Correlating Visual Speaker Gestures with Measures of Audience Engagement to Aid Video Browsing

John Ruoyu Zhang

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

©2013
John Ruoyu Zhang
All Rights Reserved

ABSTRACT

Correlating Visual Speaker Gestures with Measures of Audience Engagement to Aid Video Browsing

John Ruoyu Zhang

In this thesis, we argue that in the domains of educational lectures and political debates, speaker gestures can be a source of semantic cues for video browsing. We hypothesize that certain human gestures, which can be automatically identified through techniques of computer vision, can convey significant information that are correlated to audience engagement.

We present a joint-angle descriptor derived from an automatic upper body pose estimation framework to train an SVM which identifies *point* and *spread* poses in extracted video frames of an instructor giving a lecture. Ground-truth is collected in the form of 2500 manually annotated frames covering 20 minutes of a video lecture. Cross validation on the ground-truth data showed classifier F-scores of 0.54 and 0.39 for point and spread poses, respectively. We also derive an attribute for gestures which measures the angular variance of the arm movements from this system (analogous to arm waving).

We present a method for tracking hands which succeeds even when left and right hands are clasping and occluding each other. We evaluate on a ground-truth dataset of 698 images with 1301 annotated left and right hands, mostly clasped. Our method performs better than baseline on recall (0.66 vs. 0.53) without sacrificing precision (0.65 for both) toward the goal of recognizing clasped hands. For tracking, it results in an improvement over a baseline method with an F-score of 0.59 vs. 0.48. From this, we are able to derive hand motion-based gesture attributes such as velocity, direction change and extremal pose.

In ground-truth studies, we manually annotate and analyze the gestures of two instructors, each in a 75-minute computer science lecture using a 14-bit pose vector. We observe "pedagogical" gestures of punctuation and encouragement in addition to traditional classes

of gestures such as deictic and metaphoric. We also introduce a tool to facilitate the manual annotations of gestures in video and present results on their frequencies and co-occurrences. In particular, we find that 5 poses represent 80% of the variation in the annotated ground truth.

We demonstrate a correlation between the angular variance of arm movements and the presence of those conjunctions that are used to contrast connected clauses ("but", "neither", etc.) in the accompanying speech. We do this by training an AdaBoost-based binary classifier using decision trees as weak learners. On a ground-truth database of 4243 video clips totaling 3.83 hours, each with subtitles, training on sets of conjunctions indicating contrast produces classifiers capable of achieving 55% accuracy on a balanced test set.

We study two different presentation methods: an *attribute graph* which shows a normalized measure of the visual attributes across an entire video, as well as *emphasized subtitles*, where individual words are emphasized (resized) based on their accompanying gestures. Results from 12 subjects show supportive ratings given for the browsing aids in the task of providing keywords for video under time constraints. Subjects' keywords are also compared to independent ground-truth, resulting in precisions from 0.50–0.55, even when given less than half real time to view the video.

We demonstrate a correlation between gesture attributes and a rigorous method of measuring audience engagement: electroencephalography (EEG). Our 20 subjects watch 61 minutes of video of the 2012 U.S. Presidential Debates while under observation through EEG. After discarding corrupted recordings, we retain 47 minutes worth of EEG data for each subject. The subjects are examined in aggregate and in subgroups according to gender and political affiliation. We find statistically significant correlations between gesture attributes (particularly extremal pose) and our feature of engagement derived from EEG. For all subjects watching all videos, we see a statistically significant correlation between gesture and engagement with a Spearman rank correlation of $\rho = 0.098$ with $p < 0.05$, Bonferroni corrected. For some stratifications, correlations reach as high as $\rho = 0.297$. From these results, we conclude what gestures can be used to measure engagement.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I wish to thank my advisor John Kender for providing insightful guidance on matters relating to research, academia and beyond. The depth and breadth of his knowledge and experience is truly multimodal and inspirational. It goes without saying that this thesis would not exist without his guidance. I am deeply grateful for the opportunity to work with him.

Next, I wish to thank the other members of my thesis committee. Shih-Fu Chang for his input and collaboration through Brown Institute and Aladdin projects. Steven Feiner for his guidance throughout my PhD, having sat on my candidacy and proposal committees as well. Paul Sajda for giving me the opportunity to learn about EEG and collaborate on the political debates neural correlates project. Yang Song for being an amazing mentor and giving me the opportunity to learn about computer vision in an industrial setting through multiple internships at Google, Inc.

