duration for each interaction (Next 3-minute Sequence) were held out as independent hold-

out set for unseen video sequence. It was different from the validation set for model building because it was sampled and evaluated sequentially in time while the validation set was obtained from stratified sampling which might include frames from different time periods of the interaction.

E. Visualization

The study utilized Gradient-weighted Class Activation Mapping (Grad-CAM) [27]to visualize the class discriminative localization map of the last convolutional layer using heatmap visualization, which helped with identifying model bias or data bias in the training set. Grad-CAM can highlight important image regions for prediction.

IV. Results

By stratified sampling based on the class distribution of each interaction, the class distribution of the final train, test and validation were shown in Table I.

TABLE I
CLASS DISTRIBUTION OF TRAIN, TEST AND VALIDATION

	Train	**Test**	**Validation**
Patient	19246	9490	12320
Other	18674	9200	11950

The accuracy of the training was 98.73% (loss: 0.0354), and 98.36% (loss: 0.0559) for test set and 98.31% (loss: 0.0591) for validation set. Table II summarized the model performance in validation set using misclassification matrix and performance metrics including accuracy, sensitivity, precision and F1 score. Fig. 4 showed fast speed of convergence and comparable performance on both train and test sets.

TABLE II
MISCLASSIFICATION MATRIX AND MODEL PERFORMANCE EVALUATION FOR VALIDATION SET

Misclassification Matrix (Validation)			
	Predicted Classes		Total
Observed Classes	Other (0)	Patient (1)	
Other (0)	11749	201	11950
Patient (1)	210	12110	12320
Total	11959	12311	24270
Model Performance Evaluation (Validation)			
Accuracy	Sensitivity	Precision	F1 Score
98.31%	98.30%	98.37%	98.33%

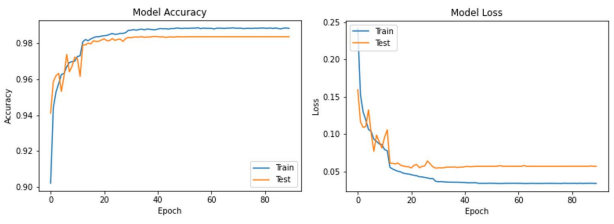

Fig. 4. Model Performance (Training Accuracy (left) and Loss(right)).

For independent hold out sets, the final model was used to predict each frame of the unseen data sequentially. The class distribution of each interaction, percentage of Patient labels among all labels (Pct Dist.) and model performance metrics were shown in Table III. For 9 of the 15 interactions, the accuracy achieved over 90%. Results of sensitivity and precision were also quite high. For doctor 1 interaction 2, the accuracy was 97.83%, with 97.83% sensitivity and 98.73% precision, which was considered to be the set with the best performance. However, for doctor 3 interaction 3, the model performed worst compared to all other interactions for accuracy and sensitivity but good in precision. The frame level inspection of this particular interaction has been provided in the discussion section.

TABLE III
CLASS DISTRIBUTION, PERCENTAGE OF PATIENT LABELS AND MODEL PERFORMANCE METRICS FOR INDEPENDENT HOLDOUT SET .

Interactions	Class Dist.	Model Performance Metrics		
	Patient%	Accuracy	Sensitivity	Precision
Doc1 – 1	42.96%	91.99%	84.85%	96.04%
Doc1 – 2	63.30%	97.83%	97.83%	98.73%
Doc1 – 3	35.58%	84.02%	94.27%	70.64%
Doc1 – 4	62.58%	72.61%	97.10%	70.38%
Doc1 – 5	94.96%	96.11%	99.98%	96.08%
Doc2 – 1	81.48%	96.40%	98.41%	97.21%
Doc2 – 2	48.73%	93.73%	99.47%	88.97%
Doc2 – 3	19.66%	80.59%	92.36%	50.33%
Doc2 – 4	38.25%	92.19%	95.98%	85.42%
Doc2 – 5	31.26%	90.14%	90.93%	80.18%
Doc3 – 1	29.33%	95.07%	92.92%	90.52%
Doc3 – 2	80.34%	91.32%	96.98%	92.58%
Doc3 – 3	71.18%	40.55%	18.21%	91.37%
Doc3 – 4	50.57%	83.22%	90.94%	79.03%
Doc3 – 5	48.49%	74.78%	76.14%	73.01%

