Predicting physician gaze in clinical settings using optical flow and positioning

Arun G. Govindaswamy*, Jacob Furst†, Daniela Raicu‡ and Enid Montague§
College of Computing and Digital Media, DePaul University
Chicago, US
Email: *aarunkarthii@gmail.com, †jfurst@cdm.depaul.edu, ‡draicu@cdm.depaul.edu, §emontag1@cdm.depaul.edu

***Abstract*—Electronic health record systems used in clini- cal settings to facilitate informed decision making, affects the dynamics between the physician and the patient during clinical interactions. The interaction between the patient and the physician can impact patient satisfaction, and overall health outcomes. Gaze during patient-doctor interactions was found to impact patient-physician relationship and is an important measure of attention towards humans and technology. This study aims to automatically label physician gaze for video interactions which is typically measured using extensive human coding. In this study, physicians' gaze is predicted at any time during the recorded video interaction using optical flow and body positioning coordinates as image features. Findings show that physician gaze could be predicted with an accuracy of over 83%. Our approach highlights the potential for the model to be an annotation tool which reduces the extensive human labor of annotating the videos for physician's gaze. These interactions can further be connected to patient ratings to better understand patient outcomes.**

***Index Terms*—physician gaze, primary care visits, patient-physician interaction, healthcare technology, computer vision**

I. Introduction

Recent advancements in health information technology (HIT) in the primary-care settings have both positive and negative impacts on patient care. Electronic health records (EHR) in clinical primary care settings provide accessible and accurate information about the patient to the physician. EHRs supports informed decision-making and medication management. Although some studies find that EHRs reduce medical errors, provide better flow of information and better documentation of patient health records, their presence in the clinical care settings can complicate clinical encounters and impact patient outcomes [1]. Several studies identify negative impact of EHRs on patient-physician interactions. For example, physicians tend to spend more time on technology rather than spending the time with patients [2] [3]. The communication between the patient and the physician is reduced due to the use of computers and adds to mutual silence while documenting [4] - [8]. EHRs can alter the way physicians work – where physicians give their visual attention to the technology present in the clinic rather than eye contact with the patient, potentially affecting the patient's communication with the doctor [5]. EHRs have been identified as an important component in physicians' burnout and affects the physicians to the extent of leaving the practice [9] - [11].

Funding Agency: NSF Division of Information & Intelligent Systems

Better understanding of the physicians' use of technology, and of the patient-physician communication and the associated patient outcomes is paramount. The patient-physician interactions can be categorized as verbal and non-verbal. The components of non-verbal interaction are facial expressions (eyebrow raising, gazing, and smiling), body posture (positioning of arms and legs), and hand gesturing (scratching, thumbs up, hand clenching) [12]. Physician gaze is an important non-verbal feature and patient emotional distress could be identified through higher levels of patient-directed gaze [13]. Identifying physician gaze of recorded patient-physician interaction has traditionally involved manual human coding. Manual video annotations are often time-consuming, labor extensive, context dependent and highly subject to the biases of human annotations [12] - [14]. Hence, this work aims to build a model to retrieve information on physician gaze on a frame level basis. This work can further be expanded to more interactions in the study leading to a robust understanding of patient outcomes in different clinical settings.

II. Related Work

Previous work by Gutstein et al., [15] - [17] used video recorded patient-physician interactions to extract motion information of the physician and the patient through optical flow algorithm [21] and You Only Look Once (YOLO) algorithm [20] to predict physician gaze. Gutstein et al., studied 6 interactions each from 2 doctors and 5 interactions from another doctor adding up to a total of 17 interactions. 3 doctor-specific models were built using an AdaBoost algorithm and reported high performance in predicting physician gaze. The work posed serious limitations due to the nature of clinical settings and camera angle. The most common issue was that of the doctor missing from one of the camera view which resulted in loss of up to 76% of frames from analysis. One other limitation of the work was the performance of these doctor-specific models on interactions from other doctors. Although the doctor specific models presented by Gutstein et al. produced high performing results, these models did not generalize well on clinical interactions which included a different doctor.

