Winter 11-15-2019

Information extraction from primary care visits to support patient-provider interactions

Daniel Baruch Gutstein
DePaul University, dgustei@depaul.edu

Information Extraction from Primary Care Visits to Support Patient-Provider Interactions

DePaul
University

Daniel Gutstein
Master in Data Science Program
DePaul University

November 15, 2019

DePaul University
College of Computing and Digital Media

MS Thesis Verification

This thesis has been read and approved by the thesis committee below according to the requirements of the School of Computing graduate program and DePaul University.

Name: Daniel Gutstein

Title of dissertation:

INFORMATION EXTRACTION FROM PRIMARY CARE VISITS TO SUPPORT PATIENT-PROVIDER INTERACTIONS

Date of Dissertation Defense: November 15, 2019

Advisor*

Dr. Daniela Raicu

1st Reader

Dr. Daniela Raicu

2nd Reader

Dr. Jacob Furst

3rd Reader

Dr. Enid Montague

** A copy of this form has been signed, but may only be viewed after submission and approval of FERPA request letter.*

Contents

Abstract

Electronic health record systems usage in clinical settings has affected the dynamic between clinicians and patients, notably physician morale and the quality of care patients receive. Recent research correlates physician burnout and negative physician attitudes with use of electronic health record systems. Understanding the relationship between electronic health record usage, physician burnout, and patient care first requires the analysis of patient-provider interactions within the context of verbal features such as turn-taking and nonverbal features such as eye contact. While previous works have sought to annotate nonverbal and verbal features via manual coding techniques and then analyze their impacts, we seek to automate the process of annotation to create a faster, more robust analytical system. This research focuses upon physician gaze and speaking annotations because nonverbal and verbal components of the interaction can be connected to eye contact and turn-taking – key features that have been linked in certain research to patient outcomes. Previously published work from within this project has demonstrated the viability of extracting image features in the form of YOLO-based person positioning coordinates and optical flow summary statistics to inform the learning of physician gaze for two physicians and six patients with over 80% minimum accuracy. This thesis expands upon the previous findings by increasing the number of patients and physicians in the analysis, diversifying the classifiers to be more robust to new data, and incorporating automatically extracted audio information in the form of mel frequency cepstral coefficients and its derivatives in addition to bounding box coordinates and optical flow summary statistics so as to enable predictions regarding physician gaze and speaking annotations on a frame

by frame basis. We thus illustrate a process of developing and implementing an automated system for multiple video labeling of physician-patient interactions. This enables us to demonstrate that a combination of audio and visual features can inform the predictions of physician gaze and speaking annotations in both testing and sequential validation data. While our approach focuses upon learning physician gaze and speaking annotations, the methodologies introduced can be extended to capture other aspects of the interaction as well as connect these interactions to patient ratings of clinical interactions, physician use of electronic health record systems, and measures of physician burnout. Ultimately, the approaches presented in this paper can aid the creation of an interactive system providing instantaneous feedback to providers during clinician visits with the intention of improving clinical care and patient outcomes while simultaneously reducing instances of physician burnout.

Chapter 1: Introduction

The proliferation of electronic health record (EHR) systems in clinical settings has affected the dynamic between clinicians and patients [1] and has been linked to both physician morale [2] and the quality of care patients receive [1]. Research findings have shown that EHR usage can facilitate the flow of accessible and 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 [3] and alter the ability of the physician and patient to communicate on an emotional level [1]. The finding by Melnick et al. [2] correlating physician dissatisfaction with EHR use and physician burnout also highlights further problems indirectly caused by unsatisfactory EHR use, such as suboptimal patient care practices as well as increased instances of medical error and malpractice [4]. Increased instances of physician burnout, in conjunction with the technological upending of the clinical interaction represented by the extent of EHR use, has consequently accentuated the need for a robust understanding of patient-provider interactions and how such interactions can be related to EHR use, physician burnout, and provider care.

A significant body of research analyzing physician behavior has relied upon a manually intensive process in which human coders have rated and/or annotated live interactions, video recordings, and audio clips of interactions according to prescribed methodologies [5]. Researchers have also used simulated sessions to evaluate physician engagement [6]. The focus of our research in this thesis includes physician-patient interactions which can be categorized as containing verbal and nonverbal elements [5, 7]. According to the Roter Interaction Analysis System [8] – a medical dialogue coding methodology – verbal characteristics include socioemotional exchange, turn-taking, task-focused exchange, tone, and affect. Nonverbal characteristics include eye contact, posture, body language, social touch, and facial emotional expression.

1.1. Verbal Interactions

Beck et al. [5] reviewed 14 verbal interaction studies and found negative patient outcomes – including poorer long-term health, nonadherence, and dissatisfaction – to be correlated with a number of physician verbal behaviors: one-way information flow, antagonistic behavior, nervousness, irritation, extensive feedback in the final phase of a visit, directedness, anxiety, expression of opinions during physical examination, dominance, and interruptions. Eide et al. [9] 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 identified a trend of patient dissatisfaction when physicians communicated in a psychosocial manner (e.g. providing reassurance of general progress) during the physical examination.

Voice characteristics (pitch, loudness, tempo, and modulation) have been explored in several studies. Little and White [10] 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. Ishikawa et al. [11] found that the physician’s ability to match the verbal speed and volume of simulated patients was correlated with patient evaluation scores. Haskard et al. [12] applied a content filtering procedure to recordings of the physician in interaction. The filtering procedure removed high and low frequencies so that words were not understandable but retained elements of vocal tone, including tempo, rhythm, volume, and voice quality. Coders applied ratings of vocal tone on a 1–9 scale. When the researchers applied principal component analysis with varimax rotation to the independent variables in the analysis, they found a correlation between vocal tone and patient satisfaction, health status, and adherence to treatment.

1.2. Nonverbal Interactions

According to Mast & Cousin [7], nonverbal 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. [5] observed that positive patient outcomes are associated with *less* mutual gaze, physician arm symmetry, body orientation, and uncrossed legs and arms. Bensing et al. [13] 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. [14] 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. [11] analyzed 89 video recordings of physician-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 verbal speed and volume of the patient, and modulating vocal tone and intonation.

1.3. Automatic Labeling: Verbal and Nonverbal Interactions

The practice of relying upon manual rating systems to analyze clinical interactions is time-consuming, labor intensive, context dependent, and highly subjective to the biases of raters [6, 7, 15]. 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 [5]. An effective automated system that can quickly ascertain interaction features and contextual factors may provide more consistent and instructive measures of physician performance.

Recent advances in human activity recognition indicate that it is possible to recognize human interaction behavior via automated processes [16, 17]. Hart et al. [6] 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.

Hart's work provides the impetus for our research hypothesis, namely that automatically extracted visual and audio cues from medical data – such as torso positioning, pixel velocity measurement, and audio pitch descriptors – can be used to predict the level of interaction between physician and patient, specifically annotations regarding physician gaze and speaking. We propose to automatically extract audio and visual features in the context of naturalistic, non-simulated interactions, with the goal of objectively and efficiently labeling video data with verbal and nonverbal annotations (characteristics) of patient-physician interactions. While in this work we focus on the verbal characteristics of speech recognition and the nonverbal characteristics of physician gaze, the presented methodologies can be extended to capture other interaction parameters, such as turn-taking and eye-contact, which certain research has shown to impact patient perceptions of physician performance. The accurate extraction of these parameters can facilitate a better understanding of the performance influence of EHR usage and its subsequent impact upon physician burnout and patient care. Our work also represents the first step toward the creation of an automated system that employs computer vision and machine learning algorithms to provide interactive feedback systems for primary care physicians. The end goal of this interactive feedback system will be to improve the efficacy of patient-provider clinical interactions within the context of EHR usage to reduce physician burnout and improve patient care.

Chapter 2: Data

There were 10 primary care physicians and 101 patients participating in the study, which was conducted through the University of Wisconsin-Madison at five primary care clinics in 2011 [18]**.** Every patient filled out a consent form authorizing participation in the study and videotaping of his or her live physician visit. Also, all patients completed questionnaires pertaining to health, demographics, and their attitudes toward physicians and hospitals. Physicians in the study provided basic information pertaining to their experiences with technology [19]. For example, physicians were asked whether the usage of computers interferes with relationships with patients. There was no set length for the videotaped interactions, which covered a gamut of health issues and were annotated by encoders after the interaction. The videos themselves and the annotations made by manual coders were the inputs for our analysis.

2.1. Video Data

The 101 clinical interactions were highly inter-dynamic, meaning that 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 physician's face (encoded as Physician-Centered) – temporally captured the visual components of each interaction at a frame rate of 29.97 frames per second. The videos were saved as MOD files. Figure 1 depicts a scene from a single interaction. The Multi-Channel frame is a collection of the Patient-Centered, Wide-frame, and Physician-Centered frames capturing a given moment in time.

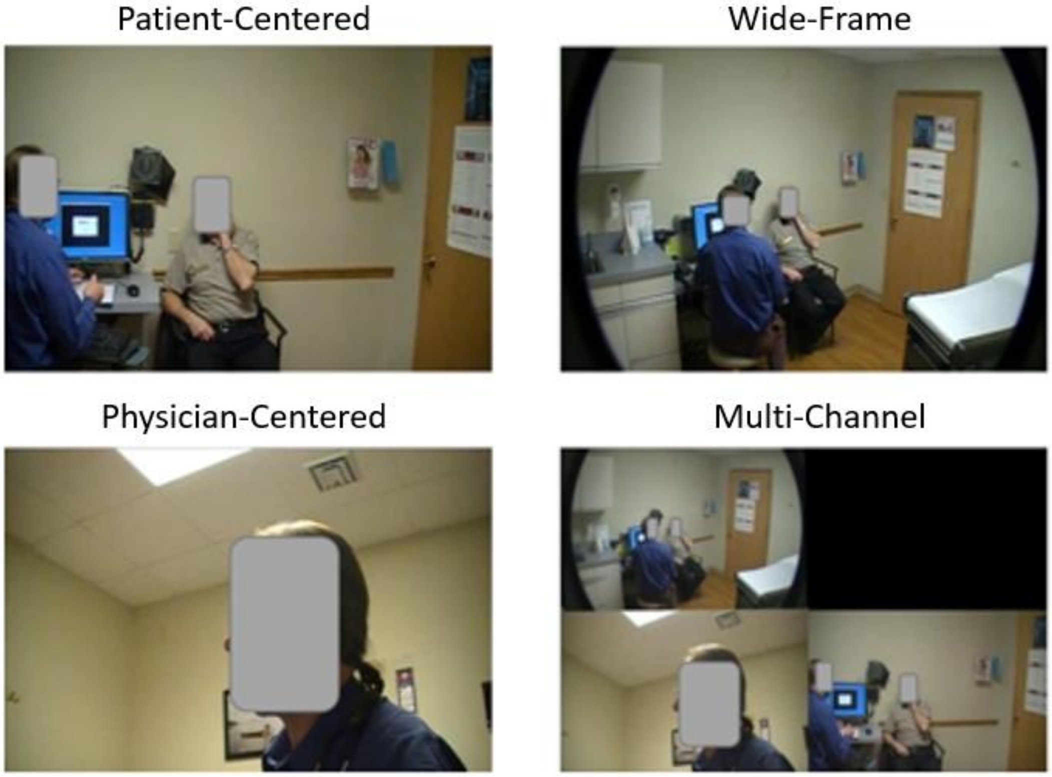


Figure 1. Interaction Video data: Examples of Patient-Centered, Doctor-Centered, Wide-Frame, and Multi-Channel Videos

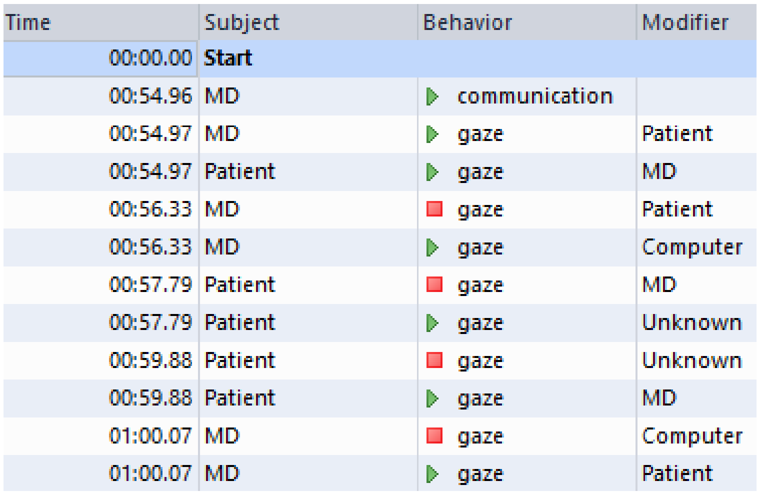
In order to work with the highest resolution videos, simplify the analysis, and focus upon those videos which recorded the most frames with facial and body movement, the work in this thesis is focused upon extracting data using sequential frames from the raw versions of the Patient-Centered and Physician-Centered videos. However, because the raw versions of the Patient-Centered and Physician-Centered videos were not perfectly aligned with one another, we used Avid Media Composer [20] to align the videos and output frames in a designated sequence of interest for each video. The process of aligning the Patient-Centered and Physician-Centered videos for analysis involved:

- Converting the mod video files into mp4 video files
- Loading the mp4 video files into Avid Media Composer
- Using corresponding motion and/or speech in Patient-Centered and Physician-Centered frames to perform alignment
- Outputting aligned sequences as a sequence of jpeg files.

