Hand-Eye Coordination: Automating the Annotation of Physician-Patient Interactions

Daniel Gutstein[#]
School of Computing
DePaul University
Chicago, USA
dbgutstein@gmail.com

Enid Montague
School of Computing
DePaul University
Chicago, USA
emontag1@cdm.depaul.edu

Jacob Furst
School of Computing
DePaul University
Chicago, USA
jfurst@cdm.depaul.edu

Daniela Raicu
School of Computing
DePaul University
Chicago, USA
draicu@cdm.depaul.edu

***Abstract*—The widespread adoption of electronic health records within clinical settings has renewed interest in understanding physician-patient interactions. Previous work analyzing clinical interactions has mostly coupled patient surveys with manually annotated video interactions provided by human coders. Physician gaze is among the components of the non-verbal interaction which has been found to impact patient outcomes. The work described in this paper illustrates an automated system for video labeling of patient-physician interactions and shows that image features (such as areas and positioning of physicians' hands) can provide important visual aids for learning physician gaze with over 90% accuracy. While our approach focuses on physician gaze, it can be extended to capture other clinical human-human and human-technology interactions as well as connect these interactions to patient ratings of clinical interactions.**

***Keywords*—Clinical Interaction, Automatic Labeling, Physician Gaze**

I. INTRODUCTION

Widespread adoption of electronic health record (EHR) systems in clinical settings has affected the dynamic between clinicians and patients. Research findings have shown that EHR usage can facilitate the flow of accessible, accurate information to patients and physicians, improve decision-making and medication management, and lead to overall improvements in health-care quality [1]. However, the presence of the EHR in the room can also influence cognitive functioning [2] and alter the ability of the physician and patient to communicate on an emotional level [1]. This technological upending of the physician-patient relationship – in both positive and negative ways – has challenged long-held doctrines regarding clinical interactions and accentuated the need for a more robust understanding of the physician-patient exchange.

The physician-patient interaction can be categorized as both verbal and non-verbal [3]. Verbal interactions can be classified into three subcategories: rapport development, data gathering, and patient education [4]. Beck *et al*. [4] reviewed 14 verbal interaction studies and found negative patient outcomes – including long-term health, adherence, satisfaction, and compression – to be correlated with 14 physician verbal behaviors. These verbal behaviors included high rates of biomedical questioning and low rates of physician feedback to patient information. In their analysis of follow up oncology visits, Eide *et al*. [5] found that informal talk during the history taking phase (rapport development/data gathering) of the interaction was associated with higher patient satisfaction ratings. The authors also determined a trend of patient dissatisfaction to be present when physicians communicated in a psychosocial manner (e.g. providing reassurance of general progress) during the physical examination.

Voice characteristics account for pitch, loudness, tempo, and modulation and have been used in several studies. Little *et. al* [6] considered 275 videotaped consultations from 25 general practice physicians. The results of their regression indicated that among other characteristics, tone of speech, physical contact, and gestures (such as head movement) have statistically significant impacts upon patient ratings of satisfaction.

According to Mast & Cousin [3], non-verbal exchanges contain three components: facial expression (e.g. eyebrow raising, gazing, and smiling), body posture (e.g. positioning of arms and legs), and hand gesturing (e.g. scratching, thumbs up, hand clenching). Beck *et al*. [4] determined that positive patient outcomes are associated with *less* mutual gaze, physician arm symmetry, body orientation, and uncrossed legs and arms. Bensing *et al.* [7] established that general practitioners with higher levels of patient-directed gaze proved to be more adept at identifying signs of patient emotional distress. Gorawara-Bhat *et al*. [8] focused upon elderly patients in a study comparing clinical exchanges with high levels versus low levels of eye-contact and clinical exchanges. Their research found minimal changes in patient-understanding and adherence between divergent eye-contact scenarios. Ishikawa *et al*. [9] analyzed 89 video recordings of doctor-simulated interactions by post-clerkship medical students to assess the connection between specific physician non-verbal behaviors such as eye contact, head movement, and body lean with patient evaluations of interactions. Their findings showed correlations between positive patient ratings and clinicians facing the patient directly, limiting unnecessary movements, nodding when listening, gazing at the patient equally when speaking and listening, matching the

Funding Agency: *NSF Division of Information & Intelligent Systems*

2471-7819/19/$31.00 ©2019 IEEE
DOI 10.1109/BIBE.2019.00123

verbal speed and volume of the patient, and modulating vocal tone and intonation.