This study is an extension of the work done by Gutstein [16] [17]. This work expands the dataset by analyzing interactions from more doctors. To negate the limitation of missing values in 70% of frames and to increase the analysis to more

interactions, the optical flow measurements and the body-positioning coordinates of the doctor present in the patient-centered camera were removed. This study aims to build a generic model which could be used to predict physician's gaze from interactions from other doctors in the study and beyond.

III. Methodology

A. Data

The current data base consists of 101 interactions between the patient and the physician. The study involves 10 doctors and 101 patients which was performed through the University of Wisconsin-Madison at five primary care clinics in 2011 [18]. Every patient in the study agreed to be videotaped and to participate in the study and signed a consent form. The 101 interactions were highly dynamic, as the lighting, camera placement, and number of people fluctuated between each interaction. These 101 interactions were captured using 3 different cameras (Fig. 1) – each placed at different positions and angles in the clinic. Patient-Centered camera – focuses on the patient's chair, Doctor-Centered camera – focuses on the doctor's face and Wide-Angle camera – captures both the patient and the doctor from a wide angle. All these cameras recorded the clinical interactions at 30 frames per second (fps). The Multi-Channel view is a collection of the Patient-Centered, Doctor-Centered and the Wide-Angle frames capturing at a given time. Only the doctor-centered and the patient-centered videos were used in the analysis as the subjects captured using wide-angle camera were at a distant and small optical flow changes could not be captured. The doctor-centered and patient-centered camera focuses on the doctor and the patient respectively capturing subtle optical flow changes.

Further, human encoders annotated the entire duration of the video for each interaction. The manual annotations encoded physician communication, physician gaze, and patient gaze through the Noldus Observer XT software [19]. The start and end time as well as duration were recorded for each of the patient and physician behaviors. There were different annotations determining where the physician gazes at a given time. This study further simplifies the physician's gaze to two levels. If the physician was deemed to be looking at the patient, then it was labeled as Patient. And, if the physician was not deemed to be looking at the patient, then it was labeled as Other. Since our analysis was performed on a frame level basis, all the original annotations were mapped to each frame.

Table I shows the 101 interactions available in the study along with their distribution per doctor, the interactions used in this analysis and the interactions used in previous work by Gutstein et al., [15] [17]. Of the 101 interactions, 18 interactions from 9 doctors were used. To have a consistent number of interactions from each doctor, we choose 2 interactions each from 9 doctors. We set a few guidelines in choosing the interactions - one, the patient stays on the right side and the doctor stays on the left side of the patient-centered camera, two - the doctor's face has to be fully captured by the doctor-centered camera (the doctor tend to move away from the camera during physical examination of the patient). We choose 2 interactions from each doctor which followed these guidelines. Of the 10 doctors, no interaction from doctor #9 followed these guidelines and hence we chose to ignore interactions from doctor #9. Hence, we have 2 interactions each from 9 doctors adding up to a total of 18 interactions.

B. Feature Extraction

Patient-centered patient, patient-centered doctor and the doctor-centered doctor are the three subjects identified from the patient-centered and the doctor-centered videos. We follow the approach used by Gutstein [15] - [17] to extract features such as bounding box co-ordinates of the patient and the physician using You Only Look Once algorithm [20] and optical flow measurements [21]. Optical flow measurements are used to estimate the motion of patient and the physician between successive frames. For each optical flow computation, 15 summary statistic variables were calculated regarding each of the following features – velocityU (x component of velocity), velocityV (y component of velocity), orientation and magnitude. The 15 summary statistics are as follows - maximum, minimum, 25^{th} percentile, 50^{th} percentile, 75^{th} percentile, sum, sum squared, skewness, kurtosis, range, mean, variance, standard deviation, covariance, and non-zero values. The statistic non-zero Values refers to the number of non-zero values for the designated feature in the region of interest (Patient-Centered Physician, Patient-Centered Patient, or Physician-Centered frame) for optical flow measurement. Due to the large number of null optical flow values regarding velocityU, velocityV, orientation, and magnitude, the variables for velocityU, velocityV, orientations and magnitude - other than Non-Zero Values were calculated for the top 25^{th} percentile of feature values with respect to the regions of interest.