I also wish to thank the members of the Columbia Computer Science Department. My previous labmates Michele Merler and Mitch Morris and my current labmate Chun-Yu (Claire) Tsai for riveting discussions. The undergraduate and master's students with whom I've had the pleasure of working with throughout the years and whom have provided me with invaluable research assistance: Cipta Herwana, Kuangye Guo, Nithin Chandrasekharan, Gaurav Agarwal, Aman Agarwal, Bonan Liu, Xiaolong Jiang, Fengjiao Jiang and Xiang Ma. My office mates Alex Berg, Chris Reiderer and Avner May for their good company. Of course, I'd also like to thank the administrative staff and the staff of Computing Research Facilities whom have supported my ancillary needs throughout the years: Twinkle Edwards, Elias Tesfaye, Lily Secora, Jessica Rosa, Patricia Hervey, Remi Moss, Cindy Walters, Daisy Nguyen, Paul Blaer, Shlomo Hershkop, Hi Tae Shin and Quy O.

One of my final projects involved a brief foray into neuroscience and EEG—fields com-

pletely foreign to me prior. In this regard, I am grateful for the incredible opportunity to collaborate with Paul Sajda, Jason Sherwin and Jacek Dmochowski, from whom I have learned so much. I'm also grateful to the rest of the Laboratory for Intelligent Imaging and Neural Computing for providing emergency assistance during my experiments.

At the beginning of my program, I was honored to be supported by a Fulbright Science and Technology Award, sponsored by the U.S. Department of State. This not only supported me financially but also gave me access to a community of like-minded peers across the range of sciences.

Toward the end of my program I also had the honor of being included in the Brown Institute family as a Brown Fellow. Like the Fulbright, the Brown Institute not only supported me financially, but provided opportunities to exchange ideas with a group of brilliant individuals. I am most grateful to Bern Girod and Mark Hansen for this opportunity.

At Google, Inc, I would like to thank my host Yang Song, who was a most patient and supportive mentor. I am also grateful to a number of other talented Googlers, especially Thomas Leung, Howard Zhou, Jingbin Wang, Zhongli Ding and Chuck Rosenberg as well as the members of both the Video Content Analysis team and the Image Content team. It is rare to have this opportunity to simultaneously have so much fun, be fed so well and learn so much.

At Alcatel-Lucent Bell Laboratories, I would like to thank Yansong (Jennifer) Ren and Fangzhe Chang for being supportive and dedicated mentors. I would also like to thank the rest of the Service Infrastructure Research Department for hosting me that summer and allowing me to take part in such exciting research in the historic Murray Hill Bell Labs.

Last but not least, I wish to thank my friends and family who have given me their love and support through these years, which were not without its challenges. To my girlfriend Ha Lee Kim, for her love, company and dedication: *saranghae!* To my parents Yaying Wang and Ruichuan Zhang, who have given me unconditional love and support all my life, I am eternally grateful. It is to them that I dedicate this thesis.

To my family.

Chapter 1

Introduction

1.1 Motivation

The explosive growth of digital video has led to an abundance of unstructured video content online. Two domains are increasingly popular.

First, educational material online in the form of Massive Open Online Courses (MOOCs) and digitally recorded video of traditional classroom lectures distributed over the Internet by post-secondary institutions[1,2]. As of writing, one online video lecture repository[3] reports 14,651 lectures by 11,138 authors.

Second, digital video of political events have been made available online for greater dissemination as well as analysis in the growing field of data sciences in journalism[4]. In this work, we focus on segments of the 2012 US Presidential Debates.

While the greater availability of video data provides countless benefits to the public, it results in the problem of information overload to viewers. Non-linear semantic video browsers such as the VAST Multimedia Browser [Haubold and Kender, 2007; Merler and Kender, 2009; Morris and Kender, 2011] present one possible solution, whereby multimedia features of videos are used to provide semantic cues to a video browser. These cues in turn

[1]Columbia Video Network, *http://www.cvn.columbia.edu*.

[2]UC Berkeley Online Learning, *http://learn.berkeley.edu, http://webcast.berkeley.edu*

[3]Videolectures.net *http://www.videolectures.net*.

[4]Brown Institute for Media Innovation, *http://brown.columbia.edu*.

offer ways for the users to browse or skim through video in a non-linear fashion.

The focus of this research is to explore one possible source of semantic cues: speaker gestures. We hypothesize that human gestures can convey significant information which can be correlated to audience engagement. The gestures can be automatically identified using techniques of computer vision and used as features and indices in video browsers.