The task of linking annotations of non-verbal behavior to video data has relied upon the standard mechanism of using human coders to label the data. This process is time-consuming, labor intensive, and highly context dependent [3, 10, 11]. Conflicting findings and the lack of consensus regarding what to measure also make it difficult to quantify and generalize the relationships between physician behaviors and patient outcomes such as satisfaction, understanding, and adherence [4].

Recent advances in human activity recognition indicate it is possible to recognize human interaction behavior via automated processes [12, 13]. Hart *et al*. [11] used staged medical interactions to analyze *simulated clinical interactions* (i.e. the actor portraying the medical practitioner would alternate between playing the part of an engaged physician and of a disconnected physician) and measured the kinetic energy outputted across two regions of interest (provider and patient) in the image data. The results showed that an increased level of motion synchrony and energy followership between the 'practitioner' and 'patient' correspond to the physician's staged active engagement with the patient. In this paper we present an application of automating the process of clinical interaction analysis via the extraction of torso and hand features to predict the object of physician gaze throughout specified sequences of interest.

II. METHODOLOGY

The purpose of this study was to evaluate the efficacy of an automated methodology for labeling video data from naturalistic, *non-simulated clinical settings* with nonverbal patient-physician interactions. A diagram of the methodology presented in this study is shown in Fig. 1.

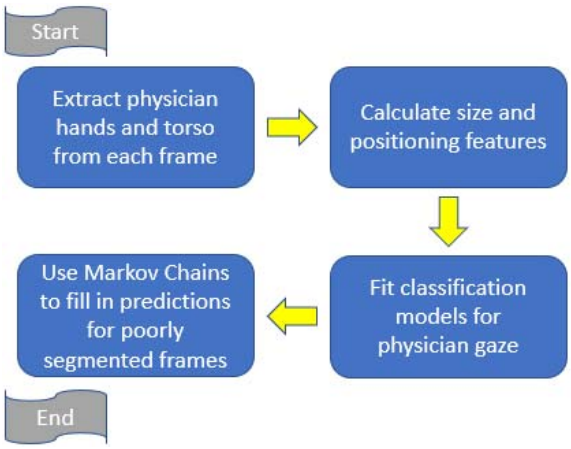

Fig. 1., Diagram of Methodology for Extracting Features from Selected Frames of Physician-Patient Interactions

Using single-view video data, we extracted visual cues (e.g. physician hand position and size) to learn physician gaze characteristics. Outcomes are presented from two phases of analysis. The first phase of classification was based upon the automatically extracted input features in frames for which the segmentation algorithm correctly identified the number of hands in Patient-Centered frames. The second phase of analysis used these predictions to produce frame-by-frame predictions for missing data – those frames for which the segmentation algorithm did *not* correctly identify the number of hands in the Patient-Centered frames. The remainder of this section includes a description of the video data, human annotations, the hand/torso feature extraction, and the two phases of the classification approach.

A. Video Data

There were 101 patients participating in this study, which was performed through the University of Wisconsin-Madison. The 101 clinical interactions were highly inter-dynamic, meaning that the settings from one interaction to another – in the form of factors such as lighting, camera placement, and number of people – fluctuated. For each clinical interaction, three video cameras – one lens centered upon the patient's chair (encoded as Patient-Centered), one wide-view lens (encoded as Wide-frame), and one lens focused upon the doctor's face (encoded as Doctor-Centered) – temporally captured the visual components of each interaction. All three videos were synchronized and combined to form a single multichannel video. In order to focus upon those videos which recorded the presence of the physician's hands, this methodology was focused upon the data in the Patient-Centered videos.

The typical clinical environment included chairs, a computer, and a desk. The standard clinical interaction consisted of a single doctor and a single patient, with the doctor assumed to be situated to the left of the patient in the scene space of the Patient-Centered videos. The videos were recorded at a rate of 29.97 frames per second, and we focused our analysis upon 10,745 frames of size 480 by 720 for each interaction in order to focus upon those durations in which the doctor was in the vicinity of his or her desk.