Since the patient-centered video captures various static subjects other than the doctor and the patient, we limited the optical flow estimation to two regions of interest in the patient-centered video sequences. The location of the doctor and the patient were identified using the YOLO algorithm. The YOLO algorithm identifies and returns one bounding box around the patient and one bounding box around the physician. Each bounding box has 4 location-based coordinate features – the starting point of the bounding box in the horizontal direction, the starting point of the bounding box in the vertical direction, the width of the bounding box and the height of the bounding box. The bounding box information of the patient-centered physician and patient-centered patient adds up to total of 8 bounding box location-based coordinate features. The optical flow estimates were confined to these two regions. Since the doctor was exclusively present in the doctor-centered video sequence, the optical flow estimates were computed from the entire frame for the doctor-centered physician. In total, 60 optical flow features for each of the regions of interest – Patient-Centered Physician, Patient-Centered Patient and Doctor-Centered Physician - were computed adding up to a total of 180 optical flow features. Figure 2 shows the identified YOLO bounding boxes in the patient-centered image

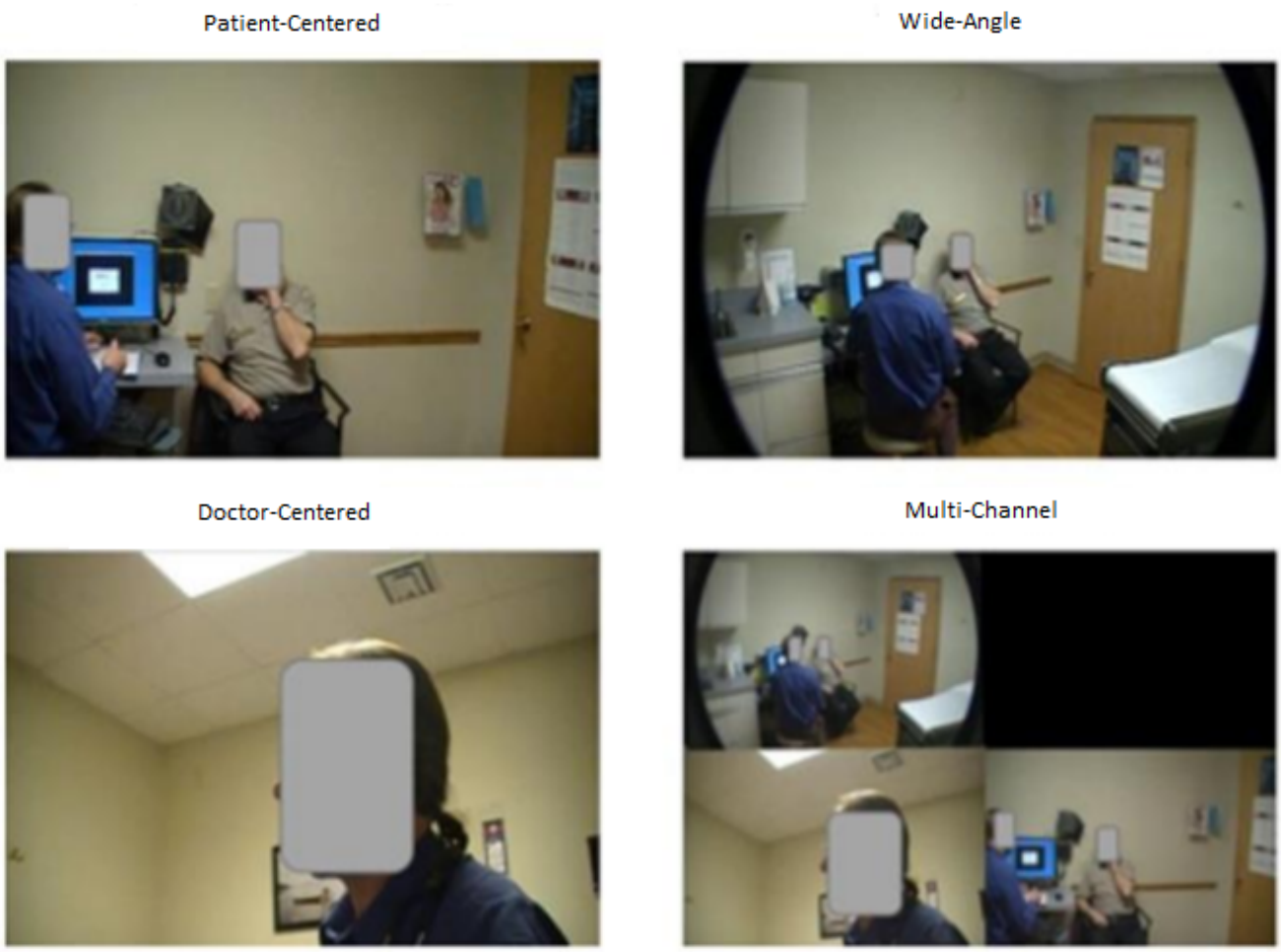

Fig. 1. Interaction video data: example of Patient-Centered, Doctor-Centered, Wide-Angle, and Multi-Channel videos from a particular time [15] [17]

and the optical flow measurements in three different regions of interest. In total, 188 features (180 optical flow measurements + 8 YOLO bounding box values) were extracted from the three subjects in two cameras.

C. Experiment

This project builds on the previous work by Gutstein et al. [15] - [17] and addresses various limitations of its work. Because of camera angle and the nature of the clinical room, the patient-centered doctor is not captured consistently and there occurs serious number of missing values for the patient-centered doctor. To eliminate this limitation, the number of dimensions of the extracted features were reduced from 188 to 124 by removing the location-based coordinate features and optical flow measurements related to patient-centered doctor. Only features related to the patient-centered patient and doctor-centered doctor were used in further model building. In this approach, 5 different experiments were performed. In each experiment, interactions from 8 doctors out of the 9 doctors were used for training, testing and validation of the model. The interactions from the other doctor were used as an additional validation set. In each of the 5 experiments, interactions from a different doctor were used as the additional validation set (Table II). 