Predicting physician gaze in clinical settings using optical flow and positioning

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

***Abstract*—Electronic health record systems used in clinical 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, gaze recognition, optical flow**

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].

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. Three doctor- specific models were built using an AdaBoost algorithm and reported high performance in predicting physician gaze. The work posed several 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 views which resulted in loss of up to 76% of frames from analysis and resulted in low generalizability power to other videos capturing interactions of these three doctors with other patients. Another 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.

Funding Agency: NSF Division of Information & Intelligent Systems

978-1-7281-8579-8/20/$31.00 ©2020 IEEE

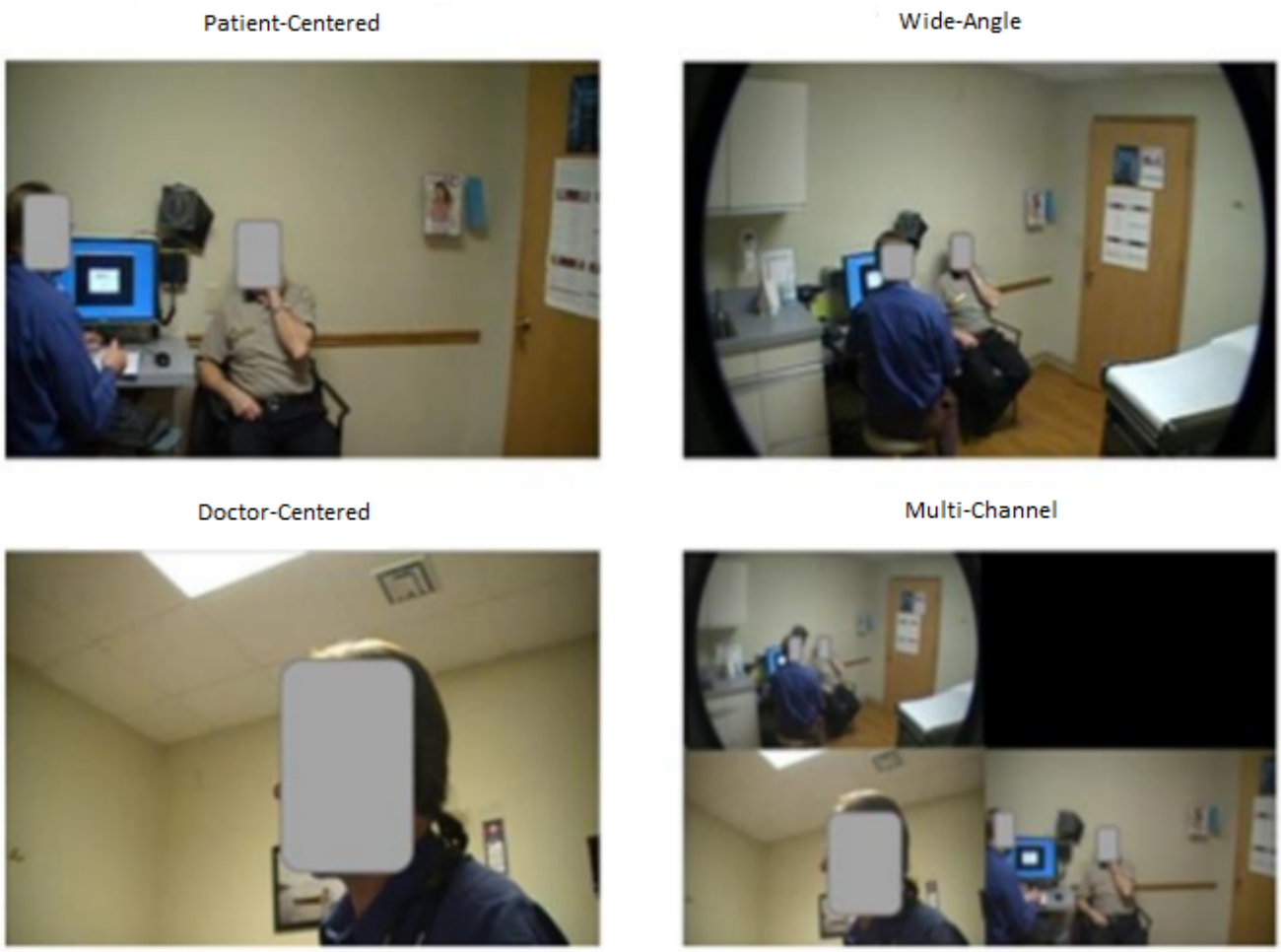

Fig. 1. Interaction video data: example of Patient-Centered, Doctor-Centered, Wide-Angle, and Multi-Channel videos from a particular time [15] [17]