Discussion From the model performance of the train, test and validation sets, with 98.31% accuracy in validation set, the model was considered to be robust to predict randomly sampled frames with labels. The independent hold out sets were intended to test the robustness when using the model for the unseen sequence for different interactions of different doctors, which was a unique contribution of the study. 80% of the interactions achieved over 80% accuracy while 75% of those interactions had accuracy higher than 90%, which indicated that the model might be comparable with human annotators for certain interactions. However, for doctor 3 interaction 3 (Doc3-3), with different results compared to all other interactions, might indicate a weakness of the model or the technology design in collecting the video data for the task of interests. Sample frames of train and independent hold out sets for Doc3-3 and doctor 1 interaction 2 (Doc1-2) were visualized and compared using the Grad-CAM technique. Doc1-2 was selected due to the high performance in all metrics. Four correctly classified frames of Doc1-2 of the training set and independent hold-out set for both class labels were shown in Fig. 5. The highlighted areas (considered important regions by the convolutional neural network) were mostly in the facial regions of the doctor such as the forehead, eyes and chin.

In this interaction, the camera captured all the regions of the face and the angle was consistent for the whole duration of interests.

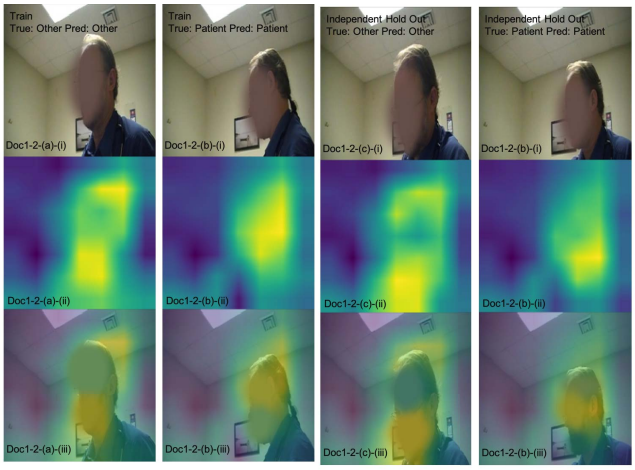

Fig. 5. Sample Frames of Doc1-2, (a) train set (true label is Other), (b) train set (true label is Patient), (c) and (d) for independent hold out set, (i) indicates original images, (ii) indicates map generated by Grad-CAM, (iii) indicates output of overlaying the map with the original image.

Four correctly classified frames of Doc3-3 from the training set and independent hold-out set for both class labels were shown in Fig. 6. Compared to the original frames of the previous doctor, more variability of the face orientation was observed. Also, the doctor might be too close or too far to the camera so that it cannot capture the full facial region. These differences between training and hold out set might bring bias, which might explain why the model cannot generalize well for the unseen sequence of this doctor. Different areas were highlighted such as the neck of the doctor and the collar of his shirt, which might not be informative for all different variations.

To further understand the model performance of Doc3-3, four frames were selected with different face orientations and different parts of the facial region presented in the camera with true label Patient from independent hold out set, as shown in Fig. 7. The four frames were chosen to see the variation of frames and difference in predictions in the hold-out set. The results of the frame of full face and profile face with no eyes informed that the model focused too much on the neck and collar. Body posture and face orientation also seemed to be influential. The frames included in the training set were quite different than those of independent hold-out set in terms of body posture, proximity to the camera and presence of eyes or the entire face region, which might introduce bias in the training. For example, from the observation of the raw data, the face of the physician might not be captured in the later stage of an interaction in the hold-out set, which was not learned by the model from the training set. For a profile face with most of the important facial region presented, the model was able to predict correctly.