1.2 Domain

In this work, we focus on videos from two domains: educational lectures and presidential debates. They share the common property that videos in both tend to focus on a single speaker at any given time, speaking and gesticulating in front of a (usually invisible) audience, which simplifies our task. The audio is also similar, as it is limited to the voice of a single speaker with no background noise. While we use the accompanying speech (either from manual transcription or ASR) in Chapter 5 and subjects are presented audio along with video as a baseline in Chapter 6, we do not directly explore the correlation between gestures and low-level audio features in this thesis.

1.2.1 Educational Lectures

We use two sources for lecture videos: the Columbia Video Network (CVN) and MIT OpenCourseWare[5] (MIT-OCW). While media from the former source are proprietary, media from the latter are available for public use. Examples of video from CVN and MIT-OCW are shown in Figures 1.1 and 1.2, respectively. We do not use videos from MOOCs as the videos of the instructors focus on little more than the face, making gesture analysis infeasible.

The videos were recorded by amateur cameramen through an ad-hoc capture system. As such, there is no pattern to their focus and camera effects such as pans and zooms occur at irregular intervals. The footage of the instructors are sometimes intermixed with shots of the instructor's slides. Minimal post-production work has been done on these videos. Editing is generally limited to the addition of a title screen listing the course name and

[5]MIT OpenCourseWare, *http://ocw.mit.edu*.

Figure 1.1: Examples of recorded lectures from the Columbia Video Network.

Figure 1.2: Examples of recorded lectures from MIT OpenCourseWare.

instructor.

The CVN videos are approximately 75 minutes long each, and are recorded at a resolution of 352×240 pixels at 29.97 frames per second. The MIT-OCW videos vary in duration and have a slightly higher resolution of 480×270 or 478×360 pixels, recorded at 15 frames per second.

The set of videos span different courses and subject matter. There can be background variation across different courses as different classrooms may be used and the instructors may prefer using either the blackboard, whiteboard, or projector screen.

There is also heavy foreground (i.e., the instructor) variation, as the lighting conditions were varied as well as the clothes and overall appearance of the instructors. All videos feature a single English-speaking instructor with varying ethnicities and genders.

1.2.2 Presidential Debates

We use recorded video of the first (original air date October 3, 2012) and final (original air date October 22, 2012) 2012 U.S. Presidential Debates between U.S. President Barack Obama and former Massachusetts Governor Mitt Romney, which are publicly available. The videos have a resolution of 640×360 pixels and are recorded at 24.58 frames per second

using a stationary camera, and tend to focus on the upper bodies of the speakers, who gesticulate frequently while they speak. Their hands appear in and out of view. Examples of the videos are shown in Figure 1.3.

Figure 1.3: Examples of the 2012 U.S. Presidential Debates.

These videos have more structure than the educational lecture videos. The number of speakers are limited to the presidential candidates and the debate moderators, and this study will only focus on the candidates. The speakers are dressed formally in a suit and tie, and the background and foreground are clearly distinguishable. In the first presidential debate, the speakers are standing in front of a podium. Shots are restricted to the moderator (with the audience behind him) and frontal and angled views of the speakers by themselves and together. In the final debate, speakers are shown sitting at a table. The camera focus is slightly off-frontal, but the setting is otherwise similar.

We restrict our analysis to these two debates (out of the total four debates that year) and discard the vice-presidential debate in order to reduce the number of speakers (for our audience engagement analysis discussed in Chapter 5), and discard the second presidential debate as the speakers are walking while carrying a microphone, therefore biasing any gesture information that can be derived.

1.3 Contributions

This thesis proposes to use speaker gestures as features for detecting moments of audience engagement. Our contributions can be summarized as follows.

1. **Methods for extracting gesture features automatically.** We propose a number of methods for extracting gestures—both poses and motions—of interest and their attributes. We propose a system for the identification of the poses of *point* and *spread*

gestures (two arms "spread" open and waving), which were identified as being often present at semantically significant moments during manual analysis of videos. We use a joint-angle descriptor derived from an automatic upper body pose estimation framework to train an SVM in order to classify extracted video frames of an instructor giving a lecture. Ground-truth is collected in the form of 2500 manually annotated frames covering approximately 20 minutes of a video lecture. Cross validation on the ground-truth data showed classifier F-scores of 0.54 and 0.39 for point and spread poses. This system can also be modified to produce an attribute for gestures which measures the angular variance of the arm movements, which we use to correlate with parts of speech. This work was initially presented in [Zhang and Kender, 2011].