B. Human Annotations

The manual annotations encoded physician communication, physician gaze, and patient gaze and were obtained using the Noldus Observer XT software [14]. Start and stop times as well as duration were recorded for each form of physician and patient behavior. Given that the computer vision and machine learning algorithms were applied on a frame-to-frame basis, the original annotations were mapped to each frame. An additional human annotator confirmed the frame labels for physician gaze after the annotations were mapped to frames. If the physician was deemed to be looking at the patient in a frame, that frame was labeled *Patient*. If the physician was deemed to not be looking at the patient in a frame, that frame was labeled *Other*. The additional annotator also encoded the number of

hands in each frame to be used as reference truths for validating the hand detection and recognition algorithms. A snippet of original manual annotations for physician communication, physician gaze, and patient gaze are shown in the visualization in Fig. 2.

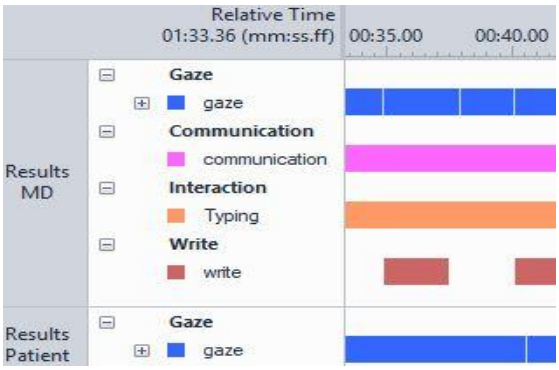

Fig. 2., Original Annotation Data Visualization Representation: 00:00:32–00:00:42 (seconds), Interaction 001

C. Hand/Torso Feature Extraction

The raw videos were acquired using the Red Green Blue (RGB) color space. Given that the RGB space provided insufficient contrast to discriminate between pixels in regions with human skin tone and pixels in regions without human skin tone, we converted the video data from each interaction to the Hue Saturation Intensity (HSI) space. Then, a combination of thresholding approaches for the HSI and RGB channels were used to segment the skin pixels from the rest of the pixels. As a post-processing step, we applied a Gaussian filter to smooth the edges of the regions [15].

Once the hands were segmented based upon skin pixel data thresholding, we used domain knowledge regarding physician and patient positioning in our clinical setting (physician sits at computer on the left side of the frame and patient sits on the right side of the frame in the vicinity of the desk) to focus on sub-regions of the frames and differentiate between the physician and patient hands. The parameters for HSI thresholding and subregion search-spacing were adjusted for each patient to account for positioning and lighting changes. Furthermore, the candidate physician hands were the two largest connected components of the segmented image.

The methodology for segmenting the physician torso was conceptually similar to the segmentation of physician hands, although the entire image search space was used and a separate combination of HSI and RGB channels were employed for the purpose of thresholding. The largest connected component (smoothed and augmented using a Gaussian filter) was classified as physician torso [16]. The candidate physician hands were confirmed as hands if the segmented hand was connected to the connected component representing the smoothed and augmented physician torso. For each interaction, unique hyper-parameters were used for the search space and HSI and RGB channels in order to account for changes in lighting, pixel intensities, and camera positioning between interactions. The segmented hands were further described using 16 automatically extracted variables, 15 of which are listed in Table 1.

The variable *Numbers of Hands Detected* is not listed in Table 1 because it can be inferred from the other variables. X and Y are the pixel coordinates representing the hand or torso regions. Three high-level features (*Number of Hands Detected*, *Left Hand Present*, *Right Hand Present)* were then extracted from each frame and used to build what we defined as *Count-Based Features (CBF)* models. These three high-level features were also included with the remaining 13 low-level hand and torso features to build what we named *All Features* (*AllF*) models.

TABLE 1: EXTRACTED VARIABLES FOR HAND AND TORSO CHARACTERIZATION

Video	**Patient-Centered**		
Body Part	*Left Hand*	*Right Hand*	*Torso*
Hand	✓	✓	
X Mean	✓	✓	✓
Y Mean	✓	✓	✓
Min (X)			✓
Min (Y)			✓
Max (X)			✓
Max (Y)			✓
Area	✓	✓	✓

D. Gaze classification

To map image features automatically extracted using computer vision techniques to annotations capturing physician behavior, we divided the classification process into two phases.