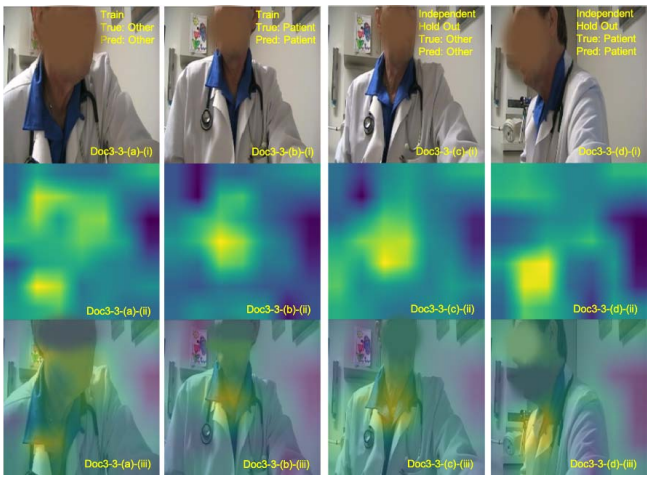

Fig. 6. Sample Frames of Doc3-3, (a) train set (true label is Other), (b) train set (true label is Patient), (c) and (d) for independent hold out set, (i) indicates original images, (ii) indicates map generated by Grad-CAM, (iii) indicates output of overlaying the map with the original image.

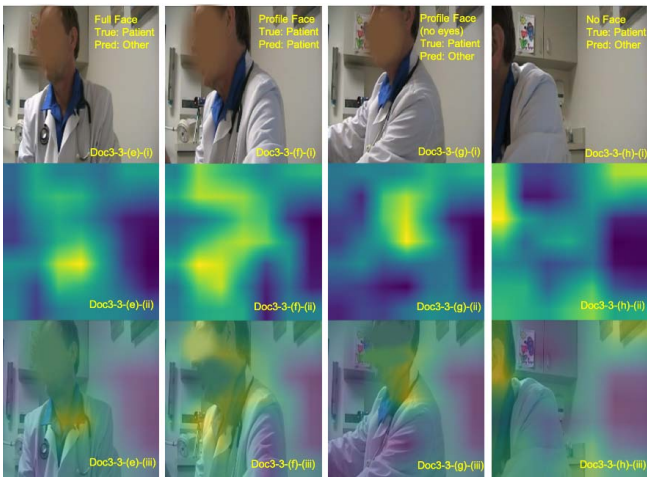

Fig. 7. Selected Frames of Doc3-3 from independent hold out set, (e) indicates a frame with full facial region, (f) indicates a frame with profile face, (g) indicates a frame with no eyes and profile face, (h) indicates no face presented, (i) indicates original images, (ii) indicates map generated by Grad-CAM, (iii) indicates output of overlaying the map with the original image.

From the observation and discussion above, careful planning of camera setting and video-recoding methods would be needed to gather data for gaze prediction using the proposed model in the future. The camera set for capturing the Physician-center video were placed with fixed camera angle on the physician's desk differently for each interaction, which might bring dataset bias based on the limited nature of the dataset used to train and evaluate the model. This might lead to false conclusions based on data collection [28]. More specifically, it brought sample selection bias which was due to differences between training and test collections related to how images were acquired [28]. The finding above might contribute to better technology design for data collection for

this particular gazing prediction task. A study [29] suggested that multiple cameras hooked to walls or side of the desks with remote control that can adjust the cameras' angle in real-time in order to have a rich collection of video data. In future work, detailed technology design suitable for improving the data collection will be provided with more evidence and findings discovered from experiments with more interactions. The findings of this research will assist in developing user-centered design methods including workflow and thematic analysis of both patients and care providers.

V. Conclusion

Incorporating deep learning techniques can help make accurate prediction of physician gaze annotations. This study applied transfer learning to gaze prediction in the clinical setting to assist the human annotation process in this field. By learning from the failed cases, the findings also might help with future design of the data collection technology for an automatic annotation process. With only videos from one camera angle and 50% of the frames annotated, the model can simplify the process and reduce the annotation time for human coders significantly. By extending the prediction to more labels including chart, computer and keyboard, the additional information gathered will further develop the prediction into more meaningful prediction of eye-contact and turn-taking in the future. More generally, automated gaze prediction can assist in studying the performance aspects of physician-patient interactions. Future work includes mapping gazing information to patient ratings, outcomes of physician behaviors as well as measures of physician burnout, which will enhance the understanding of the effects of electronic health records and computers on the physician behavior and further inform the design of the technologies in the clinical context to improve the quality of clinical encounters and reduce physician burnout.

Acknowledgment

This research was supported by NSF Division of Information & Intelligent Systems Award - "CHS: Small: Extracting affect and interaction information from primary care visits to support patient-provider interactions" (Grant No: 1816010).