We also propose a method for tracking hands to improve the accuracy of gesture attributes derived from hand motions. Our algorithm distinguishes when two hands are visually merged together (i.e., clasping) and tracks their positions by propagating tracking information from anchor frames in video. We demonstrate and evaluate on a manually labeled dataset selected primarily for clasped hands with 698 images of a single speaker with 1301 annotated left and right hands. Toward the goal of recognizing clasping hands, our method performs better than baseline on recall (0.66 vs. 0.53) without sacrificing precision (0.65 for both). We also evaluate its tracking efficacy through its ability to affect performance of a naive hand labeling heuristic, resulting in an improvement over the baseline (F-score of 0.59 vs. 0.48 baseline). Once left and right hands are distinguished and tracked, we can derive a number of related gesture attributes including velocity, direction change and extremity of pose. This work was initially presented in [Zhang and Kender, 2013].

2. **Identifying semantically relevant gestures through manual analysis.** We gather ground-truth annotations of gesture appearance using a 14-bit pose vector. We manually annotate and analyze the gestures of two instructors, each in a 75-minute computer science lecture, finding 866 gestures and identifying 126 fine equivalence classes which could be further clustered into 9 semantic classes. We observe these classes encompassing "pedagogical" gestures of punctuation and encouragement, as well as traditional classes such as deictic and metaphoric. The gestures appear to be

both highly idiosyncratic and highly repetitive. We also introduce a tool to facilitate the manual annotations of gestures in video and present results on their frequencies and co-occurrences; in particular, we find that pointing (deictic) and spreading (pedagogical) predominate, and that 5 poses represent 80% of the variation in the annotated ground truth. This work was initially presented in [Zhang *et al.*, 2010].

3. **How gestures are correlated with parts of speech.** We demonstrate a correlation between the variances of natural arm motions and the presence of those conjunctions that are used to contrast connected clauses ("but", "neither", etc.) in the accompanying speech, which we believe can indicate moments of audience interest. An AdaBoost-based binary classifier using decision trees as weak learners classifies videos according to whether its speech content contains such conjunctions using the angular variance of arm movements as a feature. Our database of 3.83 hours of video is segmented into 4243 clips, each with subtitles. We show that training on the set of all conjunctions produces a classifier that performs no better than chance, but that training on sets of conjunctions indicating contrast are capable of achieving 55% accuracy on a balanced test set. This work was initially presented in [Zhang and Kender, 2012].

4. **User interfaces for presenting gesture information to viewers.** We study two different presentation methods: an *attribute graph* which shows a normalized measure of the visual attributes across an entire video, as well as *emphasized subtitles*, where individual words are emphasized (resized) based on their accompanying gestures. Results from a user study with 12 subjects are given, with supportive ratings given for the browsing aids in the task of providing keywords for video under time constraints. Subjects' keywords are also compared to an independent ground-truth, resulting in precisions from 0.50–0.55 even when given less than half real time to view the video. This work was initially presented in [Zhang *et al.*, 2013].

5. **How gestures are correlated with audience EEG.** Due to the difficulty of gauging audience interest through intermediary measures, we seek to build a stronger argument by looking directly into the brain. We asked 20 subjects to watch a total

of 61 minutes of clips of the 2012 U.S. Presidential Debates while under observation through electroencephalography (EEG). After discarding corrupted recordings, we retain a total of 47 minutes of EEG data for each subject. The resultant EEG measurements are combined across subjects both in aggregate and in sub-groups (factored by gender and political affiliation) and processed using an existing method for measuring attentiveness. We correlate the post-processed results against gesture attributes and find statistically significant correlations between gesture attributes (particularly extremal pose) and our feature of engagement derived from EEG. For all subjects watching all videos, we see a statistically significant correlation between gesture and engagement with a Spearman rank correlation of $\rho = 0.098$ with $p < 0.05$, Bonferroni corrected. For some stratifications, correlations reach as high as $\rho = 0.297$ ($p < 0.05$, Bonferroni corrected). From these results, we conclude that certain gestures can be used to engage audiences.

1.4 Organization

The remainder of this thesis is organized as follows.

In Chapter 2 we review the literature in related areas including: the study of gestures from a psychological standpoint, extracting gesture and visual features from video, and methods of measuring audience engagement.

In Chapter 3 we propose a number of computer vision techniques for automatically extracting gesture features from video. In Section 3.1 we discuss an approach for recognizing poses of interest. In Section 3.2 we discuss a method for detecting and tracking left and right hands. The chapter concludes with Section 3.3, where we apply some of the previous results toward the creation of methods for extracting certain features which seek to encapsulate gesture attributes such as how fast an arm is moving or how unusual a pose is.