The typical clinical environment included chairs, a computer, and a desk. The sequence of chosen frames consisted of a single physician and a single patient, with the physician assumed to be situated to the left of the patient in the scene space of the Patient-Centered videos. We focused our analysis on the consecutive frames during each interaction in which the patient was present in the Patient-Centered videos and the physician was present in the Physician-Centered videos. We required that the physician be present near his or her desk throughout the chosen sequence to be able to relate the analysis to the human-technology interactions.

2.2. Manual Annotations of Human Behavior: Gaze

Manual annotations encoding physician and patient gaze were obtained using the Noldus Observer XT software [21]. For the entire duration of the 101 interactions, manual annotators provided information regarding the start and stop times for the objects of physician gaze and patient gaze. The annotators also recorded whether the physician was communicating, typing, and/or writing. Figure 2 depicts a screenshot of annotations from an interaction. Figure 3 presents a visual display summarizing annotations from the same interaction.



Time	Subject	Behavior	Modifier
00:00.00	**Start**		
00:54.96	MD	▶ communication	
00:54.97	MD	▶ gaze	Patient
00:54.97	Patient	▶ gaze	MD
00:56.33	MD	■ gaze	Patient
00:56.33	MD	▶ gaze	Computer
00:57.79	Patient	■ gaze	MD
00:57.79	Patient	▶ gaze	Unknown
00:59.88	Patient	■ gaze	Unknown
00:59.88	Patient	▶ gaze	MD
01:00.07	MD	■ gaze	Computer
01:00.07	MD	▶ gaze	Patient

Figure 2. Annotations

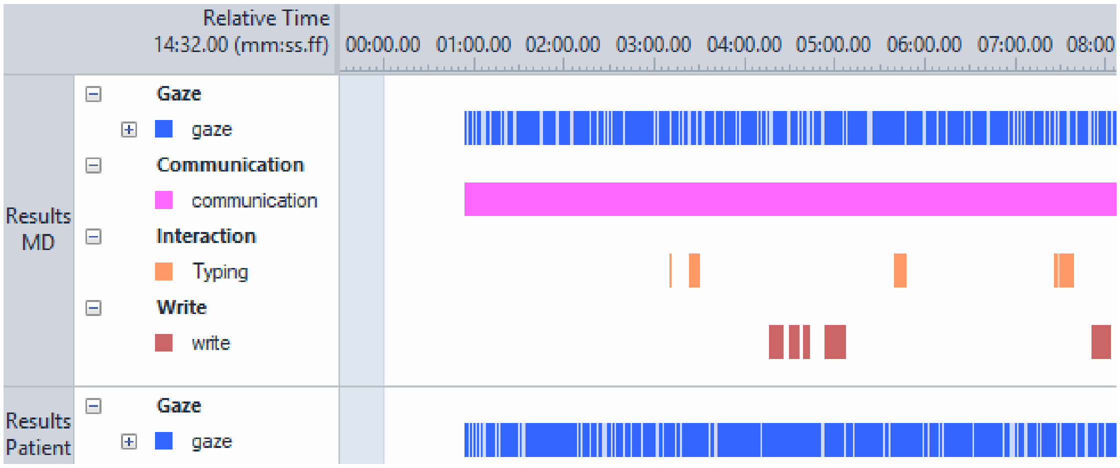


Figure 3. Annotation Visualization

The annotations for gaze provided by annotators contained start and stop times based upon the minutes and seconds of the chosen interactions. However, the computer vision and machine learning algorithms were intended to predict the object of physician gaze on a consecutive frame-to-frame basis. Therefore, the discrete formulations of start and stop times for physician gaze needed to be transformed into continuous representations of the labels. This transformation was performed according to the following process, which is depicted in Figure 4:

- Exported Observer XT file into Excel file
- For each recorded start time of physician gaze, transformed the time encoding from seconds into frames by multiplying the seconds by 29.97 (equivalent to frame rate). This new number was then rounded to the nearest whole number, which represents a frame in an interaction.
- The discrete space between two start times, A and B, was filled with the physician gaze label from start time A.

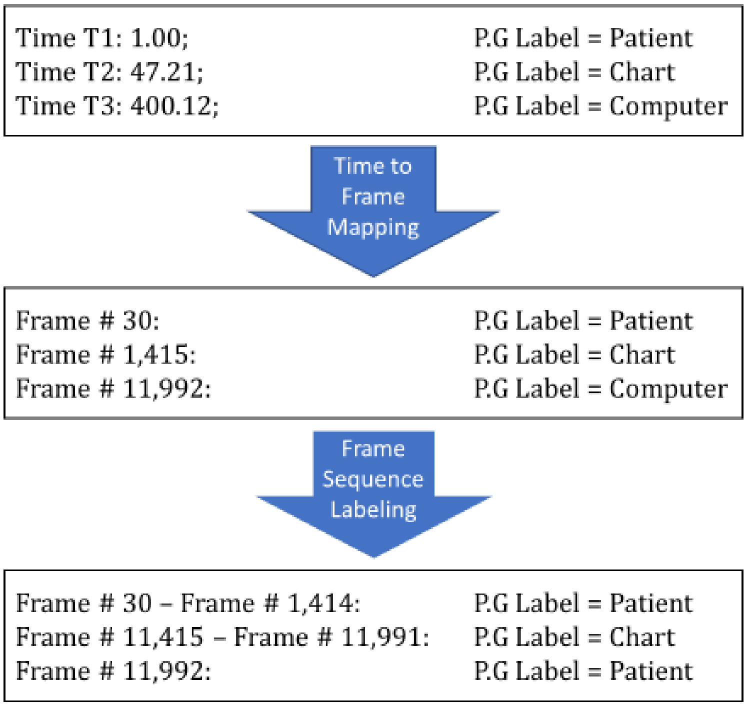


Figure 4. Flowchart for Converting Original Physician Gaze (P.G) Time Annotations into Frame # Annotations

Because the original seconds based annotations were aligned with the Multi-Channel video and not with the raw videos which formed the video data in this analysis, the annotations and videos needed to be manually aligned by comparing annotations with the actual behavior of the physician in the interaction. For the desired sequence of interest in each interaction, an additional human annotator confirmed the frame labels for physician gaze after annotations were mapped to frames. If the physician was deemed to be looking at the patient in a frame, that frame was given the label *Patient*. If the physician was deemed to not be looking at the patient in a frame, that frame was provided the label *Other*. Physician gaze annotations were used as class labels for Phases I, II, and

III of the analysis. Information regarding the distribution of annotations per interaction is listed in Tables 1–4.

Table 1. Phase I and Phase II: Manual Labels for Physician Gaze

	Physician 1			**Physician 2**		
Label	**Interaction 1 (D1_P1)**	**Interaction 2 (D1_P2)**	**Interaction 59 (D1_P3)**	**Interaction 65 (D2_P4)**	**Interaction 68 (D2_P5)**	**Interaction 71 (D2_P6)**
Patient	4,473 (42%)	4,759 (44%)	4,405 (41%)	2,488 (23%)	6,416 (60%)	2,509 (23%)
Other	6,272 (58%)	5,986 (56%)	6,340 (59%)	8,257 (77%)	4,329 (40%)	8,236 (77%)
Total	10,745	10,745	10,745	10,745	10,745	10,745

Table 2. Phase III: Physician 1 (6 Interactions) Manual Labels for Physician Gaze

	Physician 1 (6 Interactions)						
Label	**Interaction 1**	**Interaction 2**	**Interaction 59**	**Interaction 66**	**Interaction 67**	**Interaction 90**	**Total**
Patient	4,473	4,788	4,424	6,834	6,911	5,087	32,517
Other	6,297	5,982	6,346	3,936	3,859	5,683	32,103
Total	10,770	10,770	10,770	10,770	10,770	10,770	64,620

Table 3. Phase III: Physician 2 (6 Interactions) Manual Labels for Physician Gaze

Label	**Interaction 60**	**Interaction 63**	**Interaction 64**	**Interaction 65**	**Interaction 68**	**Interaction 75**	**Total**
Patient	4,291	7,546	5,574	2,488	6,453	3,449	29,801
Other	6,479	3,224	5,196	8,282	4,317	7,321	34,819
Total	10,770	10,770	10,770	10,770	10,770	10,770	64,620

Table 4. Phase III: Physician 3 (5 Interactions) Manual Labels for Physician Gaze

Label	Interaction 77	Interaction 78	Interaction 84	Interaction 98	Interaction 101	Total
Patient	5,811	8,045	4,864	5,886	6,461	31,067
Other	4,959	2,725	5,906	4,884	4,309	22,783
Total	10,770	10,770	10,770	10,770	10,770	53,850

2.3. Manual Annotations of Human Behavior: Speaking

Manual annotations encoding speech were obtained using the BORIS [22] software. The BORIS files were converted into files compatible with Noldus Observer XT software [21]. Manual annotators provided information regarding start and stop times for speakers. A visualization summarizing speaking annotations from a specified time period for a single interaction is shown in Figure 5. A screenshot of annotations from the same interaction is shown in Figure 6.

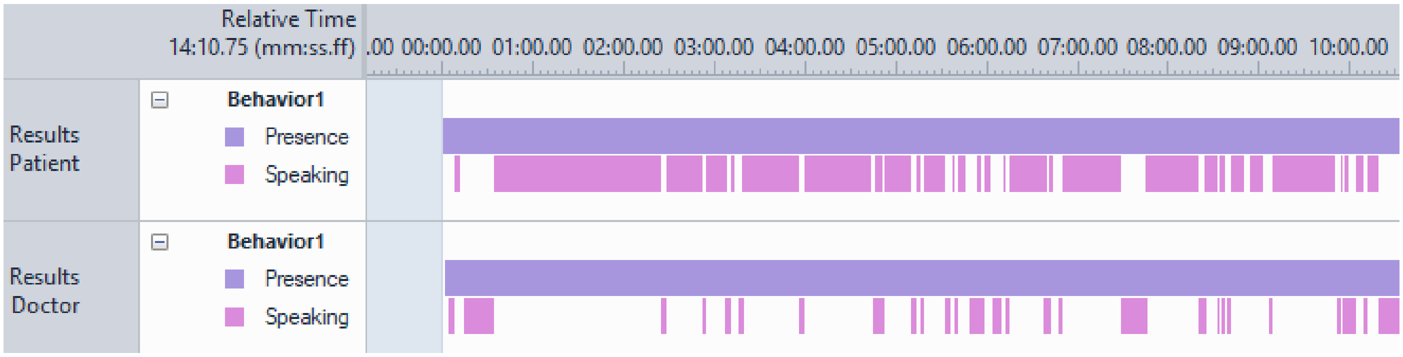
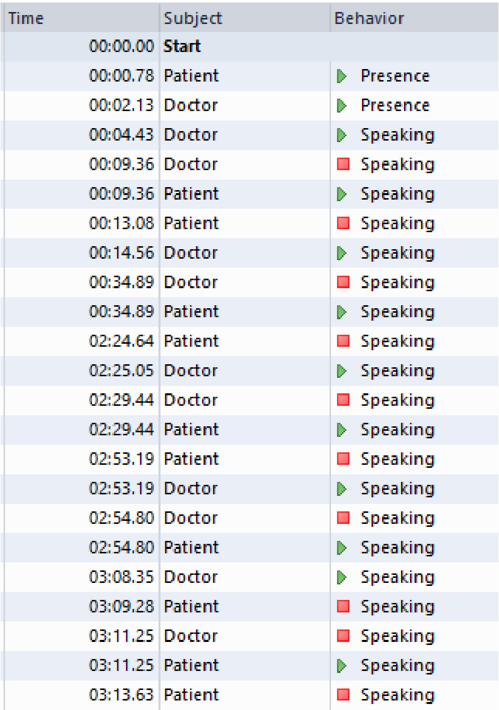


Figure 5. Visualization of Speaking Annotations



Time	Subject	Behavior
00:00.00	**Start**	
00:00.78	Patient	▶ Presence
00:02.13	Doctor	▶ Presence
00:04.43	Doctor	▶ Speaking
00:09.36	Doctor	🟥 Speaking
00:09.36	Patient	▶ Speaking
00:13.08	Patient	🟥 Speaking
00:14.56	Doctor	▶ Speaking
00:34.89	Doctor	🟥 Speaking
00:34.89	Patient	▶ Speaking
02:24.64	Patient	🟥 Speaking
02:25.05	Doctor	▶ Speaking
02:29.44	Doctor	🟥 Speaking
02:29.44	Patient	▶ Speaking
02:53.19	Patient	🟥 Speaking
02:53.19	Doctor	▶ Speaking
02:54.80	Doctor	🟥 Speaking
02:54.80	Patient	▶ Speaking
03:08.35	Doctor	▶ Speaking
03:09.28	Patient	🟥 Speaking
03:11.25	Doctor	🟥 Speaking
03:11.25	Patient	▶ Speaking
03:13.63	Patient	🟥 Speaking

Figure 6. Speaking Annotations

The annotations for speaking provided by annotators contained start and stop times based upon the minutes and seconds of the interaction. Discrete formulations of start and stop times were transformed into continuous representations of these labels. This transformation was performed according to the following process:

- Exported Observer XT file into Excel file
- For each recorded start time and stop time of speaking annotation, transformed the time encoding from seconds into frames by multiplying the seconds encoding by 29.97 (equivalent to frame rate). This new number was rounded to the nearest whole number, which is representative of a frame in an interaction.