This study is an extension of the work done by Gutstein [16] [17] and addresses the two limitations posed. We address the first limitation by careful analysis of feature importance and removal of doctor specific motion information from one of the camera views. These removed features were found to be redundant by careful feature selection techniques and the removal of these features did not decrease the model performance. By removal of these features, the new methodology could be extended to all the 101 interactions in the database. We address the second limitation by building a generic model using interactions involving multiple doctors and patients. In the long run, the model can be used as an annotation tool for automatic labeling of physician gaze when analyzing physician-patient interactions in clinical settings.

III. Methodology

A. Data

The current database consists of 101 interactions between 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 three 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 distance and thus, 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

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

Doctor Index	Interaction indices	Selected Interactions	Previous work 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

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 investigates the automatic labeling of the physician's gaze using two levels. The problem at hand is a binary classification task where physician's gaze is classified. 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. Of all the frames available for analysis, the physician's gaze was directed at the patient for 45% of the frames and physician's gaze directed elsewhere 55% of the frames.

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 tends 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. Therefore, we have 2 interactions each from 9 doctors adding up to a total of 18 interactions.

B. Feature Extraction

First, we detect the patients and the doctors in the patient-centered videos and the doctors from the doctor-centered videos. We follow the approach used by Gutstein [15] - [17] to extract features such as bounding box coordinates of the patient and the physician location using You Only Look Once algorithm [20] and optical flow measurements [21].

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. Fig. 2 shows the identified YOLO bounding boxes in the patient-centered image 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 regions in two cameras.

Optical flow measurements are used to estimate the motion of patient and the physician between successive frames. For each region-based optical flow computation, 15 summary statistic variables were calculated to aggregate the values of each of the following optical flow 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 measurements. 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.

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]

C. Physician Gaze Prediction

Because of camera angle and the nature of the clinical room, the patient-centered doctor region was not detected consistently and therefore, many of these regions were missing. To eliminate this limitation, the physician regions detected in the patient-centered videos were not considered and thus, the feature space was reduced from 188 to 124 by removing the location-based coordinate features and optical flow measurements related to patient-centered doctor. Therefore, we only used features related to the patient-centered patient and doctor-centered doctor regions for model building. To validate the robustness of our approach, we chose to perform 5 different experiments. 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 other words, in each experiment, the idea is to build a model using interactions from 8 doctors for training, testing and validating the model. In addition to the validation set, we build an additional validation set using interactions from one doctor which was not a part of the training. The training, testing and validation data comes from random sampling of data from 8 doctors and the additional validation set comes from the other doctor. The purpose of the additional validation set is to evaluate the generalizability of the model to completely new data differing in terms of clinical arrangements, camera positioning and doctor. In each of the 5 experiments, interactions from a different doctor were used as the additional validation set (Table II). Two interactions from each of the other 8 doctors were split into training and testing data (70% of all the frames) and validation data (30% of all the frames). For each experiment, one random forest model was trained and tuned and the performance of the model is reported in Table IV. To be noted is that the model obtained optimal results when the number of features used to train each tree in a random forest model was 11 out of the 124 features.

We also investigated the prediction power of the model when considering the time component of the video data. In other words, build the model on frame sequences of a certain length and make predictions for the next sequence of frames in the interaction. We performed 6 different experiments. 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 The combinations are summarized in Table III. A random forest classifier was built on sequences of annotated frames to predict the annotations in future sequences.

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

Model	Interactions used for additional validation set	
	Doctor	***Interactions***
Model R1	Doctor 1	Interactions 01, 02
Model R2	Doctor 3	Interactions 17, 29
Model R3	Doctor 4	Interactions 09, 10
Model R4	Doctor 6	Interactions 34, 49
Model R5	Doctor 7	Interactions 38, 55

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

Model	Duration of the video used		
	Training	***Testing***	***Validation***
Model S1	00:01 – 04:00	04:01 – 05:00	05:01 – 06:00
Model S2	01:01 – 05:00	00:01 – 01:00	05:01 – 06:00
Model S3	02:01 – 06:00	01:01 – 02:00	00:01 – 01:00
Model S4	00:01 - 05:00	05:01 – 05:30	05:31 – 06:00
Model S5	00:01 – 00:30	00:31 – 05:30	05:31 – 06:00
Model S6	01:01 – 06:00	00:31 – 01:00	00:01 – 00:30