In Chapters 4 and 5, we describe our progress in studying the relationship between speakers' gestures and the attentiveness of their audience. We begin with Chapter 4, where we propose a tool for labeling gestures and poses in video which in turn allow researchers manually examine lecture videos and identify meaningful gestures and their properties.

From these observations, we derive statistics and discuss results. Inspired by some of these results, we look at the correlation between gestures and measures of audience engagement in Chapter 5. This includes looking at correlations between gestures and parts of speech in Section 5.1, and how to present gesture features to viewers in a video browser by proposing different user interface elements and performing user studies in Section 5.2.

In Chapter 6, we examine in greater detail the correlation between speaker gestures' and viewers' engagement by monitoring 20 subjects, stratified by political affiliation and gender, using electroencephalography (EEG) while they watch clips of presidential debate videos both with and without audio. We show statistically significant correlations between gesture attributes (as described in Section 3.3) and viewers' neural activity even in the presence of speech. We also identify the most important gesture attribute to be extremal poses which speakers can use to recapture audiences' attention.

Finally, we summarize our results, draw conclusions and discuss possible directions for future work in Chapter 7.

Chapter 2

Related Work

In the following, we will review the literature in a number of fields including: the representation of gestures and the study of their semantics as done in the fields of psychology and linguistics, techniques for automatic gesture and pose recognition from computer vision, as well as the study of measuring engagement and attentiveness with a focus on methods using neural activity. This thesis lies at the intersection of these fields.

2.1 Representation of Gestures

We review three prominent schemes for the representation of gestures: FORM, ANVIL and CoGesT.

2.1.1 FORM

A necessary first step is determining the appropriate representation of gestures. One such representation, FORM, encompasses both kinematic and temporal information whereby gestures are represented using graphs of nodes lying on the same timeline as shown in Figure 2.1 [Martell, 2002]. The nodes represent timestamps while the arcs represent events spanning the time between two nodes. Each arc contains a series of *tracks*—essentially a list of information. Two tracks are available for each body part. One track describes the location, scale and orientation of a part, another describes the motion of the part. Objects placed in the tracks are attribute-value pairs that can describe temporal or physical data.

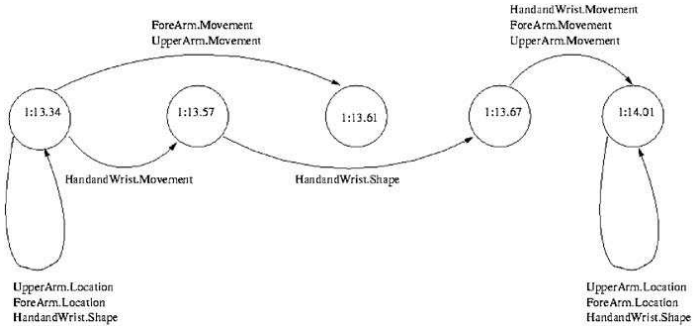


Figure 2.1: Representation of gestures in FORM [Martell, 2002], where nodes represent timestamps and arcs represent events (i.e., motion) spanning the time between nodes.

As FORM is a framework and designed to be extensible, tracks can be added as needed. An arm gesture scheme is given as an example in [Martell, 2002]. To evaluate the scheme, a user study was performed examining inter- and intra-annotator agreement revealing high intra-annotator agreement but low inter-annotator agreement.

For our work, we choose to build a simpler model of gesture representation by deriving from the multi-phasic temporal model of [Kendon, 1980]. This approach lacks the fine-grained temporal modeling capability but allows us to build simpler methods for automatically extracting gestures and their attributes.

2.1.2 ANVIL

ANVIL (ANnotation of VIdeo and Language) is proposed as an annotation scheme for identification and analysis of gestures [Kipp, 2001]. Annotations are made by attaching anchored attribute-value pairs via text input to tracks. These annotations are hierarchical: elements of one track that are time-anchored are primary. Elements can also be defined relative to elements of other tracks. These are secondary. Annotation schemes for ANVIL must be defined in XML (which can easily represent hierarchy through nested tags).

Like FORM, ANVIL is also meant to be generic and extensible, and the use of multiple tracks allows support for multimodal annotation (i.e., annotations regarding text and audio can also be made). However, the given tool for ANVIL annotations (Figure 2.2a) is not similarly extensible and requires purely text input, when a graphical input method may be

ideal for some kinematic information (e.g., drawing a pose).