- The discrete space between two start times, A and B, was filled with the speaking label from start time A.
- Four possible values for each frame
  - Only Patient speaking, labeled ***Patient***
  - Only Physician speaking, labeled ***Physician***
  - Both Patient and Physician speaking, labeled ***Both***
  - Neither Patient nor Physician speaking, labeled ***Silence***

The original manual annotations were performed by capitalizing upon only the annotators' hearing and intuition. However, using Avid Media Composer [20], the exchange between the patient and physician could be further dissected by analyzing the audio wavelengths to determine starts and stops in speech. A visualization of audio wavelengths can be seen in Figure 7. The portion of the wavelength with minimal amplitude is indicative of lower audio levels recorded by the microphones and implies silence, whereas the wavelength with greater amplitude is indicative of higher audio levels recorded by the microphones and implies speech. Thus, a sudden transfer from a high amplitude portion of the signal to a low amplitude signal can often be considered a stoppage in speech if the low amplitude signal persists, whereas a sudden transfer from a long period of a low amplitude portion of the signal can be often be considered an initiation of speech.

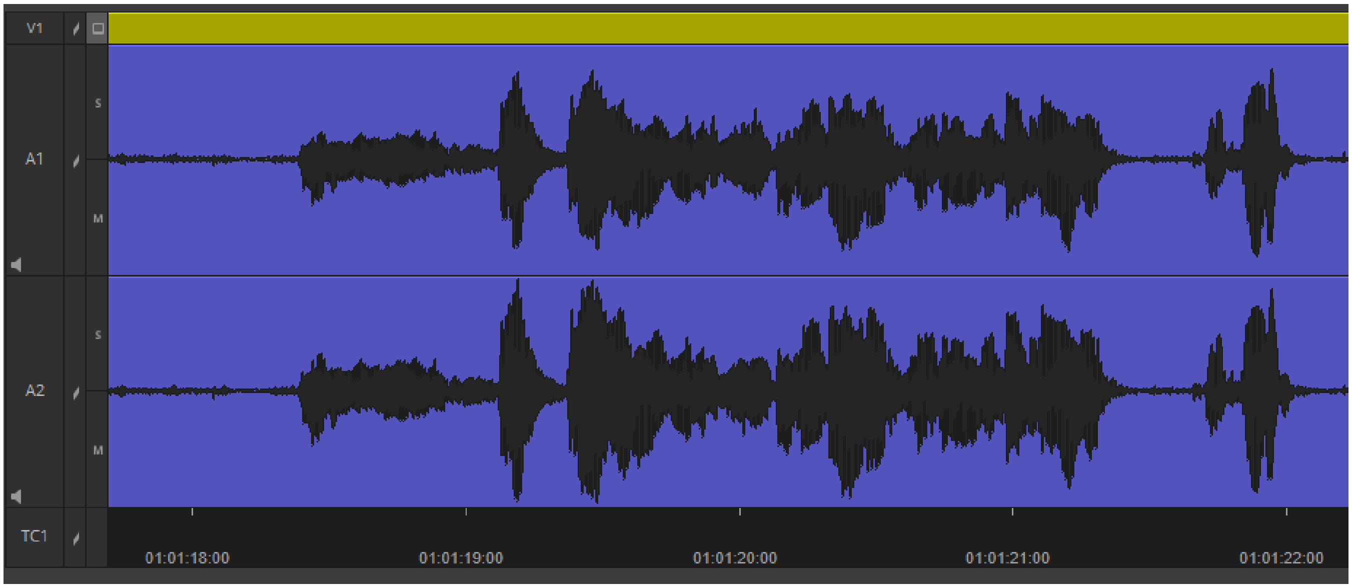


Figure 7. Interaction 2 Audio Wavelengths

Speaking annotations were used as class labels for Phase III of the analysis. Information regarding the distribution of the annotations per interaction is listed in Tables 5–7.

Table 5. Phase III: Physician 1 (6 Interactions) Manual Labels for Speaking Annotations

Label	Int 1	Int 2	Int 59	Int 66	Int 67	Int 90	Total
Both	162	107	226	522	369	271	1,386
Physician	1,551	2,136	5,879	5,710	4,432	4,633	19,708
Patient	6,686	6,654	4,069	3,901	3,747	4,790	25,057
Silence	2,371	1,873	596	637	2,222	1,076	7,699
Total	10,770	10,770	10,770	10,770	10,770	10,770	64,620

Table 6. Phase III: Physician 2 (6 Interactions) Manual Labels for Speaking Annotations

Label	Int 60	Int 63	Int 64	Int 65	Int 68	Int 75	Total
Both	331	87	779	627	681	713	2,505
Physician	5,014	3,916	3,477	6,257	7,702	6,826	26,366
Patient	4,664	4,374	4,798	2,403	1,793	2,016	18,032
Silence	761	2,393	1,716	1,483	594	1,215	6,947
Total	10,770	10,770	10,770	10,770	10,770	10,770	64,620

Table 7. Phase III: Physician 3 (5 Interactions) Manual Labels for Speaking Annotations

Label	Int 77	Int 78	Int 84	Int 98	Int 101	Total
Both	177	88	261	357	214	1,097
Physician	3,519	1,962	4,711	2,075	855	13,122
Patient	4,012	7,154	2,720	7,376	8,397	29,659
Silence	3,062	1,566	3,078	962	1,304	9,972
Total	10,770	10,770	10,770	10,770	10,770	53,850

The relationship between physician gaze annotations and speaking annotations for each physician (Six interactions for Physician 1, Six interactions for Physician 2, Five interactions for Physician 3) is summarized in Tables 8–10.

Table 8. Physician 1 Gaze and Speaking Annotations Across Corresponding Patient Interactions

	Physician Gaze		
Speaking	Patient	Other	Total
Both	582	1,075	1,657
Doctor	13,053	11,288	24,341
Patient	12,882	16,965	29,847
Silence	5,586	3,189	8,775
Total	32,103	32,517	64,620

Table 9. Physician 2 Gaze and Speaking Annotations Across Corresponding Patient Interactions

	Physician Gaze		
Speaking	Patient	Other	Total
Both	1,946	1,272	3,218
Doctor	20,955	12,237	33,192
Patient	7,067	12,981	20,048
Silence	4,851	3,311	8,162
Total	34,819	29,801	64,620

Table 10. Physician 3 Gaze and Speaking Annotations Across Corresponding Patient Interactions

	Physician Gaze		
Speaking	Patient	Other	Total
Both	544	553	1,097
Doctor	6,731	6,391	13,122
Patient	9,485	20,174	29,659
Silence	6,023	3,949	9,972
Total	22,783	31,067	53,850

Chapter 3: Phase I. Interaction Classification using Thresholding Approaches for Feature Extraction

In Phase I of our project analysis, which was published at the 2019 IEEE Bioinformatics and Bioengineering Conference [23], we extracted physician-related features from Patient-Centered videos by using thresholding-based segmentation approaches in the Hue, Saturation, Intensity (HIS) and Red, Green, Blue (RGB) [24] spaces within search spaces. The efficacy of segmentation was validated with a manual labeling process. Features related to physician hands and physician torso from frames deemed to have experienced accurate segmentation of physician hands were used to classify the object of physician gaze from those same frames. Markov Chains were then employed using the combined sequential test and validation predictions of physician gaze in order to fill in predictions on those frames that were poorly segmented, and which were thus not originally subject to classification.

3.1. Phase I. Physician Hands Thresholding-based Segmentation

For the segmentation of physician hands, we used domain knowledge regarding physician and patient positioning in our clinical setting (physician sits at the 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 hands of the physician and patient. The parameters for HSI/RGB thresholding and sub-region search-spacing were then adjusted for each patient to account for positioning and lighting changes. Furthermore, we assumed that the physician's hands were the two largest connected components of the segmented image. As a post-processing step for segmenting physician hands, we applied a Gaussian filter to smooth the edges of the regions [24].

3.2. Phase I. Physician Torso Thresholding-based Segmentation

The methodology for segmenting physician torso was conceptually akin 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 [25]. 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.

3.3. Phase I Extracted Features

The high-level and low-level features extracted for accurately segmented frames are listed in Table 11.

Table 11. Thresholding Features Extracted. Two of three low-level features in blue (*Number of Hands Present* not listed). Thirteen high-level features in red.

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

The three high-level features were 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. The process for extracting features in Phase I is described in Figure 8.

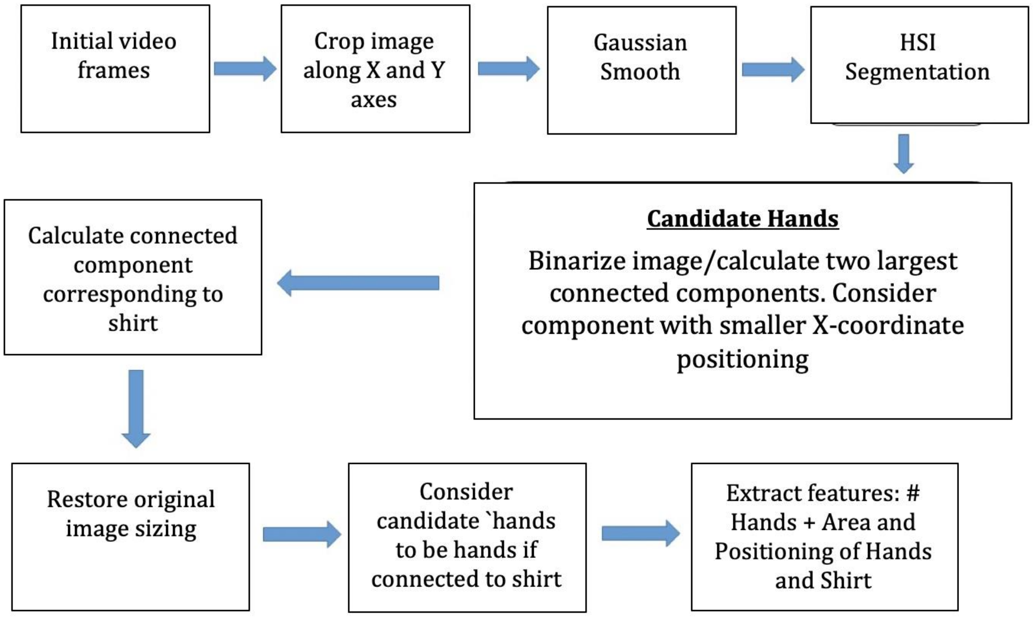
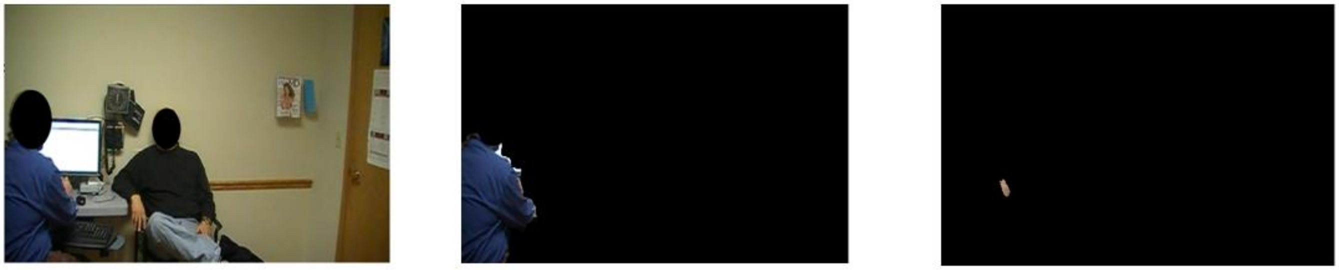

Initial video frames
Crop image along X and Y axes
Gaussian Smooth
HSI Segmentation
Candidate Hands
Binarize image/calculate two largest connected components. Consider component with smaller X-coordinate positioning
Calculate connected component corresponding to shirt
Restore original image sizing
Consider candidate `hands to be hands if connected to shirt
Extract features: # Hands + Area and Positioning of Hands and Shirt

Figure 8. Phase I Flowchart of Feature Extraction for Hands and Torso

3.4. Phase I. Segmentation Validation

Before including the features from the segmentations of physician hands and torso described into a classification scheme, we needed to validate the efficacy of each segmentation. To do this, we manually labeled each frame in the six interactions of interest to determine how many physician hands were present in the frame. For a given frame, if the number of physician hands detected by the segmentation scheme matched the actual number of hands detected by the human labeling process, the extracted features pertaining to physician hands and torso were used by the Phase I Stage I classifier to make a prediction for physician gaze regarding that “accurately segmented frame.” However, for a given frame, if the number of physician hands detected by the segmentation scheme did **not** match the actual number of hands detected by the human labeling process, the entire frame was considered a missing value. The Phase I Stage I classifier ignored the extracted features and did not attempt to make a prediction for physician gaze regarding that frame with the "inaccurately segmented frame."