2 interactions from each of the other 8 doctors were split into training (46%), testing (24%) and validation (30%) data set. For each experiment one random forest model was trained and tuned and the performance of the model is reported in Table IV.

One other aim of the project is to use this model as a tool for annotating the videos for physician's gaze. Another approach was followed where sequences of frames from all the 18 interactions were used for training, testing, and validation. 6 different experiments were performed. In the first 3 experiments, a sequence of 4 minutes was used for training, and a 1-minute sequence of frames was used each for testing and validation. In the other 3 experiments, a sequence of 5 minutes was used for training, and a 30-second sequence of frames was used each for testing and validation. In all these experiments, different combinations of the duration of the video were used for training, testing and validation. The

TABLE I
INTERACTIONS AVAILABLE PER DOCTOR, DATA FOR THIS STUDY AND RELATIONSHIP TO PREVIOUS WORK

Doctor Index	Interaction index distribution per physician	Interactions used in this analysis	Interactions used by Gutstein [15] - [17]
1	01, 02, 59, 66, 67, 73, 74, 89 - 91	01, 02	01, 02, 59, 66, 67, 90
2	03 - 08, 35 - 37	06, 36	-
3	17, 21, 26 - 31, 39, 40	17, 29	-
4	09 - 20, 25	09, 10	-
5	22 - 24, 41 - 43, 51 - 54	41, 42	-
6	32 - 34, 45, 48 - 50, 69, 80, 81	34, 49	-
7	38, 44, 46, 47, 55 - 58, 61, 62	38, 55	-
8	60, 63 - 65, 68, 71, 72, 75, 76, 92	64, 65	60, 63, 64, 65, 68, 75
9	70, 85 - 88, 93 - 97	-	-
10	77 - 79, 82 - 84, 98 - 101	77, 78	77, 78, 84, 98, 101

Fig. 2. An example of a frame with bounding boxes based on YOLO, and marked optical flow vectors in patient-centered and doctor-centered views [15] [17]

TABLE II
INTERACTIONS USED IN THE ADDITIONAL VALIDATION SET FOR EACH EXPERIMENT

Model	Interactions used for additional validation set	
	Doctor	*Interactions*
Model 1	Doctor 1	Interactions 01, 02
Model 2	Doctor 3	Interactions 17, 29
Model 3	Doctor 4	Interactions 09, 10
Model 4	Doctor 6	Interactions 34, 49
Model 5	Doctor 7	Interactions 38, 55

combinations are summarized in Table III and the results of the random forest model is shown in Table V.