IV. Results

The results of the random forest classifier are presented in 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

TABLE IV
PERFORMANCE OF THE MODELS ON DIFFERENT DATA SETS

Model	Optimal random forest model parameters				Performance of the model on different data sets			
	No. of Trees	*No. of features*	*Max. Depth*	*Min. samples for split*	*Training*	*Testing*	*Validation*	*Additional Validation*
Model R1	400	11	30	10	98.24%	83.54%	83.60%	41.54%
Model R2	350	11	35	10	98.43%	83.58%	83.88%	30.46%
Model R3	400	11	30	5	98.30%	83.75%	84.01%	30.52%
Model R4	400	11	35	10	98.18%	83.70%	84.01%	42.34%
Model R5	450	11	30	10	98.29%	83.84%	84.19%	36.22%

TABLE V
PERFORMANCE OF THE MODELS ON DIFFERENT SEQUENTIAL DATA SETS

Model	Optimal random forest model parameters				Performance of the model on different data sets		
	No. of Trees	*No. of features*	*Max. Depth*	*Min. samples for split*	*Training*	*Testing*	*Validation*
Model S1	400	11	30	10	96.56%	66.19%	58.93%
Model S2	400	11	40	10	96.38%	64.09%	66.12%
Model S3	450	11	40	10	96.78%	62.33%	59.53%
Model S4	450	11	35	10	95.09%	66.86%	65.04%
Model S5	450	11	40	5	95.10%	62.23%	68.78%
Model S6	450	11	40	10	95.08%	69.73%	60.01%

(average accuracy of 83.93% by 5 models on validation set from Table IV). However, the performance of the model on the additional validation set was significantly low. 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. Studies also show that there exist distinguishable patterns of gaze between the doctors. 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 conclude that this model could be used to predict physician's gaze on any unseen data within learnt clinical settings and camera projections and within doctors already learnt. 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.

The evaluation of the model using sequences of frames is shown in Table V. The results suggest 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. There different models built using 4 minutes of video interactions were able to produce an average accuracy of 61.52%, and 64.61% using 5 minutes of data. In addition to it, in our previous analysis, which is not reported in this work, we had noted that the models trained using 3 minutes of video had produced an average accuracy of 58.84%. The data that we use here to classify physician gaze are optical flow motion of the patient and the doctor. It is very clear from our analysis that the performance of the models increased with additional duration of video. Through our video data analysis, it was found that the doctors and the patient exhibit variety of motions throughout the interaction (example: looking at the chart, using EHR technology, typing over the keyboard, performing physical examination, talking to the patient and much more). The 101 interactions on an average last over 28 minutes approximately and 6 minutes of data we use is very small to capture the different motions. The analysis suggests that the performance on the unseen sequences could be improved when the model learns different motions of the doctor and patient.

V. CONCLUSION AND FUTURE WORK

The importance of the patient-centered doctor, patient-centered patient and doctor-centered doctor optical flow estimates in gaze recognition were studied and it was found that the patient-centered doctor were redundant by feature selection techniques and the removal of patient-centered doctor information had no impact on the performance of the gaze recognition model. Further, the interactions from multiple 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 average accuracy of 83.93%. 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. The average video duration for 101 interactions is 28 minutes and being able to annotate the the remaining 22 minutes of videos with 6-minute of labelled data shows that 80% of human labor could be reduced. Given the tremendous amount of human labor which goes in manually annotating the videos, we show that this methodology could be used reduce the human labor by approximately 80%. This works shows promise in

terms of reducing human labor and can be extended to other interactions in the database and beyond. This work has used 18 interactions of the 101 interactions in the study and this work can be expanded to the other interactions of the study.

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.

[14] Hart, Y., Czerniak, E., Karnieli-Miller, O., Mayo, A. E., Ziv, A., Biegon, A., Citron, A., & Alon, U. (2016). Automated video analysis of nonverbal communication in a medical setting. Frontiers in Psychology, 7, Article 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. D. and T. Kanade. "An Iterative Image Registration Technique with an Application to Stereo Vision." International Joint Conferences on Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1981, 674–679.