The process of validating the accuracy of the segmentation process for each frame is depicted in Figures 9 and 10.



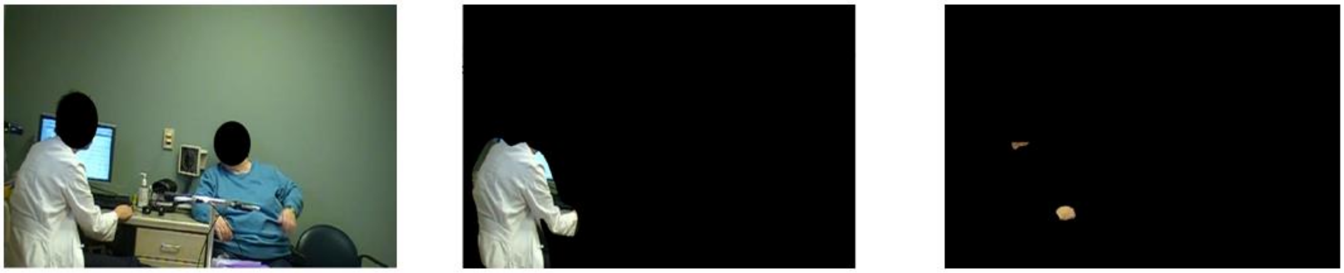
D1_P2 Input Frame: Human Labeler = One Physician Hand

D1_P2 Segmented Torso

D1_P2 Segmented Hand: One Hand Detected

Figure 9. Phase I Interaction D1_P2 Segmentation Validation Process: Accurate Segmentation

In Figure 9, the human labeler labeled there to be one physician hand. Similarly, the automated segmentation system detected one hand. Therefore, all the desired features for the segmented torso and segmented hand for the input frame were included in the Phase I Stage I classification scheme.



D1_P2 Input Frame: Human Labeler = One Physician Hand

D2_P4 Segmented Torso

D1_P2 Segmented Hand: Two Hands Detected

Figure 10. Phase I Interaction D2_P4 Segmentation Validation Process: Inaccurate Segmentation

In Figure 10, the human labeler labeled there to be one physician hand. Conversely, the automated segmentation system detected two hands. Therefore, the frame was not included in the Phase I Stage I classification scheme.

3.5. Phase I. Classification of Physician Gaze

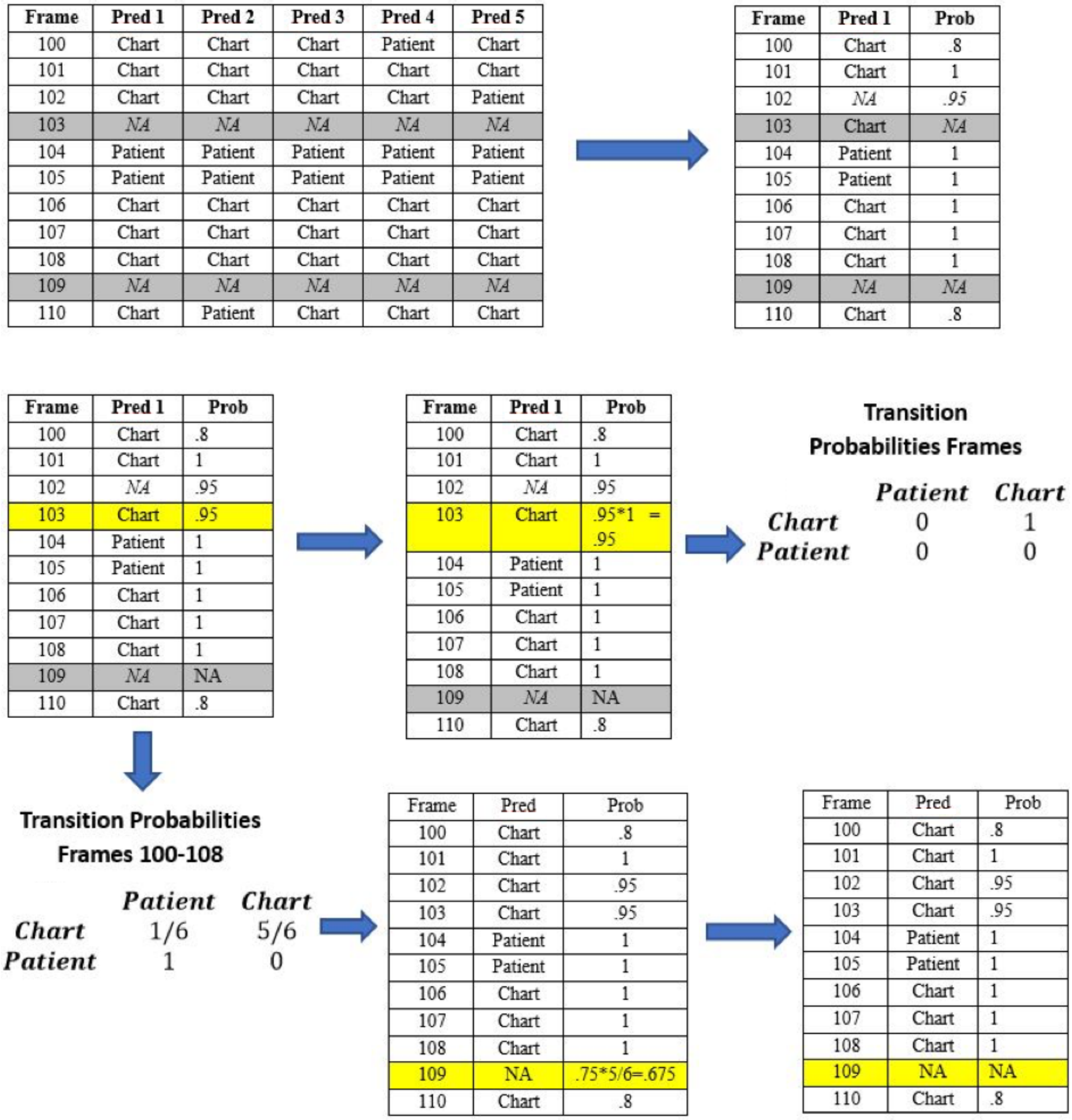
To map image features automatically extracted using computer vision techniques to annotations capturing physician behavior, we divided the classification process into two stages. Phase I Stage I was intended to make predictions for physician gaze on frames with accurate segmentations. The methodology described in Phase I Stage II was used to fill in missing values in the data by making predictions on those frames with poor segmentation of the original frames.

3.6. Phase I Stage I. Decision Tree and AdaBoost Classification of Accurately Segmented Frames

For accurately segmented frames, we individually fitted and validated a simple classifier – decision trees (DT) [26] – and a more advanced classifier – AdaBoost (AB) [27] – for each patient-physician interaction. From a set of 10,745 frames in each interaction, 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 I. To achieve class balance, each interaction's model was fitted with an equal number of 'Other' labeled frames and 'Patient' labeled frames. For every 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.

3.7. Phase I Stage II. Markov Chains on Inaccurately Segmented Frames to fill in Missing Values

In Phase I Stage II, for the optimal classifier, 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 [28] 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 a 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. The process of filling in missing values with Markov Chains is shown in Figure 11.



Frame	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5
100	Chart	Chart	Chart	Patient	Chart
101	Chart	Chart	Chart	Chart	Chart
102	Chart	Chart	Chart	Chart	Patient
103	NA	NA	NA	NA	NA
104	Patient	Patient	Patient	Patient	Patient
105	Patient	Patient	Patient	Patient	Patient
106	Chart	Chart	Chart	Chart	Chart
107	Chart	Chart	Chart	Chart	Chart
108	Chart	Chart	Chart	Chart	Chart
109	NA	NA	NA	NA	NA
110	Chart	Patient	Chart	Chart	Chart

Frame	Pred 1	Prob
100	Chart	.8
101	Chart	1
102	NA	.95
103	Chart	NA
104	Patient	1
105	Patient	1
106	Chart	1
107	Chart	1
108	Chart	1
109	NA	NA
110	Chart	.8

Frame	Pred 1	Prob
100	Chart	.8
101	Chart	1
102	NA	.95
103	Chart	.95
104	Patient	1
105	Patient	1
106	Chart	1
107	Chart	1
108	Chart	1
109	NA	NA
110	Chart	.8

Frame	Pred 1	Prob
100	Chart	.8
101	Chart	1
102	NA	.95
103	Chart	.95*1 = .95
104	Patient	1
105	Patient	1
106	Chart	1
107	Chart	1
108	Chart	1
109	NA	NA
110	Chart	.8

Transition Probabilities Frames

	Patient	Chart
Chart	0	1
Patient	0	0

Transition Probabilities Frames 100-108

	Patient	Chart
Chart	1/6	5/6
Patient	1	0

Frame	Pred	Prob
100	Chart	.8
101	Chart	1
102	Chart	.95
103	Chart	.95
104	Patient	1
105	Patient	1
106	Chart	1
107	Chart	1
108	Chart	1
109	NA	.75*5/6=.675
110	Chart	.8

Frame	Pred	Prob
100	Chart	.8
101	Chart	1
102	Chart	.95
103	Chart	.95
104	Patient	1
105	Patient	1
106	Chart	1
107	Chart	1
108	Chart	1
109	NA	NA
110	Chart	.8

Figure 11. Phase I Stage II: Markov Chain Flow Chart

3.8. Phase I. Results and Analysis

Table 12 presents mean training accuracy (40 iterations) for the classification of physician gaze on each training set. Tables 13–15 present mean accuracy, sensitivity, and precision (40 iterations) for the classification of physician gaze on the test and validation (Val) sets within the interaction that each classification algorithm was trained upon. The results are compared across the CBF Models and the AllF Models.

Table 12. Phase I Stage I Mean Training Accuracy: Physician Gaze Classifiers

	Interaction					
Classifier	D1_P1	D1_P2	D1_P3	D2_P4	D2_P5	D2_P6
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 13. Phase I Stage I Mean Test and Validation (Val) Accuracy: Physician Gaze Classifiers

	Interaction											
	D1_P1		D1_P2		D1_P3		D2_P4		D2_P5		D2_P6	
Classifier	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val
CBF DT	71%	90%	64%	53%	58%	65%	60%	85%	66%	73%	50%	41%
AllF DT	86%	90%	88%	89%	86%	86%	75%	71%	83%	81%	79%	75%
CBF AB	72%	91%	64%	53%	58%	65%	60%	85%	66%	73%	50%	41%
AllF AB	93%	94%	95%	95%	96%	97%	79%	78%	88%	88%	84%	82%

Table 14. Phase I Stage I Mean Test and Validation (Val) Sensitivity: Physician Gaze Classifiers

	Interaction											
	D1_P1		D1_P2		D1_P3		D2_P4		D2_P5		D2_P6	
Classifier	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val
CBF DT	49%	51%	95%	95%	45%	47%	22%	22%	79%	80%	62%	61%
AllF DT	81%	79%	84%	83%	85%	85%	81%	79%	79%	78%	83%	82%
CBF AB	48%	50%	95%	95%	45%	47%	22%	22%	79%	80%	62%	61%
AllF AB	92%	92%	94%	93%	96%	95%	80%	79%	86%	86%	86%	84%

Table 15. Phase I Stage I Mean Test and Validation (Val) Specificity: Physician Gaze Classifiers

	Interaction											
	D1_P1		D1_P2		D1_P3		D2_P4		D2_P5		D2_P6	
Classifier	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val
CBF DT	90%	46%	58%	40%	68%	45%	92%	64%	63%	82%	68%	42%
AllF DT	90%	48%	91%	81%	86%	72%	74%	37%	86%	94%	76%	21%
CBF AB	91%	48%	58%	40%	68%	45%	92%	64%	63%	82%	68%	42%
AllF AB	94%	60%	96%	92%	96%	93%	79%	42%	91%	97%	82%	27%

For five of six interactions, AllF AB achieved the highest accuracy and sensitivity scores on testing and validation (at or exceeding 82%). Regarding D2_P5, for which the 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. D2_P6 also had low performance in terms of precision on the validation data.

Table 16 summarizes the effect of Markov Chains on the performance of AllF AB predictions for each interaction made on a frame by frame basis and subsequent performance metrics. The accuracy percentages listed in Table 16 refer to the efficacy of the algorithm across the complete sequence of 10,745 frames.

Table 16. Phase I Combined Validation and Test Predictions: Number of Frame-by-Frame Physician Gaze Predictions and Percentage of Predictions out of 10,745 total labels

	Accuracy			
Interaction	***Segmentation***	***Classification***	***Refined Segmentation***	***Refined Classification***
D1_P1	72%	69%	93%	88%
D1_P2	89%	85%	93%	89%
D1_P3	70%	68%	98%	95%
D2_P4	18%	14%	75%	60%
D2_P5	50%	45%	79%	70%
D2_P6	87%	73%	97%	81%

For 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%. For 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%, while the minimum accuracy (D2_P4) met a 60% threshold.