References

[1] R. S. Beck, R. Daughtridge, and P. D. Sloane, "Physician-patient communication in the primary care office: a systematic review." *The Journal of the American Board of Family Practice*, vol. 15, no. 1, pp. 25–38, 2002.

[2] Y. Hart, E. Czerniak, O. Karnieli-Miller, A. E. Mayo, A. Ziv, A. Biegon, A. Citron, and U. Alon, "Automated video analysis of non-verbal communication in a medical setting," *Frontiers in psychology*, vol. 7, p. 1130, 2016.

[3] M. S. Mast and G. Cousin, "The role of nonverbal communication in medical interactions: Empirical results, theoretical bases, and methodological issues," *The oxford handbook of health communication, behavior change and treatment adherence*, pp. 38–53, 2013.

[4] N. Ambady, J. Koo, R. Rosenthal, and C. H. Winograd, "Physical therapists' nonverbal communication predicts geriatric patients' health outcomes." *Psychology and aging*, vol. 17, no. 3, p. 443, 2002.

[5] A. King and R. B. Hoppe, ""best practice" for patient-centered communication: a narrative review," *Journal of graduate medical education*, vol. 5, no. 3, pp. 385–393, 2013.

[6] M. S. Mast, "On the importance of nonverbal communication in the physician–patient interaction," *Patient education and counseling*, vol. 67, no. 3, pp. 315–318, 2007.

[7] E. Montague, P.-y. Chen, J. Xu, B. Chewning, and B. Barrett, "Nonverbal interpersonal interactions in clinical encounters and patient perceptions of empathy," *J Participat Med*, vol. 5, p. e33, 2013.

[8] M. S. Aruguete and C. A. Roberts, "Participants' ratings of male physicians who vary in race and communication style," *Psychological reports*, vol. 91, no. 3, pp. 793–806, 2002.

[9] E. Krupat, R. Frankel, T. Stein, and J. Irish, "The four habits coding scheme: validation of an instrument to assess clinicians' communication behavior," *Patient education and counseling*, vol. 62, no. 1, pp. 38–45, 2006.

[10] T. A. D'Agostino and C. L. Bylund, "The nonverbal accommodation analysis system (naas): Initial application and evaluation," *Patient education and counseling*, vol. 85, no. 1, pp. 33–39, 2011.

[11] E.-L. Nelson, E. A. Miller, and K. A. Larson, "Reliability associated with the roter interaction analysis system (rias) adapted for the telemedicine context," *Patient education and counseling*, vol. 78, no. 1, pp. 72–78, 2010.

[12] W. M. Caris-Verhallen, A. Kerkstra, and J. M. Bensing, "Non-verbal behaviour in nurse–elderly patient communication," *Journal of advanced nursing*, vol. 29, no. 4, pp. 808–818, 1999.

[13] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 557–566.

[14] R. D. Dias, A. Gupta, and S. J. Yule, "Using machine learning to assess physician competence: A systematic review," *Academic Medicine*, vol. 94, no. 3, pp. 427–439, 2019.

[15] D. Gutstein, E. Montague, J. Furst, and D. Raicu, "Optical flow, positioning, and eye coordination: Automating the annotation of physician-patient interactions," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 943–947.

[16] T. Singh and D. K. Vishwakarma, "Video benchmarks of human action datasets: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 1107–1154, 2019.

[17] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[18] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[19] P. H. Zimmerman, J. E. Bolhuis, A. Willemsen, E. S. Meyer, and L. P. Noldus, "The observer xt: A tool for the integration and synchronization of multimodal signals," *Behavior research methods*, vol. 41, no. 3, pp. 731–735, 2009.

[20] N. A. Murphy, "Using thin slices for behavioral coding," *Journal of Nonverbal Behavior*, vol. 29, no. 4, pp. 235–246, 2005.

[21] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision. 2015," *arXiv preprint arXiv:1512.00567*, 2015.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition. 2015," *arXiv preprint arXiv:1512.03385*, 2016.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[28] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, "A deeper look at dataset bias," in *Domain adaptation in computer vision applications*. Springer, 2017, pp. 37–55.

[29] O. Asan and E. Montague, "Using video-based observation research methods in primary care health encounters to evaluate complex interactions," *Informatics in primary care*, vol. 21, no. 4, p. 161, 2014.