In Phase 1, we individually fitted and validated a simple classifier – decision trees (DT) [17] – and a more advanced classifier – AdaBoost (AB) [18] – for each of the patient-physician interactions. The hyper-parameters of the decision tree and AdaBoost classifiers were tuned accordingly for each physician. Those frames in each interaction for which the number of hands identified by the feature extraction system did not match the number of hands encoded by the human annotator were not classified in Phase 1. We present the classification performances regarding physician gaze in terms of sensitivity and precision for the classification of gazing at the patient and of the accuracy upon the testing and validation sets within the interaction upon which each classifier was trained.

In Phase 2 of the classification, for the optimal classifier from Phase 1, we performed predictions of physician gaze on a frame-by-frame basis based upon the mode of the predicted labels for each frame in the testing and validation sets.

Probabilities for each prediction were derived from the homogeneity rates of the predicted labels (e.g. four frame predictions of physician gazing at chart and one prediction of physician gazing at patient resulted in a final frame prediction of chart with 80% probability). The temporal automated labels and probabilities were then augmented using localized first order Markov Chains [19] to predict physician gaze labels for frames in each interaction which did *not* experience the accurate segmentation of hands in the computer vision feature extraction phase. For any label in the dataset, if the probability of the label failed to meet the 70% probability threshold, physician gaze either remained unlabeled or was changed to unlabeled for the frame. The Markov Chains transition matrix was derived from a maximum of the previous 50 frames, and the maximum number of consecutive filled in values was set to 51.

III. RESULTS

We present our preliminary results for two physicians and three patients for each physician, resulting in a total of 6 interactions. The first phase of the process of classifying physician gaze in terms of either gaze to the *Patient* or *Other* was applied to those frames, from a set of 10,745 frames in each interaction, for which the number of components registered as hands by segmentation algorithm matched the number of hands registered by a human coder. The distributions of physician gaze labels – in terms of *Patient* and *Other* – are shown for the matching frames for each of the six interactions in Table 2. To achieve class balance, each interaction's model was fitted with an equal number of *Other* labeled frames and *Patient* labeled frames. For each interaction, the validation data consisted of a random subset of 20% of the frames from the balanced data together with those frames which were originally removed from the model fitting process for the purpose of achieving class balance. The remaining 80% of the data consisted of training and test data. The algorithms were run 40 times upon the training, test, and validation data, with the training and test data being split randomly for each iteration according to a 66%:34% ratio. Physician 1 is associated with *Interaction 1*, *Interaction 2*, and *Interaction 59,* while Physician 2 is associated with *Interaction 65*, *Interaction 68*, and *Interaction 71*. The listing *Int* in Tables 2–11 is an abbreviation for *Interaction*.

TABLE 2: COMBINED MANUAL LABELS FOR PHYSICIAN GAZE

Label	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
Patient	4473	4759	4405	2488	6416	2509
Other	6272	5986	6340	8257	4329	8236

Table 3 lists the number of frames in each interaction for which the number of hands identified by the feature extraction system matched the number of hands manually encoded by the human annotator.

TABLE 3: NUMBER OF FRAME LABELS AND PERCENTAGE OF FRAME LABELS OUT OF 10,745 TOTAL LABELS WITH SEGMENTATIONS CORRECTLY CALCULATING NUMBER OF PHYSICIAN'S HANDS

Label	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
Patient	1970	4210	3252	707	3043	2168
Other	5805	5245	4272	1221	2304	7222
Total	7775 (72%)	9455 (88%)	7524 (70%)	1928 (18%)	5347 (50%)	9390 (87%)

The frames represented in Table 3 (frames with accurate hand segmentations) were the frames upon which the findings from Tables 4 – 10 were derived. Table 4 presents the results of the mean training accuracy (40 iterations) for the classification of physician gaze on each training set. Tables 5 – 10 present the mean accuracy, sensitivity, and precision (40 iterations) for the classification of physician gaze on the test and validation sets within the interaction that each classification algorithm was trained upon. The results are compared across the CBF Models and the AllF Models. For the interactions involving Physician 1, with regard to both DT and AB for the CBF Model and the AllF Model, the minimum leaf size was set to 8 and a maximum of 64 splits were allowed; 50 decision trees were determined as the optimal number of DTs for the AdaBoost (AB) classifier. For the interactions involving Physician 2, with regard to both DT and AB for the CBF Model and the AllF Model, the minimum leaf size was set to 8 and a maximum number of splits to 16; 25 decision trees were determined to be the optimal number of trees for the AdaBoost classifier.