3.9. Phase I. Conclusions

The combination of feature segmentation, AdaBoost classification, and Markov chains applied in the Phase I analysis demonstrated that a combination of computer vision and machine learning algorithms could be used to successfully predict physician gaze with at least 60% accuracy across all frames. On correctly segmented frames, the accuracy of AdaBoost AllF was at least 77% on all interactions for both the test and validation data, while the use of Markov chains allowed for the classification system to be successfully employed even across a series of frames with poor segmentations.

While the Phase I analysis and results bore a degree of success in introducing automation to the process of manually annotating features from clinical interactions, the degree of automation was seriously inhibited by an intensive labeling process to determine the number of physician hands in each image, which was necessary to account for the limitations of the segmentation system. Further, the segmentation system itself required manually setting thresholding and search space parameters for each interaction. The goal of Phase II, introduced in Chapter 4, would be to increase the degree of automation in the feature extraction and labeling phases by eliminating the problem of poor segmentation via the application of YOLO and optical flow algorithms.

Chapter 4: Phase II. Interaction Classification using Body Positioning and Optical Flow Approaches for Feature Extraction

In Phase I of our thesis analysis, we successfully utilized a combination of thresholding approaches and classification techniques to make predictions regarding frame annotations of physician gaze with at least 60% accuracy in sequences from six clinical interactions. However, the process required intensive manual labeling and did not provide predictions for all frames of interest due to bad segmentation. These problems caused by poor segmentation were addressed in Phase II via the application of the YOLO [29, 30] algorithm and optical flow [31] measurements.

4.1. Phase II. Optical Flow

In Phase II of our project analysis, published at the 2019 IEEE International Conference on Bioinformatics and Biomedicine[32], we demonstrated that the YOLO [29, 30] algorithm and optical flow [31] measurements could solve the problem of poor segmentation that inhibited the Phase I analysis and thus increase automation and provide annotation predictions for all frames while improving the average and minimum accuracy rate for predicting physician gaze. Optical flow is defined as “the velocity field in the image plane due to the motion of the observer, the motion of objects in the scene, or apparent motion which is a change in the image intensity between frames that mimics object or observer motion [33].” For the Lucas Kanade [31] method, the relationship between temporal and spatial gradients of intensity is exploited to calculate the velocity vector between frames [33]. According to the Brightness Constancy Constraint, by which the projection of the same points looks the same in every frame:

$$I(x, y, t-1) = I(x + u(x, y), y + v(x, y), t); \qquad (1)$$

where [x, y] = image coordinates, t = time, and [u, v] = velocity vector representing displacement between t – 1 and t.

Upon linearizing the right side of the Brightness Constancy equation using a Taylor expansion series:

$$I(x, y, t-1) = I(x, y, t) + I_x \times u(x, y) + I_y \times v(x, y) \qquad (2)$$

Hence:

$$I_x \times u + I_y \times v + I_t \approx 0; \qquad (3)$$

where I_x = Spatial Gradient with respect to x and I_y = Spatial Gradient with respect to y.

According to the assumption of Spatial Coherence, which states that points move like their neighbors, the velocity vector [u, v] can be solved by providing the same value of [u, v] to the pixel's neighbors. Thus:

$$\begin{bmatrix} I_{x1} & I_{y1} \\ I_{x2} & I_{y2} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ I_{x25} & I_{y25} \end{bmatrix} \begin{bmatrix} \mathrm{U} \\ \mathrm{V} \end{bmatrix} = - \begin{matrix} I_{t1} \\ I_{t2} \\ \cdot \\ \cdot \\ \cdot \\ I_{t25} \end{matrix} \qquad (4)$$

In this formulation, the components of the velocity vector – *u* and *v* – can be solved via least squared regression [33]. For synchronized frames in each interaction, we calculated optical flow measurements in the Patient-Centered and Doctor-Centered videos.

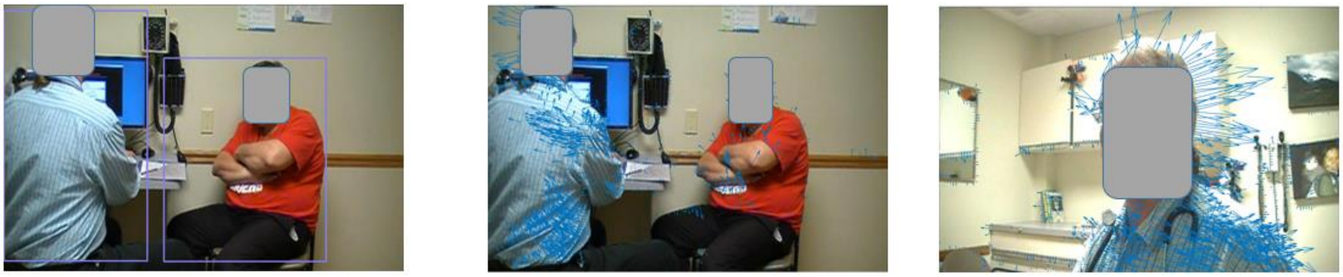
For each optical flow computation, we calculated 14 summary statistic variables regarding each of the following variables: velocity U, velocity V, orientation, and magnitude. The values for velocity U and velocity V refer to the values for *u* and *v* in Equation 4. Orientation [34] was determined from the phase angles [35] of the vectors of *u* and *v* from Equation 4. The magnitude [24] of optical flow was also derived from the values for *u* and *v* in Equation 4 [34]. The pertinent summary statistics are as follows: Maximum, Minimum, 25th Percentile, 50th Percentile, 75th Percentile, Sum, 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 with regard to velocity U, velocity V, orientation, and magnitude, their variables – with the exception of *# Non-Zero Values* – were calculated for the top 25th percentile of feature values in the regions of interest.

4.2. Phase II. YOLO: Human Boundary Boxes

Upon each of Patient-Centered image, we applied the YOLO (You Only Look Once) [29, 30] algorithm, which uses an individual convolutional neural network to predict class probabilities for a series of conditional bounding boxes for specified objects and human beings. An SxS grid is created for each input image, and grid cells are tasked with predicting bounding boxes and creating confidence scores describing the confidence of creating an object and describing accuracy of the predicted box. Furthermore, each cell contains predicted class probabilities. The predicted bounding boxes and confidence scores, as well as class probabilities, are used as inputs to estimate final bounding boxes.

Using domain knowledge, we assumed that when multiple persons were detected in an image, the corresponding bounding box for the physician had the leftmost bounding box X coordinate. Further, the second bounding box corresponded to the patient. When only a single person was detected in a Patient-Centered image, that person was assumed to be the patient. The (X, Y) coordinates of the corner coordinates of these bounding boxes represented the eight positioning variables. In the Patient-Centered frames, the calculations of optical flow described in section 4.1. were quarantined to the regions within the Patient and Physician bounding boxes. In the Physician-Centered frames, in which the physician was generally exclusively present, the calculations of optical flow described in section 4.1. were applied to the entire frame region. The application of YOLO and optical flow velocity vectors in Physician-Centered and Patient-Centered frames for Interaction D1_P3 are depicted by the frames in Figure 12.



Patient-Centered YOLO Bounding Boxes

Patient-Centered Optical Flow

Doctor-Centered Optical Flow

Figure 12. Phase II Interaction D1_P3: Bounding Boxes and Optical Flow Measurement

A summary of the variables extracted using YOLO and optical flow in Phase II is shown in Table 17.

Table 17. Phase II Variables: YOLO body positioning coordinates (4 Coordinates) and optical flow summary statistics (14 Each for Magnitude, Orientation, Velocity X, and Velocity Y)

	Patient-Centered		**Physician-Centered**
Variable	Patient	Physician	Entire Image
Body Positioning	✓	✓	
O.F Magnitude	✓	✓	✓
O.F Orientation	✓	✓	✓
O.F Velocity X	✓	✓	✓
O.F Velocity Y	✓	✓	✓

4.3. Phase II. YOLO/Optical Flow based Classification

To map optical flow and body positioning coordinate features automatically extracted using the Lukas Kanade and YOLO algorithms to annotations capturing physician behavior, we applied AdaBoost [27] classifiers to three data formulations.

The first data formulation (referred to as OF_56) for each interaction consisted of 56 optical flow measurements derived from the Physician-Centered frames. The second data formulation (referred to as OF_168) for each interaction consisted of the same data as 0F_56 plus 112 optical flow measurements from the Patient-Centered frames (168 variables total). The third data formulation (referred to as OF_176) comprised of 168 optical flow variables in addition to eight Patient and Physician bounding box coordinate variables derived from the Patient-Centered frames. (176 variables total). The hyper-parameters of the AdaBoost classifiers (*MNS* is an abbreviation for *Maximum Number of Splits*, *MLS* is an abbreviation for *Minimum Leaf Size*, and *Trees* is a reference to the number of AdaBoost cycles used in the classification process) were tuned accordingly for each of the three data formulations within each physician in order to optimize the classifiers without resulting in overfitting. Depictions of the three classification processes are shown in Figures 13–15.

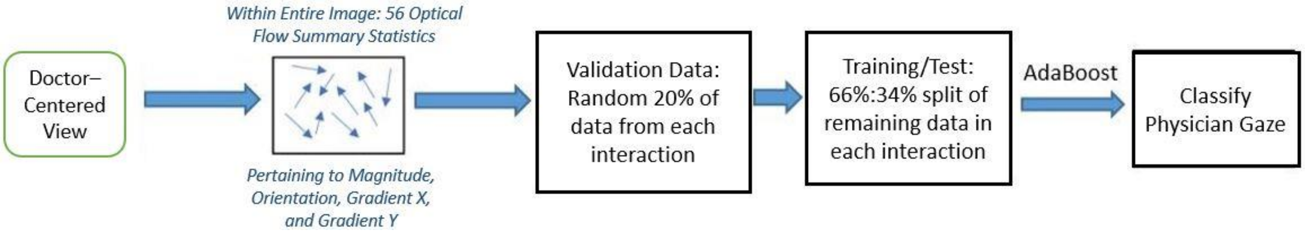

Doctor–Centered View
Within Entire Image: 56 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Validation Data: Random 20% of data from each interaction
Training/Test: 66%:34% split of remaining data in each interaction
AdaBoost
Classify Physician Gaze

Figure 13. Phase II Classification of Physician Gaze: Physician-Centered Optical Flow (OF_56)

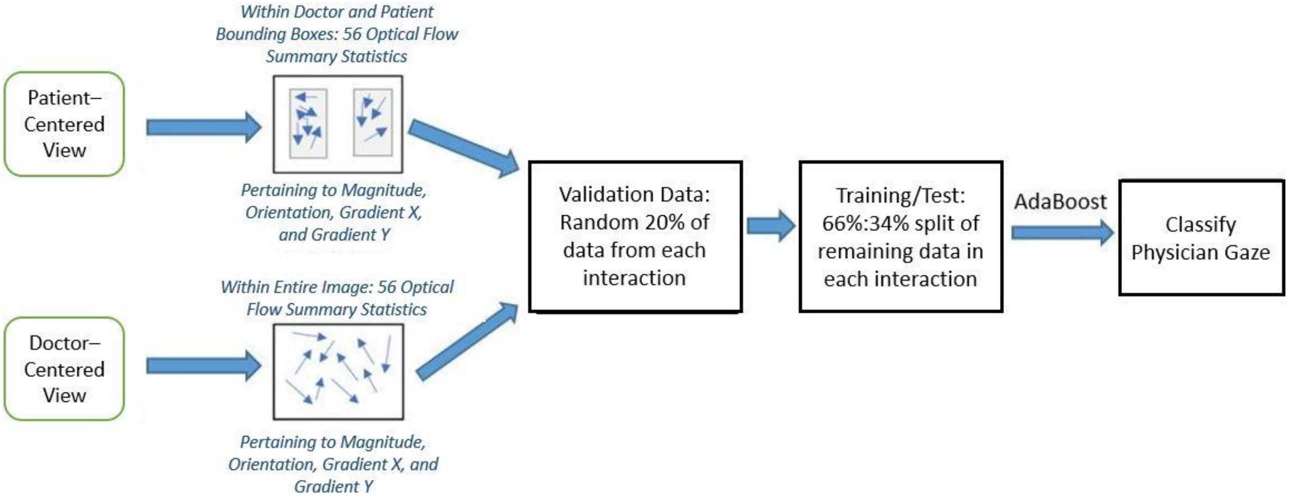

Patient–Centered View
Within Doctor and Patient Bounding Boxes: 56 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Doctor–Centered View
Within Entire Image: 56 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Validation Data: Random 20% of data from each interaction
Training/Test: 66%:34% split of remaining data in each interaction
AdaBoost
Classify Physician Gaze

Figure 14. Phase II Classification of Physician Gaze: Physician-Centered Optical Flow + Patient-Centered Optical Flow (OF_168)