TABLE III
DIFFERENT SEQUENCES OF FRAMES USED FOR TRAINING, TESTING AND VALIDATION

Model	Duration of the video used		
	Training	*Testing*	*Validation*
Model 1	00:01 – 04:00	04:01 – 05:00	05:01 – 06:00
Model 2	01:01 – 05:00	00:01 – 01:00	05:01 – 06:00
Model 3	02:01 – 06:00	01:01 – 02:00	00:01 – 01:00
Model 4	00:01 - 05:00	05:01 – 05:30	05:31 – 06:00
Model 5	00:01 – 00:30	00:31 – 05:30	05:31 – 06:00
Model 6	01:01 – 06:00	00:31 – 01:00	00:01 – 00:30

IV. RESULTS

The results of the Random Forest classifier for the methodology discussed above is shown in the Table IV. The models show consistent performance on the training, testing and validation data across models. The models predict physician's gaze on any unseen data within the interactions it was trained on with relatively high accuracy. We show that this model could be used to predict physician's gaze with over 83% accuracy on unseen data. However, the striking part is the performance of the model on the additional validation set. The interactions used for the additional validation set is of a new doctor which was not seen by the model during training. It was learnt through manual analysis of the video interactions that there lie differences in the camera angle, the projections, the objects present in the clinical setting, the difference in the room itself. This significant drop in the performance of the model could be explained that the model did not capture the underlying differences in the clinical setting and the camera projections which is crucial in a computer vision problem. Even though we show poor results on the additional validation set, the performance of the model on validation set is enough to believe that this model could be used to predict physician's gaze on any unseen data within learnt clinical settings and camera projections. We show that to predict the physician's gaze on completely new interaction, a 6-minute video with human annotations on physician's gaze is enough to retrain the model with additional data.

TABLE IV
PERFORMANCE OF THE MODELS ON DIFFERENT DATA SETS

Model	Training	Testing	Validation	Additional Validation
Model 1	98.24%	83.54%	83.60%	41.54%
Model 2	98.43%	83.58%	83.88%	30.46%
Model 3	98.30%	83.75%	84.01%	30.52%
Model 4	98.18%	83.70%	84.01%	42.34%
Model 5	98.29%	83.84%	84.19%	36.22%

The idea of using the model as an annotation tool was further tested and the performance of the models are shown in Table V. The performance of the model on validation set suggests that the hypothesis of using the model as an annotation tool could not be accepted. However, a closer look suggests that with increase in the duration of sequences for training, the performance on the validation set improved. The first 3 models use 4 minutes of sequences for training, whereas the last 3 models use 5 minutes for training. Clearly, the performance of the models on validation set increased. The results suggest the performance on the unseen sequences could be improved when the model learns different motions of the doctor and patient. The model can be used as a tool to annotate further sequences with additional data sequences for training and this remains the limitation of this study.

TABLE V
PERFORMANCE OF THE MODELS ON DIFFERENT SEQUENTIAL DATA SETS

Model	Training	Testing	Validation
Model 1	96.56%	66.19%	58.93%
Model 2	96.38%	64.09%	66.12%
Model 3	96.78%	62.33%	59.53%
Model 4	95.09%	66.86%	65.04%
Model 5	95.10%	62.23%	68.78%
Model 6	95.08%	69.73%	60.01%