TABLE 4 MEAN TRAINING ACCURACY (ACC): PHYSICIAN GAZE CLASSIFIERS

Classifier	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
CBF DT	71%	64%	59%	60%	66%	51%
AllF DT	89%	89%	88%	79%	84%	80%
CBF AB	72%	64%	59%	60%	66%	51%
AllF AB	100%	100%	100%	100%	95%	90%

TABLE 5 MEAN TEST ACCURACY (ACC): PHYSICIAN GAZE CLASSIFIERS

Classifier	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
CBF DT	71%	64%	58%	60%	66%	50%
AllF DT	86%	88%	86%	75%	83%	79%
CBF AB	72%	64%	58%	60%	66%	50%
AllF AB	93%	95%	96%	79%	88%	84%

TABLE 6: MEAN TEST SENSITIVITY: PHYSICIAN GAZE AT PATIENT

Classifier	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
CBF DT	49%	95%	45%	22%	79%	62%
AllF DT	81%	84%	85%	81%	79%	83%
CBF AB	48%	95%	45%	22%	79%	62%
AllF AB	92%	94%	96%	80%	86%	86%

TABLE 7: MEAN TEST PRECISION: PHYSICIAN GAZE AT PATIENT

Classifier	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
CBF DT	90%	58%	68%	92%	63%	68%
AllF DT	90%	91%	86%	74%	86%	76%
CBF AB	91%	58%	68%	92%	63%	68%
AllF AB	94%	96%	96%	79%	91%	82%

TABLE 8: MEAN VALIDATION ACCURACY (ACC): PHYSICIAN GAZE

Classifier	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
CBF DT	90%	53%	65%	85%	73%	41%
AllF DT	90%	89%	86%	71%	81%	75%
CBF AB	91%	53%	65%	85%	73%	41%
AllF AB	94%	95%	97%	78%	88%	82%

TABLE 9: MEAN VALIDATION SENSITIVITY: PHYSICIAN GAZE AT PATIENT

Classifier	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
CBF DT	51%	95%	47%	22%	80%	61%
AllF DT	79%	83%	85%	79%	78%	82%
CBF AB	50%	95%	47%	22%	80%	61%
AllF AB	92%	93%	95%	79%	86%	84%

TABLE 10: MEAN VALIDATION PRECISION: PHYSICIAN GAZE AT PATIENT

Classifier	Int 1	Int 2	Int 59	Int 65	Int 68	Int 71
CBF DT	46%	40%	45%	64%	82%	42%
AllF DT	48%	81%	72%	37%	94%	21%
CBF AB	48%	40%	45%	64%	82%	42%
AllF AB	60%	92%	93%	42%	97%	27%

For five of the six interactions, AllF AB achieved the highest accuracy and sensitivity scores on testing and validation (at or exceeding 82%). Regarding Int 65 with Physician 2, for which AllF AB model did not achieve the best accuracy and sensitivity scores on testing and validation, an analysis of the results showed that the feature extraction phase itself performed poorly. Int 71 with Physician 2 also had low performance in terms of precision on the validation data. Table 11 summarizes the effect of the Markov Chains on the performance of AllF AB predictions for each interaction made on a frame by frame basis and the subsequent performance metrics. The percentages listed in Table 11 refer to the efficacy of the algorithm across the complete sequence of 10,745 frames. The listing *Pred* in Table 11 is an abbreviation for *Prediction*.

For the interactions involving Physician 1, the application of Markov Chains to fill in missing values from the AllF AB predictions produced an average of 1,733 additional accurate predictions. The mean percentage of frames (out of 10,745) accurately predicted for the three interactions involving Physician 1 before filling in missing values was 74.85%. After filling in missing values via the application of Markov Chains, the mean percentage of frames (out of 10,745) accurately predicted for the three interactions involving Physician 1 increased to 90.98%.