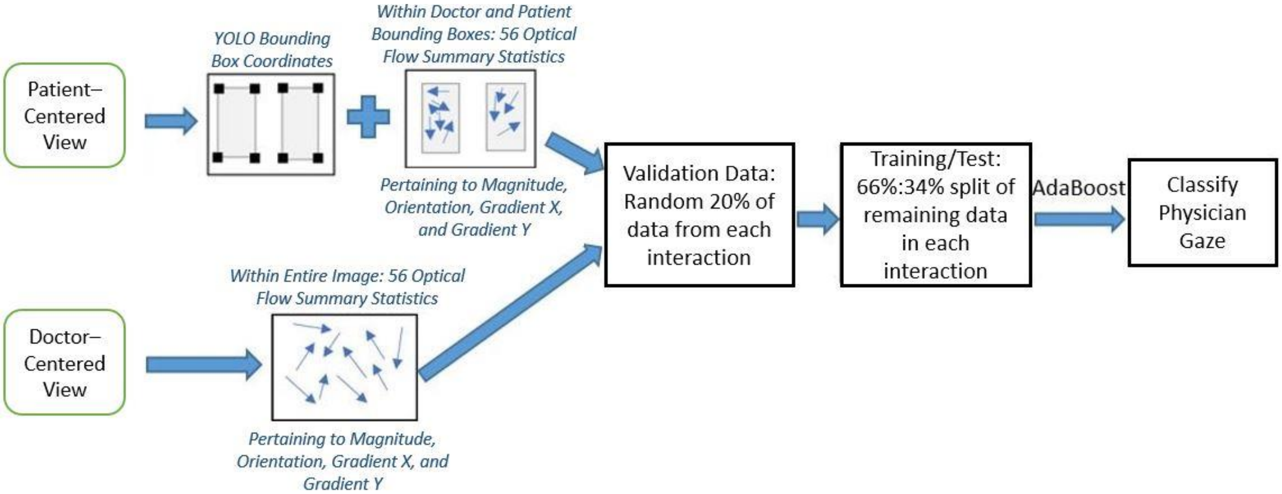

Patient–Centered View
YOLO Bounding Box Coordinates
Within Doctor and Patient Bounding Boxes: 56 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Doctor–Centered View
Within Entire Image: 56 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Validation Data: Random 20% of data from each interaction
Training/Test: 66%:34% split of remaining data in each interaction
AdaBoost
Classify Physician Gaze

Figure 15. Phase II Classification of Physician Gaze: Patient-Centered Bounding Box Coordinates + Patient-Centered Optical Flow + Physician-Centered Optical Flow (OF_176)

4.4. Phase II. Results and Analysis

Table 18 and Table 19 present the mean accuracy (Acc), sensitivity (Sns), and precision (Prc) scores for the classification of physician gaze on the test and validation sets within the interaction that each classification algorithm was trained upon.

Table 18. Phase II Classification of Test Data

	OF_56			**OF_168**			**OF_176**		
Classifier	*Acc*	*Sns*	*Prc*	*Acc*	*Sns*	*Prc*	*Acc*	*Sns*	*Prc*
D1_P1	76%	74%	77%	89%	87%	90%	93%	92%	94%
D1_P2	72%	71%	72%	84%	85%	84%	88%	88%	88%
D1_P3	75%	76%	75%	82%	84%	82%	87%	88%	87%
D2_P4	65%	62%	66%	79%	80%	78%	83%	84%	82%
D2_P5	77%	78%	76%	79%	79%	79%	84%	83%	84%
D2_P6	70%	70%	70%	73%	73%	73%	81%	81%	81%

Table 19. Phase II Classification of Validation Data

	OF_56			**OF_168**			**OF_176**		
Classifier	*Acc*	*Sns*	*Prc*	*Acc*	*Sns*	*Prc*	*Acc*	*Sns*	*Prc*
D1_P1	78%	75%	54%	89%	88%	74%	93%	93%	82%
D1_P2	73%	72%	54%	85%	85%	71%	88%	88%	77%
D1_P3	75%	76%	48%	82%	83%	58%	87%	88%	67%
D2_P4	67%	62%	13%	78%	79%	22%	82%	83%	27%
D2_P5	79%	80%	92%	80%	80%	93%	84%	84%	95%
D2_P6	68%	71%	15%	72%	75%	17%	80%	82%	24%

For all six interactions, the OF_176 classifier, which incorporated positioning data from the Patient-Centered videos and optical flow from the Patient-Centered and Physician-Centered videos, proved to be the most robust of the three forms of classification. The OF_176 classifiers had the highest average accuracy, sensitivity, and precision scores on testing and validation in each interaction. More specifically, the OF_176 classifiers exhibited excellent stability, as the average accuracy, sensitivity, and precision scores were at or exceeding 80% for each interaction – such as 93% for Interaction 1. Furthermore, on the validation data, the OF_176 classifier produced average accuracy and sensitivity scores at or exceeding 82% for each interaction. The OF_176 classifier also reached or exceeded a threshold precision score of 67% on the validation data for four of the six interactions. However, for Interactions 65 and 71, the combined average precision score was 25.5%, which may have been due to variability in the validation data which was not accounted for in the training and test data.

4.5. Phase II. Conclusions

The usage of YOLO and optical flow for feature extraction, as well as the application of AdaBoost classification, demonstrated that computer vision and machine learning algorithms combined can be used to successfully predict physician gaze with at least 80% on both test and validation data. The usage of YOLO and optical flow in Phase II addressed the major areas of concern which emerged from the Phase I analysis, namely the problem of poor segmentation that had caused the individualized setting of hyper-parameters and excessive amounts of manual labeling (regarding the number of physician hands in each image). The Phase II analysis thus achieved an increased degree of automation as well as improved accuracy scores, as the minimum accuracy on the combined test and validation data (measured across all 10,745 frames) increased from 60% in Phase I to 80% in Phase II.

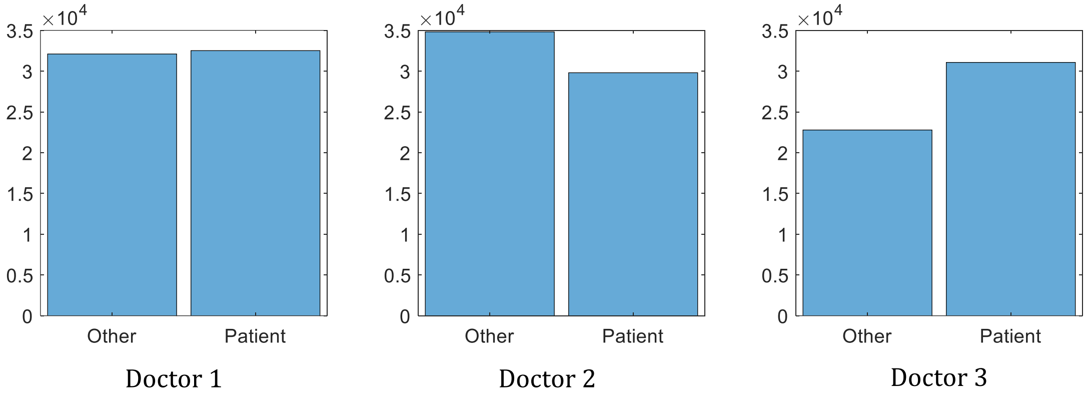
The improved automation and accuracy of the Phase II process – by using optical flow and YOLO for feature extraction – was overwhelmingly successful on the same six interactions which were featured in Phase I. However, the findings were limited to a total of two doctors and a total of six interactions. Further, there was a degree of instability in the validation results. The goal of Phase III, introduced in the upcoming chapter, became to address the issues of instability and limited test case by expanding and diversifying the classifiers to include more interactions and doctors. The issue of instability would also addressed via the addition of pitch-related audio features to aid the prediction manual annotations.

Chapter 5: Phase III. Interaction Classification using Improved Audio Features and Improved Phase II Processes

Phase III of this project analysis, published for the first time in this thesis, seeks to improve and expand upon the findings of Phase II in terms of the expansiveness and stability of the classification scheme. The improved Phase II process – via the usage of optical flow and YOLO for feature extraction – succeeded in introducing increased automation and improving the power of predicting the annotations of physician gaze. However, the findings were limited to a total of two doctors and a total of six interactions, while the classification only predicted annotations of physician gaze. Further, there was a degree of instability in the results of the validation. Also, in order to validate a system's efficacy to perform feature extraction and classification for the purpose of providing simultaneous feedback to physicians, it is necessary to demonstrate the efficacy of the model on video data which can simulate an additional interaction. In Phase III, we address these issues by expanding the data gathering and annotation and by further developing the schema of feature extraction, validation data organization, and classification. The changes implemented in Phase III are explained in sections 5.1.–5.5.

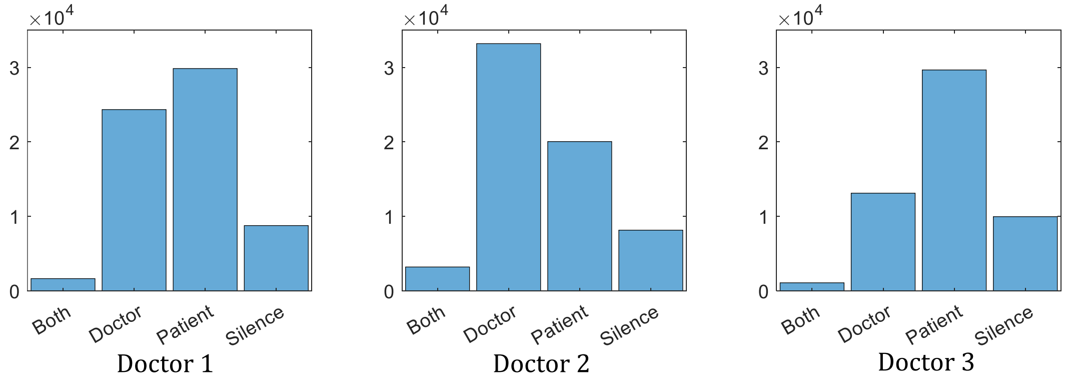
5.1 Phase III. Annotation and Data Gathering

Phase I and Phase II used manual labels of physician gaze annotations as a dependent variable to be predicted. In Phase III, we have added manually labeled speaking annotations. Furthermore, instead of limiting the annotation and subsequent data gathering and analysis to the same two doctors and six interactions (10,745 frames per interaction) used in Phase I and Phase II, Phase III focuses upon 17 interactions (10,770 frames per interaction) for three physicians, with six interactions for Physician 1, six interactions for Physician 2, and five interactions for Physician 3. The distribution of physician gaze annotations for each doctor is illustrated in Figure 16. The distribution of speaking annotations for each doctor is illustrated in Figure 17.



Doctor 1 Doctor 2 Doctor 3

Figure 16. Distribution of Physician Gaze Annotations: Doctors 1 – 3



Doctor 1 Doctor 2 Doctor 3

Figure 17. Distribution of Speaking Annotations: Doctors 1 – 3

5.2. Phase III. Feature Extraction

In the sequences of interest, we sought to automatically extract audio features from each Doctor-Centered video interaction that could be mapped to the 10,770 frames of interest. Mel Frequency Cepstral Coefficients (MFCCs) are an integral factor of numerous speech automatic speech recognition systems [36]. The computation of MFCCs utilizes framing and Fourier transforms to extract coefficients which are correlated with perceptions of pitch [37]. The procedure for extracting MFCCs and mapping them to the video frames is to:

1) Apply pre-emphasis filter to amplify high frequency aspects of the signal
2) Slice signal into overlapping frames
3) Upon every frame
    - Apply window function
    - Perform Short-Term Fourier Transform and derive power spectrum
    - Compute filter banks
    - Apply Discrete Cosine Transform
    - Perform mean normalization

The 14 Mel Frequency Cepstral Coefficients, along with 14 delta (change in coefficients) coefficients and 14 deltaDelta (change in delta) coefficients were calculated using MATLAB's Audio Toolbox [37, 38, 39].

The variables extracted using YOLO, optical flow, and variables related to MFCCs are summarized in Table 20. In Phase III, an additional optical flow summary statistic not calculated in Phase II – sum of squared – was also computed.

Table 20. Phase III Variables: YOLO body positioning (4 Coordinates), Optical Flow Summary Statistics (15 Each for Magnitude, Orientation, Velocity X, and Velocity Y), and Audio Features (14 variables for each audio feature)

	Patient-Centered		**Physician-Centered**
	Patient	Physician	Entire Image
Body Positioning	✓	✓	
O.F Magnitude	✓	✓	
O.F Orientation	✓	✓	
O.F Velocity X	✓	✓	✓
O.F Velocity Y	✓	✓	✓
MFCC	✓	✓	✓
Delta	✓	✓	✓
Delta-Delta	✓	✓	✓

Table 21 summarizes the feature extraction techniques which are used in each phase of the analysis.