V. Comparison and Discussion

In this study, two limitations of previous work by Gutstein were addressed. Gutstein used 6 interactions each from 2 doctors and 5 interaction from other doctor adding up to a total of 17 interaction in his approach. These interactions were used to build doctor-specific models in predicting physician gaze. Although the performance of these models were over 91% accuracy, these doctor-specific models did not generalize well on interactions from other doctors. The performance of the doctor-specific models on interactions from other doctors were as low as 47% which is less than a random guess (Table VI. It could be seen from the table that three different models pertaining to doctor #1, doctor #8 and doctor #10 performed with high accuracy on interactions it was trained on. However, the models did not generalize well on interactions from other doctors. Another limitation of the work was in expanding the work to more interactions. The study used optical flow measurements from patient-centered patient, patient-centered doctor, and doctor-centered doctor. However, the patient-centered doctor were inconsistent across the 101 interactions available in the study. As a result, the interactions where the patient-centered doctor were missing from the camera had to be ignored. This resulted in removing almost 70% of the data in other interactions in the study.

TABLE VI
PERFORMANCE OF THE THREE DOCTOR-SPECIFIC MODELS ON INTERACTIONS FROM OTHER DOCTORS

Dataset	*Doctor #1 Model*	*Doctor #8 Model*	*Doctor #10 Model*
Doctor 1	93.73%	47.02%	45.02%
Doctor 8	47.39%	91.98%	53.29%
Doctor 10	51.04%	48.94%	91.71%

Hence, to negate this limitation, the optical flow measurements and the body-positioning coordinates regarding the patient-centered doctor were removed from analysis. This helped to extend the analysis to many more interactions in the study. Further, the interactions from all the doctors were used to build a random forest model and our results show that our generic model could be used to predict physician gaze with over an accuracy of 83%. The results show that the model can only be used within interactions from doctors it was trained on. The results show that to predict the physician's gaze on completely new interaction, a 6-minute video with human annotations on physician's gaze is required to retrain the model with additional data.

VI. Limitations and Future Work

Retrieving information about the physician's gaze using the motion estimates as shown is a challenging work given the dynamic nature of the video: different clinical settings, different projections of the camera, different patients' motion, and different motion behavior of the doctors. In this study, we show that the information regarding the physician's gaze could be retrieved with 83% accuracy. We also show that this model has the potential to reduce the extensive human labor of annotating the videos for labels. We showed the potential of the model and we show that with additional data we could achieve the aim of using this model as an annotating tool. The usage of 6 minutes from each interaction is seen as a limitation. Another limitation is the performance of the model on completely unseen interaction. These limitations are a work in progress and gives researchers plenty of scope because of the difficulty of the problem and the technological discoveries that could be made on human-computer interaction in primary-care settings.

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] Pelland, K.D., Baier, R.R., and Gardner, R.L.: "It is like texting at the dinner table': a qualitative analysis of the impact of electronic health records on patient–physician interaction in hospitals', BMJ Health & Care Informatics, 2017, 24, (2), pp. 216.

[2] Asan, O., D. Smith, P., and Montague, E.: 'More screen time, less face time – implications for EHR design', Journal of Evaluation in Clinical Practice, 2014, 20, (6), pp. 896-901

[3] Park, S.Y., Lee, S.Y., and Chen, Y.: 'The effects of EMR deployment on doctors' work practices: A qualitative study in the emergency department of a teaching hospital', International Journal of Medical Informatics, 2012, 81, (3), pp. 204-217

[4] Street, R.L., Liu, L., Farber, N.J., Chen, Y., Calvitti, A., Zuest, D., Gabuzda, M.T., Bell, K., Gray, B., Rick, S., Ashfaq, S., and Agha, Z.: 'Provider interaction with the electronic health record: The effects on patient-centered communication in medical encounters', Patient Education and Counseling, 2014, 96, (3), pp. 315-319

[5] Margalit, R.S., Roter, D., Dunevant, M.A., Larson, S., and Reis, S.: 'Electronic medical record use and physician–patient communication: An observational study of Israeli primary care encounters', Patient Education and Counseling, 2006, 61, (1), pp. 134-141

[6] Dowell, A., Stubbe, M., Scott-Dowell, K., Macdonald, L., and Dew, K.: 'Talking with the alien: interaction with computers in the GP consultation', Australian Journal of Primary Health, 2013, 19, (4), pp. 275-282