TABLE 11: COMBINED VALIDATION AND TEST PREDICTIONS: NUMBER OF FRAME-BY-FRAME PHYSICIAN GAZE PREDICTIONS AND PERCENTAGE OF PREDICTIONS OUT OF 10,745 TOTAL LABELS

	AllF_AB		AllF_AB + Markov	
Int	*Correct Pred*	*Total Pred*	*Correct Pred*	*Total Pred*
Int 1	7,375 (69%)	7,775 (72%)	9,472 (88%)	9986 (93%)
Int 2	9,455 (88%)	9,090 (85%)	9,601 (89%)	9987 (93%)
Int 59	7,299 (68%)	7,524 (70%)	10,254 (95%)	10571 (98%)
Int 65	1,546 (14%)	1,928 (18%)	6,421 (60%)	8008 (75%)
Int 68	4,784 (45%)	5,347 (50%)	7,558 (70%)	8448 (79%)
Int 71	7,792 (73%)	9,390 (87%)	8,676 (81%)	10456 (97%)

For the interactions involving Physician 2, the application of Markov Chains to fill in missing values from the AllF AB predictions produced an average of 2,844 additional accurate predictions. The mean percentage of frames (out of 10,745) accurately predicted for the three interactions involving Physician 2 before filling in missing values was 43.81%. After filling in missing values via the application of Markov Chains, the mean percentage of frames (out of 10,745) accurately predicted for the three interactions involving Physician 2 increased to 70.28%.

IV. CONCLUSION

Our results demonstrate that a combination of machine learning techniques can be applied to image features automatically extracted from single-view video data to learn physician behavior such as gazing at a patient in a clinical setting. These preliminary results create the premises for exploring new computer vision algorithms to encode single-view video data for automatically capturing human-human interaction and human-machine interaction.

As shown by the results for some physician-patient video data interactions, the segmentation and feature extraction steps need to be refined further to take into account their sensitivity to changes in lighting and patient-physician positioning, as well as variations among interactions within the same physician data. We will be exploring the utility of the YOLO (You Only Look Once) [20] algorithm to improve the robustness of our segmentation approach as well as optical flow [21] to complement the feature measurements for approximating the physician-patient position and body movement across multiple physician-patient interactions.

Furthermore, based on the research of Schneider *et al*. [22] – whose findings determined that HIV infected patients who provided higher ratings in the form of overall

satisfaction, willingness to recommend a physician, and physician trust were more likely to adhere to medication plans – in the long run we will also look into mapping positioning information and energy flows to patient ratings and outcomes. In the long term, we expect that the applications of these techniques will enhance the understanding of the effects of different forms of EHRs on the physician-patient relationships, and further inform the design of more efficient, effective EHRs to enhance the quality of the physician-patient interaction. Ultimately, the proposed work has the potential to inform and aid the design of technologies for capturing interactions from video data and providing real-time feedback to physicians in clinical settings.

ACKNOWLEDGEMENTS

This research was supported by NSF Division of Information & Intelligent Systems Award #1816010, "CHS: Small: Extracting affect and interaction information from primary care visits to support patient-provider interactions."

REFERENCES

[1] R.S Margalit, D. Roter, M.A Dunevant, S. Larson, and S. Reis, "Electronic medical record use and physician-patient communication: an observational study of Israeli primary care encounters," Patient Education and Counseling. vol. 64, pp. 134-141, April 2006.

[2] B. Karsh, "Beyond Usability: Designing Effective Technology Implementation Systems to Promote Patient Safety," Quality & Safety in Health Care, vol. 13, pp. 388-394, October 2004.

[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] 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, pp. 25-38, January 2002.

[5] H. Eide, P. Graugaard, K. Holgersen, and A. Finset, "Physician communication in different phases of a consultation at an oncology outpatient clinic related to patient satisfaction," Patient Education and Counseling, vol. 51, pp. 259-266, November 2003.

[6] P. Little, P, White, H. Everitt, S. Gashi, A. Bikker, and S. Mercer. "Verbal and Non-Verbal Behavior and Patient Perception of Communication in Primary Care: an Observational Study," British Journal of General Practice," vol. 65, pp. 357-635, May 2015

[7] J.M. Bensing, J.J. Kerssens, and M.V.D Pasch, "Patient Directed Gaze as a Tool for Discovering and Handling Psychosocial Problems in General Practice," Journal of Nonverbal Behavior, vol. 19, pp. 223-242, Winter 1995.