Table 21. Feature Extraction Techniques: Phase I, Phase II, Phase III

	Phase I	**Phase II**	**Phase III**
Search Space Definition	✓		
HSI/RGB Based Thresholding	✓		
YOLO Positioning		✓	✓
Optical Flow		✓	✓
Audio Features			✓

5.3. Phase III. Validation Data

The different interactions, even within the same physician, often contain varying camera angles, motion characteristics, and spatial relationships. This variability in the data inhibits the ability of a classifier trained upon one interaction to predict behavior in other interactions. We created simulated validation data by removing entire sequences of consecutive frames from the data before training the classifier. The “new” sequential data possesses a measure of separation from the interaction, since a great number of the frames in the validation data are not temporally close to frames in the training and test data. Yet, the factors of the patient and camera angle were held constant, and therefore, we make inferences about the durability of the model in terms of introducing new interaction data based upon its performance on sequential validation data.

5.4. Phase III. Classification

Each classifier in Phase II was separately trained using data from a single interaction. In Phase III, wherein the number of interactions per physician and the number of physicians were both increased, we trained and tested the labels for all the interactions within each physician. In other words, the data to predict labels for the six interactions for Physician 1 were trained and tested together, the data to predict labels for the six interactions for Physician 2 were trained and tested together, and the data to predict labels for the five interactions for Physician 3 were trained and tested together. Three classifiers were trained in Phase III.

1) Visual features (optical flow and YOLO body positioning bounding boxes) were used to aid the predictions of physician gaze in Test and Validation data. This process is depicted in Figure 18.
2) Audio features (MFCC, LOC, delta, and delta delta) were used to aid the predictions of speaking annotations in Test and Validation data. This process is depicted in Figure 19.
3) A combination of visual features (Optical flow and YOLO positioning measurements) and audio features (MFCC, LOC, delta, and delta delta) were used to aid the predictions of physician gaze and speaking annotations in Test and Validation data. This process is depicted in Figure 20.

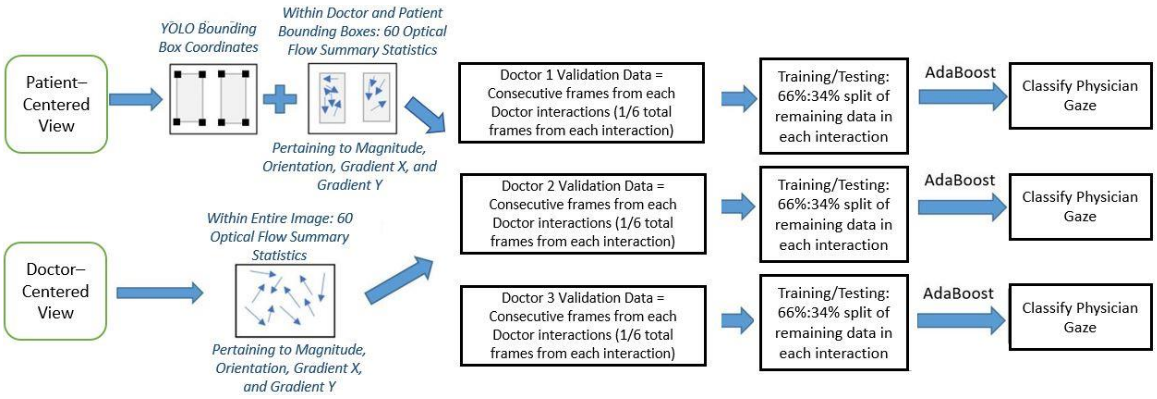

Patient–Centered View
YOLO Bounding Box Coordinates
Within Doctor and Patient Bounding Boxes: 60 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Doctor–Centered View
Within Entire Image: 60 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Doctor 1 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Doctor 2 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Doctor 3 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Training/Testing: 66%:34% split of remaining data in each interaction
AdaBoost
Classify Physician Gaze

Figure 18. Phase III Classification of Physician Gaze: Patient-Centered YOLO Bounding Box Coordinates + Patient-Centered Optical Flow + Physician-Centered Optical Flow

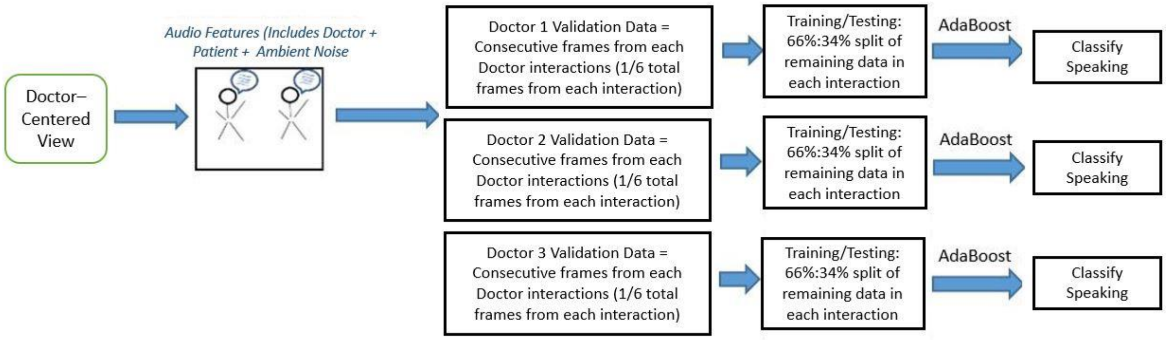

Doctor–Centered View
Audio Features (Includes Doctor + Patient + Ambient Noise
Doctor 1 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Doctor 2 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Doctor 3 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Training/Testing: 66%:34% split of remaining data in each interaction
AdaBoost
Classify Speaking

Figure 19. Phase III Classification of Speaking: MFCC Based Audio Features

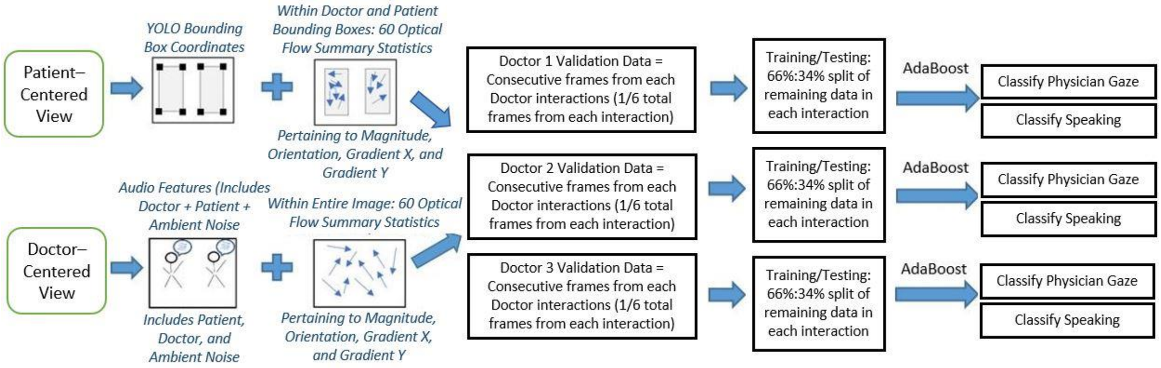

Patient–Centered View
YOLO Bounding Box Coordinates
Within Doctor and Patient Bounding Boxes: 60 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Doctor–Centered View
Audio Features (Includes Doctor + Patient + Ambient Noise
Includes Patient, Doctor, and Ambient Noise
Within Entire Image: 60 Optical Flow Summary Statistics
Pertaining to Magnitude, Orientation, Gradient X, and Gradient Y
Doctor 1 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Doctor 2 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Doctor 3 Validation Data = Consecutive frames from each Doctor interactions (1/6 total frames from each interaction)
Training/Testing: 66%:34% split of remaining data in each interaction
AdaBoost
Classify Physician Gaze
Classify Speaking

Figure 20. Phase III Classification of Physician Gaze and Speaking: Patient-Centered YOLO Bounding Box Coordinates + Patient-Centered Optical Flow + Physician-Centered Optical Flow + MFCC based audio features

5.5. Phase III Results and Analysis

Tables 22–25 present the mean accuracy, sensitivity, and specificity scores for the classification of physician gaze and speaking annotations on test and defined validation data. There are two forms of validation data, the first formed by sequential frames in each interaction and the second formed by a random assortment of frames in each interaction. The results are presented according to each physician. For predicting physician gaze annotations, the model including only visual parameters (pertaining to optical flow and positioning) is defined as *Visual: P.G.* The model incorporating audio (pertaining to MFCCs) and visual information (pertaining to optical flow and positioning) to predict physician gaze annotations is defined as *Audio + Visual: P.G*. For predicting speaking annotations, the model including only audio parameters (pertaining to MFCCs) is defined as *Audio: Speaking.* The model incorporating audio (pertaining to MFCCs) and visual information (pertaining to optical flow and positioning) to predict speaking annotations is defined as *Audio + Visual: Speaking*.

Table 22. Phase III Annotation Prediction Results: Training Data (P.G = Physician Gaze)

	Physician 1			**Physician 2**			**Physician 3**		
Prediction Method	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
Visual: P.G	99%	99%	99%	98%	97%	99%	96%	96%	95%
Audio + Visual: P.G	100%	100%	100%	100%	100%	100%	100%	100%	100%
Audio: Speaking	90%	93%	92%	84%	80%	95%	93%	98%	93%
Audio + Visual: Speaking	97%	99%	97%	92%	93%	97%	97%	99%	97%

Table 23. Phase III Annotation Prediction Results: Test Data (P.G = Physician Gaze)

	Physician 1			**Physician 2**			**Physician 3**		
Prediction Method	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
Visual: P.G	94%	93%	94%	92%	89%	94%	88%	91%	85%
Audio + Visual: P.G	97%	96%	97%	96%	95%	97%	95%	96%	94%
Audio: Speaking	84%	88%	87%	79%	72%	93%	87%	94%	86%
Audio + Visual: Speaking	92%	96%	93%	86%	86%	95%	91%	96%	90%

Table 24. Phase III Annotation Prediction Results: Sequential Validation Data (P.G = Physician Gaze)

	Physician 1			**Physician 2**			**Physician 3**		
Prediction Method	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
Visual: P.G	77%	84%	68%	70%	69%	72%	67%	73%	61%
Audio + Visual: P.G	77%	82%	70%	76%	67%	83%	60%	64%	55%
Audio: Speaking	76%	83%	80%	64%	43%	90%	74%	91%	69%
Audio + Visual: Speaking	78%	83%	84%	71%	56%	93%	77%	93%	75%

Table 25. Phase III Annotation Prediction Results: Random Validation Data (P.G = Physician Gaze)

	Physician 1			**Physician 2**			**Physician 3**		
Prediction Method	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**	**Acc**	**Sens**	**Spec**
Visual: P.G	93%	92%	93%	91%	88%	93%	87%	89%	84%
Audio + Visual: P.G	96%	95%	96%	95%	94%	96%	93%	94%	91%
Visual: Speaking	93%	92%	93%	91%	88%	93%	87%	89%	84%
Audio + Visual: Speaking	90%	95%	92%	85%	85%	94%	89%	96%	89%

For all three physicians, with respect to the test labels for both physician gaze and speaking annotations, the Audio + Visual models proved to be most robust/statistically significant in terms of accuracy, sensitivity, and specificity. In terms of physician gaze for the Audio + Visual model performed for each physician, classification met or exceeded 94% accuracy, sensitivity, and specificity on each physician. In terms of speaking for the Audio + Visual model performed for each physician, classification met or exceeded 86% accuracy, sensitivity, and specificity on each physician.

The purpose of introducing sequential frames within the validation data was to test how each classification model performed when introduced to a large chain of data points which were removed in time and spatial proximity from the test data. In this way, we hoped to simulate the usage of validation data from additional interactions while keeping the camera positioning, doctor, and patient constant. To demonstrate that the large differences between test data and sequential validation data were due to unincorporated information within the models rather than overfitting, we also provide results for randomly assorted validation data. The random validation data did not necessarily contain many data points which were far disparate, both temporally and spatially, from the test data.

Regarding the classification of physician gaze annotations on the sequential validation data, for both Physician 1 and Physician 2 the accuracy met or exceeded 75%, while for Physician 3 the accuracy was 60%. In terms of predicting speaking annotations on the sequential validation data, for all three physicians the accuracy met or exceeded 71%. Thus, the sequential validation data, while proving somewhat more susceptible than the test data to error due to the introduction of new information, classifies both physician gaze and speaking annotations in a manner that can aid the extraction of features in the future creation of a clinician feedback system. The superiority of the classification results on the random validation data over the classification results on the sequential validation data can be attributed more to the uniqueness of the sequential validation data in comparison to the training and test data than to overfitting.

Chapter 6: Discussion

This work represents the initial step toward the creation of an automated system to provide interactive feedback systems for physicians in primary care environments with the goal of improving the efficacy of patient-provider clinical interactions. While in this work we focused on the verbal characteristic of speech recognition and the nonverbal characteristic of physician gaze, the presented methodologies can be extended to capture other parameters of the clinical interactions such as eye-contact, turn-taking, and screen sharing, which have themselves been shown in certain studies to impact patient outcomes [1, 2].

The goal of Phase I was to demonstrate that feature extraction techniques could be used as part of a classification process to predict an aspect of the clinical interaction. We successfully segmented features by leveraging search spaces and thresholding within the HSI and RGB color in conjunction with a classification system of AdaBoost and Markov Chains. The results of our analysis in Phase 1 demonstrated that – for two physicians and a total of six patients – we could perform two-class predictions regarding the object of physician gaze with an excess of 90% accuracy. However, due to the problem of poor segmentation in the feature extraction phase, significant amounts of manual labeling and hyper-parameter setting were necessary to facilitate the feature extraction process.