[Kipp *et al.*, 2007] extend their work with ANVIL by proposing a specific scheme for manually transcribing gestures for the purpose of animating avatars. In this scheme, attributes containing spatial and temporal data (following the tri-phasic gestural model proposed by [Kendon, 1980]) are assigned to the tracks described in the ANVIL annotation scheme. The authors of this work implemented and evaluated their scheme using the ANVIL tool. For evaluation, 420 gestures of two speakers from 18 minutes of TV video are annotated. Evaluation is performed through a user study whereby collected annotations are examined by recreating the gestures. The authors note that phases were annotated with 72% reliability.

Kipp et al.'s proposed scheme presents some strengths over other proposed representations as it captures temporal and spatial data of gestures, with particularly accurate spatial data. However, just one minute of source material takes 90 minutes to encode, given short annotator training.

Our tool proposed in Chapter 4 takes some inspiration from the ANVIL annotation tool. However, to improve annotator speed, it is tailor-made according to our representation model and uses a graphical input method (i.e., posing by avatar) rather than text input.

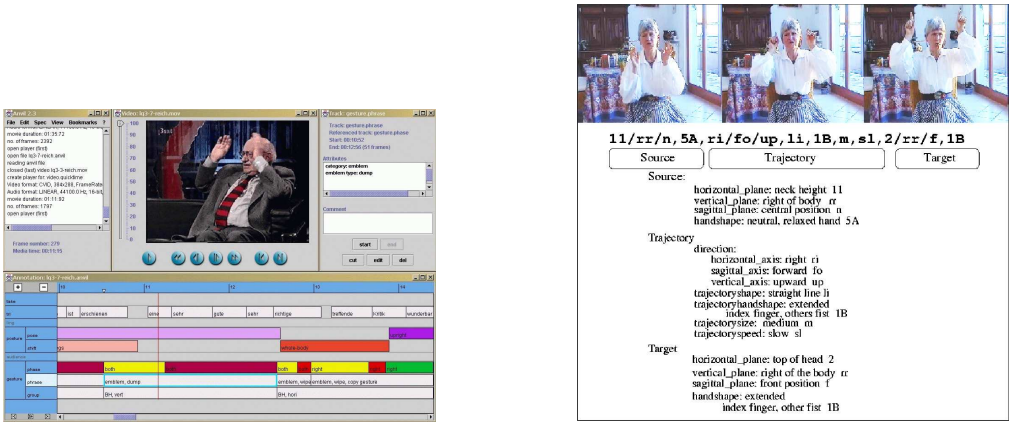
2.1.3 CoGesT

The CoGesT annotation scheme is proposed for capturing conversational speech-related gestures by [Gut *et al.*, 1993]. In this scheme, each gesture is defined according to the following properties: phase (with a source, trajectory and target specified where trajectory is parameterized by direction and shape), hand shape (defined according to an index of hand poses), symmetry (whether the hand gestures are parallel or mirror), and modifiers. Addition information such as speed, number of repetitions, etc., can be added to scheme via the modifiers. An example of a transcription can be seen in Figure 2.2b. The authors evaluate their scheme by comparing inter-annotator agreement and find high-agreement for segmentation at 86% but low agreement for the actual descriptions at 23.4%.

CoGesT presents an advantageous approach whereby form and function of gestures are distinguished. While it is meant to capture conversational gestures, it may be applicable in a classroom setting where we imagine the teacher as having a one-sided conversation with

each student, simultaneously.

However, CoGesT only supports arm and hand gestures with no justification given for this choice. The software tool used to annotate videos also only uses text annotations which proved to be a poor choice, as "typing errors" was listed as one of the reasons for poor inter-annotator consistency. Arm and hand gestures are a cornerstone of this thesis, but ultimately the CoGesT representation and tool were not used in favor of a simpler representation and a graphical tool which is less prone to input errors.



(a) (b)

Figure 2.2: (a) In ANVIL [Kipp, 2001], annotations are made by attaching anchored attribute-value pairs to tracks. (b) An example of a transcription made using the CoGesT scheme [Gut *et al.*, 1993], describing source, trajectory and target states.

2.2 Semantic Relevance of Gestures

In order to review the relationship between gestures and semantics, we examine literature related to the taxonomy of gestures (as specific classes of gestures convey more meaning than others) and how gestures are used to communicate in general and in the classroom and political debates.