[8] R. Gorawara-Bhat, D.L. Dethmers, M.A. Cook, "Physician Eye Contact and Elder Patient Perceptions of Understanding and Adherence, Patient Education and Counseling," vol. 92, pp. 375–38, September 2013.

[9] H. Ishikawa H. Hashimoto M. Kinoshita S. Fujimori T. Shimizu E. Yano, "Evaluating Medical Students' Non-Verbal Communication During the Objective Structured Clinical Examination," Medical Education Journal, vol. 40, pp. 1180–1187, November 2006.

[10] 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, Pp. 161-170. 2014.

[11] Y. Hart, E. Czerniak, O. Karneili-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, August 2016.

[12] G. Skantze, "Real-Time Coordination in Human-Robot Interaction Using Face and Voice," AI Magazine, vol. 37, 19–31, 2016.

[13] D. Bohus, E. Kamar, and E. Horvitz, "Towards Situated Activity Management; Representation, Inference, and Decision Making," Microsoft Research.

[14] P.H. Zimmerman, L. Bolhuis, A. Willemsen, E.S. Meyer, and L.P.J.J Noldus, "The Observer XT: A Tool for the integration and Synchronization of Multimodal Signals," Behavioral Research Methods, vol. 41, pp. 731–735, August 2009.

[15] R. Gonzalez and R. Woods, "Digital Image Processing (4th Edition)," Pearson Education, 2017.

[16] A. Rosenfield and J.L Pealtz, "Sequential Operations in Digital Picture Processing. Journal of the Association for Computing Machinery," vol. 13, 471–494, October 1966.

[17] W.A Belson, "Matching and Prediction on the Principle of Biological Classification," Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 8, 65–75, June 1959.

[18] Y. Freund and R.E Schapire, "Experiments with a New Boosting Algorithm, Machine Learning: Proceedings of the Thirteenth International Conference," pp. 148-156, 1996.

[19] D. Koller and K. Friedman, "Probabilistic Graphical Models: Principles and Techniques," The MIT Press, July 2009.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016 IEEE Conference on Computer Vision and Pattern Recognition," pp. 779–788, 2016.

[21] B. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision. Proceedings of Imaging Understanding Workshop," pp. 121–130, April 1981.

[22] J. Schneider, S.H. Kaplan, S. Greenfield, W. Li, I.B. Wilson, "Better physician-patient relationships are associated with higher reported adherence to antiretroviral therapy in patients with HIV infection," Journal of General Internal Medicine, vol. 19, pp. 1096-1103, November 2004.

[23] O. Asan, H.N. Young, B. Chewning, and E. Montague, "How physician electronic health record screen sharing affects patient and doctor non-verbal communication in primary care," Patient Counseling and Education, vol. 98, pp. 310–316, March 2015.

[24] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L. Guibas, and L.Fei-Fei, "Human Action Recognition by Learning Bases of Action Attributes and Parts," 2011 International. Conference on Computer Vision, pp. 1331–1338, November 2011.

[25] D.T Nguyen, B.S. Hua, L.F. Yu, and S.K. Yeung, "A Robust 3D-2D Interactive Tool for Scene Segmentation and Annotation," IEEE Transactions on Visualization and Computer Graphics, vol. 24, pp. 3005–3018, November 2017.

[26] E. Montague and O. Asan, "Dynamic Modeling of Patient and Physician Eye Gaze to Understand the Effects of Electronic Health Records on Doctor–Patient Communication and Attention," International Journal of Medical Informatics, vol. 83, pp. 225–234, March 2014.

[27] G. Littlewort, M. Bartlett, L.P. Salamanca, and J. Reilly, "Measurement of Children's Facial Expressions During Problem Solving Tasks," Ninth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 298-305, March 2011.

[28] F. Sener, C. Bas, and N. Ikizler-Cinbis, "On Recognizing Actions in Still Images via Multiple Features," European Conference on Computer Vision, pp. 263–272, 2012.

[29] J.D Robinson, "Getting Down to Business Talk, Gaze, and Body Orientation During Openings of Doctor-Patient Consultations," Human Communication Research, vol. 25, pp. 97–123, March 2006.