The goal of Phase II became to improve upon the findings of Phase I by increasing both automation and classification results. We analyzed the same two physicians and six patients from

Phase I while expanding the data collection to synchronized multiple cameras (Patient-Centered and Physician-Centered) and decreasing the amount of manual input necessary in the feature extraction and classification stages. The usage of YOLO and optical flow allowed us to automatically extract energy level measurements specific to Patient and Physician in the Patient-Centered videos without adjusting hyper-parameters between interactions, whereas the application of optical flow in the Physician-Centered videos automatically extracted energy flow information from the entirety of each Physician-Centered frame. After training and testing an AdaBoost classifier within each interaction, we achieved accuracy scores in excess of 90% on the test and validation data (from within same interaction). In Phase II we demonstrated that within the framework of two physicians and six patients, the use of YOLO and optical flow could solve the problem of poor segmentation while improving the classification results in the prediction of physician gaze.

However, the findings from Phase II were limited by two factors. First, the classification models were successfully trained, tested, and validated according to each patient. The generalizability of the results was thereby limited, and thus the goal of Phase III became to create a more expansive, diversified system of analysis while also enlarging the nature of the analysis to include automatically extracted audio information, an additional optical flow summary statistic, the prediction of speaking annotations, and sequential validation data.

In Phase III, which operated upon three physicians, each with 5-6 patients, we demonstrated that audio data could be used to predict speaking annotations with at least 77% accuracy, while visual data from images could be used to predict physician gaze with at least 90% accuracy. By combining the audio and visual data, we were able to make predictions at or in excess of 95% for physician gaze and at or in excess of 86% for speaking annotations. Furthermore, on sequential validation data, the classification system combining audiovisual information predicted physician gaze with an average accuracy at or in in excess of 60%, while the predictions for speaking annotations met or exceeded 71%. Across a broadened spectrum of physicians, the Phase III classification system demonstrated the capability to use audiovisual information to automate verbal and non-verbal annotations of human behavior.

Chapter 7: Conclusions and Future Work

The goal of this thesis has been to build toward the creation of a system to extract audio and visual features in the context of naturalistic, non-simulated interactions, with the final goal of objectively and efficiently labeling video data with verbal and nonverbal characteristics of patient-provider interactions. In this work, we effectively demonstrated how the usage of the YOLO algorithm, optical flow, and MFCC coefficients can overcome the problem of poor segmentation to extract features regarding body positioning, energy output, and audio pitch perception. These features can then be mapped via classification systems to accurately predict physician gaze and speaking annotations on both test and sequential validation data. Moving forward, we seek to gather additional information and advance our algorithms to further develop our predictions of physician gaze and speaking annotations into meaningful predictions of eye-contact and turn-taking. From there, based upon the research of Schneider et al. [40] – 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 – we will seek to map positioning information, energy flows, and audio pitch perception to patient ratings and outcomes of physician behavior as well as measures of physician burnout in the context of EHR usage. We expect that the use of MFCCs, optical flow, and YOLO will enhance the understanding of the effects of different forms of EHRs on physician behavior and further inform the design of more efficient, effective EHRs to enhance the quality of the physician-patient interaction so as to also reduce physician burnout. Ultimately, the proposed work has the potential to inform and aid the design of technologies for capturing interactions from multiple view video data and providing real-time feedback to physicians within the context of EHR usage to facilitate reduced levels of physician burnout and improved measures of patient care.

Acknowledgement

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, S. Reis, “Electronic medical record use and physician-patient communication: an observational study of Israeli primary care encounters,” Patient Education and Counseling. Volume 64, pp. 134–141, April 2006. https://www.sciencedirect.com/science/article/pii/S073839910500090X.

[2] E.R. Melnick, L.N. Dyrbye, C.A. Sinsky, M. Trockel, C.P. West, L. Nedelec, M.A. Tutty, and T. Shanafelt, “The Association Between Perceived Electronic Health Record Usability and Professional Burnout Among US Physicians,” Mayo Clinic Proceedings, Volume 0, September 2019. https://www.mayoclinicproceedings.org/article/S0025-6196(19)30836-5/pdf.

[3] B. Karsh, “Beyond Usability: Designing Effective Technology Implementation Systems to Promote Patient Safety,” Quality & Safety in Health Care, Volume 13, pp. 388–394, October 2004. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1743880/.

[4] C.P. West, L.N. Dyrbye, and T.D. Shanafelt, “Physician burnout: contributors, consequences, and solutions,” Journal of Internal Medicine. Volume 283, pp. 516–529, March 2018. https://onlinelibrary.wiley.com/doi/full/10.1111/joim.12752.

[5] R.S. Beck, R. Daughtridge, and P.D. Sloane, “Physician-Patient Communication in the Primary Care Office: A Systematic Review,” Journal of the American Board of Family Practice, Volume 15, pp. 25–38, January 2002. https://pdfs.semanticscholar.org/47a5/56f6701148ae43c3d0c11797b34b8b927bdc.pdf.

[6] Y. Hart, E. Czerniak, O. Karnieli-Miller, A.E. Mayo, A. Ziv, A. Beiegon, A. Citron, and U. Alon, "Automated Video Analysis of Non-verbal Communication in a Medical Setting," Frontiers in Psychology," Volume 7, August 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4993763/.

[7] 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. https://www.researchgate.net/publication/282849046_The_role_of_nonverbal_communication_in_medical_interactions_Empirical_results_theoretical_bases_and_methodological_issues.

[8] D. Roter and S. Larson, "The Roter Interaction Analysis System (RIAS), Utility and Flexibility for Analysis of Medical Interactions," Journal of Patient Education and Counseling, Volume 46, issue 4, 2002. https://www.sciencedirect.com/science/article/pii/S0738399102000125

[9] 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, Volume 51, pp. 259–266, November 2003. https://www.sciencedirect.com/science/article/pii/S0738399102002252.

[10] P. Little, P. White, j. Kelly, 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," Volume 65, pp. 357–635, May 2015. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4439825/.

[11] H. Ishikawa, H. Hashimoto, M. Kinoshita, S. Fugimori, T. Shimizu, and E. Yano, "Evaluating Medical Students' Non-Verbal Communication During the Objective Structured Clinical Examination," Medical

Education Journal, Volume 40, pp. 1180–1187, November 2006. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2929.2006.02628.x.

[12] K.B. Haskard, S.L Williams, M.R. DiMatteo, J. Heritage, and R. Rosenthal, "The Provider's Voice: Patient Satisfaction and the Content-filtered Speech of Nurses and Physicains in Primary Medical Care," Journal of Nonverbal Behavior. Volume 32, pp. 1–20, March 2008. https://link.springer.com/article/10.1007/s10919-007-0038-2.

[13] J.M. Bensing, J.J. Kerssens, and M. van der Pasch,"Patient Directed Gaze as a Tool for Discovering and Handling Psychosocial Problems in General Practice," Journal of Nonverbal Behavior, Volume 19, pp. 223–242, Winter 1995. https://link.springer.com/article/10.1007/BF02173082.

[14] R. Gorawara-Bhat, D.L. Dethmers, and M.A. Cook, "Physician Eye Contact and Elder Patient Perceptions of Understanding and Adherence, Patient Education and Counseling," Volume 92, pp. 375–380, September 2013. https://www.sciencedirect.com/science/article/pii/S073839911300089X.

[15] O. Asan and E. Montague, "Using Video-Based Observation Research Methods in Primary Care Health Encounters to Evaluate Complex Interactions," Informatics In Primary Care, Volume 21, pp. 161-170. 2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350928/.

[16] G. Skantze, "Real-Time Coordination in Human-Robot Interaction Using Face and Voice," AI Magazine, Volume 37, pp. 19–31, 2016. http://www.speech.kth.se/prod/publications/files/4086.pdf.

[17] D. Bohus, E. Kamar, and E. Horvitz, "Towards Situated Activity Management; Representation, Inference, and Decision Making," Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/2015/07/sigdial_11.pdf.

[18] O. Asan, J. Xu, and E. Montague, "Dynamic Comparison of Physicians Interaction Style with Electronic Health Records in Primary Care Settings," Journal of General Practice, Volume 2, 2013.

[19] J. O. Asan, H.N. Young, B. Chewning, and E. Montague, "How physician electronic health record screen sharing affects patient and physician non-verbal communication in primary care," Patient Counseling and Education, Volume 98, pp. 310–316, March 2015.https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4319541/.

[20] Avid Media Composer. (2019.7), Avid Technology. Accessed: July 28, 2019, [Video Editing Software].

[21] P.H. Zimmerman, J.E. Bolhuis, A. Willemsen, E.S. Meyer, L.P.J.J. Noldus, "The Observer XT: A Tool for the integration and Synchronization of Multimodal Signals," Behavioral Research Methods, Volume 41, pp. 731–735, August 2009. https://link.springer.com/article/10.3758%2FBRM.41.3.731.

[22] O. Friard and M. Gamba. "BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations," Methods in Ecology and Evolution, Volume 7, pp. 1325–1330, 2016. https://link.springer.com/article/10.3758%2FBRM.41.3.731.

[23] D. Gutstein, E. Montague, J. Furst, and D. Raicu, "Hand-Eye Coordination: Automating the Annotation of Physician-Patient Interactions," 19th IEEE International Conference on BioInformatics and BioEngineering, October 2019.

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

[25] A. Rosenfield and J.L Pealtz, "Sequential Operations in Digital Picture Processing. Journal of the Association for Computing Machinery," Volume 13, 471–494, October 1966. https://dl.acm.org/citation.cfm?id=321357.

[26] W.A Belson, “Matching and Prediction on the Principle of Biological Classification,” Journal of the Royal Statistical Society: Series C (Applied Statistics), Volume 8, 65–75, June 1959. https://www.jstor.org/stable/2985543?seq=1#metadata_info_tab_contents.

[27] Y. Freund and R.E Schapire, “Experiments with a New Boosting Algorithm, Machine Learning: Proc. of the Thirteenth International Conference,” pp. 148–156, 1996. https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf.

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

[29] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conf. on Computer Vision and Pattern Recognition,” pp. 779–788, 2016, https://arxiv.org/abs/1506.02640..

[30] A. Ponnusamy, “YOLO Object Detection with OpenCV and Python,” July 2018. [Online]. [Accessed June 21, 2019], https://pysource.com/2019/06/27/yolo-object-detection-using-opencv-with-python/.

[31] B. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision. Proc. of Imaging Understanding Workshop,” pp. 121–130, April 1981, https://www.ri.cmu.edu/publications/an-iterative-image-registration-technique-with-an-application-to-stereo-vision-darpa/.

[32] D. Gutstein, E. Montague, J. Furst, and D. Raicu, “Optical Flow, Positioning, and Eye Coordination: Automating the Annotation of Physician-Patient Interactions,” IEEE International Conference on BioInformatics and Biomedicine, November 2019.

[33] R. Jain, R. Kasturi, B.G. Schunck ,“Machine Vision,” McGraw-Hill, 1995. http://www.cse.usf.edu/~r1k/MachineVisionBook/MachineVision.pdf.

[34] “estimateFlow,” *Mathworks,* [Online]. [Accessed: June 10, 2019]. https://www.mathworks.com/help/vision/ref/opticalflowhs.estimateflow.html#buqx_u5-1-opticFlow.

[35] “How can I determine the angle between two vectors in MATLAB,” *Mathworks,* February 2016. [Online]. [Accessed: Aug. 14, 2019]. https://www.mathworks.com/matlabcentral/answers/101590-how-can-i-determine-the-angle-between-two-vectors-in-matlab.

[36] H. Fayek. “Speech Procesing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficents (MFCCs) and What’s In-Between,” April 2016, https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

[37] H. Fayek. “Speech Procesing for Machine Learning: Filter banks, Mel-Frfequency Cepstral Coefficents (MFCCs) and What’s In-Between,” April 2016, https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html.

[38] Md. Sahidulla and G. Saha, “Design, Analysis, and Experimental Evaluation of Black Based Transformation in MFCC Computation for Speaker Recognition,” Journal of Speech Communication, Volume 54, pp. 543–565, May 2012, https://www.sciencedirect.com/science/article/pii/S0167639311001622?via%3Dihub.

[39] “MFCC, Extract mfcc, log energy, delta, and delta-delta of audio signal,” *Mathworks,* [Online]. [Accessed: October 3, 2019], https://www.mathworks.com/help/audio/ref/mfcc.html.

[40] Schneider, S.H. Kaplan, S. Greenfield, W. Li, and I.B. Wilson, “Better physician-patient relationships are associated with higher reported adherence to antiretroviral therapy in patients with HIV infection,”

Journal of Gen. Internal Medicine, Volume 19, pp. 1096–1103, November 2004. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1494791/.

[41] G. Littlewort, M. Bartlett, L.P. Salamanca, and J. Reilly, "Automated 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. https://ieeexplore.ieee.org/document/5771418.