[7] Alkureishi, M.A., Lee, W.W., Lyons, M., Press, V.G., Imam, S., Nkansah-Amankra, A., Werner, D., and Arora, V.M.: 'Impact of Electronic Medical Record Use on the Patient–Doctor Relationship and Communication: A Systematic Review', Journal of General Internal Medicine, 2016, 31, (5), pp. 548-560

[8] Linzer, M., Poplau, S., Babbott, S., Collins, T., Guzman-Corrales, L., Menk, J., Murphy, M.L., and Ovington, K.: 'Worklife and Wellness in Academic General Internal Medicine: Results from a National Survey', Journal of General Internal Medicine, 2016, 31, (9), pp. 1004-1010

[9] Friedberg, M.W., Chen, P.G., Van Busum, K.R., Aunon, F., Pham, C., Caloyeras, J., Mattke, S., Pitchforth, E., Quigley, D.D., Brook, R.H., Crosson, F.J., and Tutty, M.: 'Factors Affecting Physician Professional Satisfaction and Their Implications for Patient Care, Health Systems, and Health Policy', Rand Health Q, 2014, 3, (4), pp. 1-1

[10] Sinsky, C.A., Dyrbye, L.N., West, C.P., Satele, D., Tutty, M., and Shanafelt, T.D.: 'Professional Satisfaction and the Career Plans of US Physicians', Mayo Clinic Proceedings, 2017, 92, (11), pp. 1625-1635

[11] Babbott, S., Manwell, L.B., Brown, R., Montague, E., Williams, E., Schwartz, M., Hess, E., and Linzer, M.: 'Electronic medical records and physician stress in primary care: results from the MEMO Study', Journal of the American Medical Informatics Association, 2013, 21, (e1), pp. e100-e106

[12] Bensing, J.M., Kerssens, J.J., and van der Pasch, M.: 'Patient-directed gaze as a tool for discovering and handling psychosocial problems in general practice', Journal of Nonverbal Behavior, 1995, 19, (4), pp. 223-242

[13] Cousin, M.S.M.a.G.: 'The Role of Nonverbal Communication in Medical Interactions: Empirical Results Theoretical Bases and Methodological Issues' (2013. 2013)

[14] Hart, Y., Czerniak, E., Karnieli-Miller, O., Mayo, A.E., Ziv, A., Biegon, A., Citron, A., and Alon, U.: 'Automated Video Analysis of Non-verbal Communication in a Medical Setting', Front Psychol, 2016, 7, pp. 1130-1130

[15] Gutstein, D.: 'Information Extraction from Primary Care Visits to Support Patient-Provider Interactions', DePaul University, 2020

[16] Gutstein, D., Montague, E., Furst, J.D., and Raicu, D.S.: 'Hand-Eye Coordination: Automating the Annotation of Physician-Patient Interactions', 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 2019, pp. 657-662

[17] Gutstein, D., Montague, E., Furst, J.D., and Raicu, D.S.: 'Optical Flow, Positioning, and Eye Coordination: Automating the Annotation of Physician-Patient Interactions', 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 943-947

[18] Haskard, K.B., Williams, S.L., DiMatteo, M.R., Heritage, J., and Rosenthal, R.: 'The Provider's Voice: Patient Satisfaction and the Content-filtered Speech of Nurses and Physicians in Primary Medical Care', Journal of Nonverbal Behavior, 2008, 32, (1), pp. 1-20

[19] Zimmerman, P.H., Bolhuis, J.E., Willemsen, A., Meyer, E.S., and Noldus, L.P.: 'The Observer XT: a tool for the integration and synchronization of multimodal signals', Behav Res Methods, 2009, 41, (3), pp. 731-735

[20] Redmon, J., Divvala, S., Girshick, R., and Farhadi, A.: 'You Only Look Once: Unified, Real-Time Object Detection', 'Book You Only Look Once: Unified, Real-Time Object Detection' (2016, edn.), pp. 779-788

[21] Lucas, B., and Kanade, T.: 'An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI)' (1981. 1981)