2.2.1 Taxonomies of Gestures

Seminal work on the relationship between gestures and language done by [McNeill, 1992] identifies five classes: *iconics*, *metaphorics*, *beats*, *cohesives* and *deictics*. Iconic gestures attempt to illustrate the semantic content of speech, e.g., holding a fist in front and slightly turning it when talking about a steering wheel. Metaphorics are similar to iconics, but whereas iconics describe concrete objects or events, metaphorics are used to depict abstract ideas. Beat gestures are typically simple gestures of emphasis, e.g., a light "beat" of a hand in the air. McNeill describes cohesives as composite gestures (i.e., they consist of the other types of gestures) which signal continuities in thematically related but temporally separated discourse; e.g., a speaker makes a certain gesture when describing an event, makes a different gesture when making a side note, and then returns to the original gesture to signal that they have returned to the original topic. The last class, deictic gestures, are pointing gestures.

These classes appear frequently in the literature and we will develop them further in our own analyses of gestures.

[Pozzer-Ardenghi and Roth, 2004] further develops the research on gesture taxonomies by analyzing classroom videos and proposes eight classes of gestures. The task in all the recorded lectures in the study was to communicate meaning of photographs or figures in a lecture environment.

The eight classes proposed were mutually exclusive and based on function. They are as follows, and can be related to McNeill's classes:

- *Representing.* Gestures used to represent objects or phenomena associated with a photograph.
- *Emphasizing.* Iconic gestures representing something directly in the photo, but referring to an object and so serves a deictic function.
- *Highlighting.* Tends to be a circular motion without referring to anything specific.
- *Pointing.* Deictic gestures.
- *Outlining.* Very specific deictic gestures, but also iconic.
- *Adding.* Like outlining, but traces an abstract object.

- *Extending.* "Added" object can be outside the bounds of the photo.
- *Positioning.* Speaker is positioned in a way that puts her "inside" the photo to provide perspective.

The classes were derived by analyzing videos of classroom lectures, so the gestures (and classifications) arose naturally. Many of these classes will overlap with our own proposed taxonomy as described in Chapter 4. One reason the gestural classes suggested here is insufficient for our purposes is that some of these classes would be difficult to distinguish automatically, e.g., for outlining and adding, objects in the photos would need to be identified and matched against gestures.

2.2.2 Gestures in Communication

It has long been known that gestures play an important role in everyday communication, particularly in some cultures [Donadio, 2013].

[Eisenstein and Davis, 2006] addresses the question of whether or not gestures are multimodal. Human raters are trained to assign one of these labels to gestures as shown in a video: *deictic*, *action* (iconic), *other* (beat) or *don't know*. Their labels are evaluated for inter-annotator agreement. Then, the visual and auditory modalities are alternately removed and the experiment is performed again. The conclusion of the authors is that neither modality alone is sufficient to classify gestures, although vision-only gesture classification is significantly stronger. In addition, an automatic classification method based on linguistics was introduced. For each gesture, a feature vector was constructed using words that appeared within windows surrounding the stroke phase of the gesture. This automatic classifier was found to outperform human classifiers when using audio only, but not as good as humans using both audio and visual information.

The results of this study leads us to focus on gesture recognition using vision-based techniques, although it suggests the potential efficacy of using multimodal data as well to aid in recognition methods. We will also examine the correlation of gestures with parts of speech in Chapter 5.

We further explore the possibility of correlating gestures with semantic significance in

communication by examining the work of [Bavelas *et al.*, 1995]. Here, hand gestures used during dialogue are analyzed. In the first experiment, a person describes a cartoon to another. Their interaction is videotaped and the number of gestures are counted. It was shown the frequency of gestures decreased when the addressee would not see them. The authors identified four classes of objectives of gestures in dialogue: marking the delivery of information, citing the other's contribution, seeking (a response from the addressee) and turn-taking (meant to coordinate the "turn-taking" of speaking during conversations). In a second experiment, the authors attempt to predict an addressee's responses based on the speaker's gestures. A high rate of success is reported.

The correlation between gestures and possible meanings in communication is the main contribution of this work. While a taxonomy of hand gestures is introduced here, they may be unusable for the proposed project as they are classified by intent and function and which are difficult to detect using machine vision.

2.2.3 Gestures in Education

In studies more specifically related to teaching, [Roth and Bowen, 1998] explore the relationships between semiotics, graphs and gestures in education by analyzing an ecology lecture. Three types of shifts in speech are identified which the authors contend correlate to points of misunderstanding between the lecturer and the students: semantic, temporal and structural. Semantic shifts occur when the speech describes something general while the gesture is toward something specific or vice versa. Temporal shifts occur in events where the gestures and utterances are asynchronous. Structural shifts occur when the lecturer digresses in his utterances from the current gesture and topic without changing gestures.

In this work, students, seminars as well as lecturers were studied. The observation that shifts in speech that may be points of misunderstanding between students and lecturers is a significant result that may suggest their usefulness as cues in video browsers. However, analyses here were restricted to lectures from a single course. Furthermore, significant events are identified by the researchers and are subjective.

In another study, [Roth and Lawless, 2002] analyzed videos of various lectures and made observations regarding body orientation, proximity (to an inscription such as a slide,

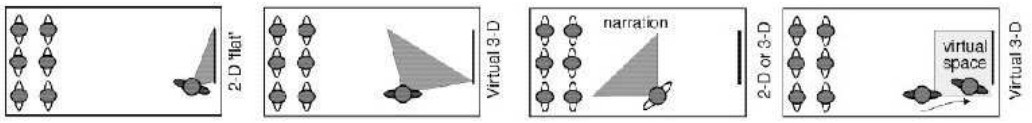


Figure 2.3: Visualization of 2D and 3D inscriptions from [Roth and Lawless, 2002], i.e., where an instructor stands and references an inscription (e.g., a blackboard or photograph) relative to the classroom consisting of 6 listeners. These are used to argue for the importance of a lecturer's position in conveying information.

photo or figure) and gestures, particularly iconic and deictic gestures. In particular, three situations were identified: the use of 2D inscriptions, 3D inscriptions (like 2D inscriptions, but the position of the speaker conveys information) and without using inscriptions (Figure 2.3). The primary contribution of this work was to argue for the importance of body orientation and position in teaching. This affects our proposed research as it implies the importance of body positions and orientation when recognizing and classifying gestures, which is accounted for in one of our studies described in Chapter 4.

[Roth, 2001] presents a survey of research in gestures across lectures in different subjects. The survey argues for the importance of hand and arm gestures relative to body motion and position and argues that gestures can be distinguished from other body movements by their adherence to the multi-phase model proposed by [Kendon, 1980]. Roth et al. further review gestures relative to speech by providing a spectrum, with increasing independence from speech: gesticulations, language-like gestures, pantomime, emblems, sign language.

Significant references are made to McNeill's gestural classes, which are frequently used by education researchers. The authors contend that McNeill's classes of gestures are not equally represented in particular events. Iconic gestures appear more often as references to story events. Deictic, metaphoric and beat gestures appear in references to story structure.

Furthermore, the survey presents two theories for function and production of gestures. First: gestures are a byproduct of speech, in that they do not convey additional meaning. Second: gestures are generated along with speech and are a different way of expressing the same meaning. As of publication, no evidence strongly supports one over the other, although studies done with children indicate that some meaning could be conveyed through

gestures that were sometimes not conveyed through speech.

This study is comprehensive: the gestures surveyed spans across diverse subjects as anthropology, linguistics and psychology. However, the few physical descriptions of the gestures, which are generally distinguished in terms of function and meaning, render them infeasible for direct use in an automatic classification setting and leads us to explore an alternate taxonomy.

The overarching results of Roth et al. and others reviewed here support the feasibility of using gestures as semantic cues in the education domain.

2.2.4 Gestures in Politics

A number of studies have also been done on gestures in the context of political speeches and debates. In many cases, gesture analysis experts (usually psychologists and linguists) are hired in an attempt to infer personality traits about political candidates through their gesticulations while speaking, as done in the 2012 U.S. Presidential debates [Xaquin *et al.*, 2012]. However, our focus is on how the gestures of the speakers elicit reactions from their audience. In the following, we review some literature examining the relationship between the gestures of political candidates and the corresponding spoken content.

[Bull, 1986] examined the relationship between the style and content of political speeches with the gesticulations of the speaker. Three British political speeches made by three different speakers with varying levels of oratory reputation were videotaped and analyzed. The author annotated gestures using his own system [Bull and Connelly, 1985] and annotated speech content according to purpose (e.g., an attack on another speaker, advocacy, naming, addressing, etc.) After analysis, the author noted that gestures were correlated to vocal intonation and often used to elicit and control applause, although this result varied heavily based on speaker. The deliberate use of gestures to control the audience is another reason we believe that gestures are valid as semantic cues.

More recently, [Casasanto and Jasmin, 2010] examined the co-occurrence of certain hand gestures and the sentiment (i.e., positivity or negativity) of the accompanying speech. The authors demonstrate their hypothesis that the sentiment of a person's spoken clauses and the handedness of their gestures are highly correlated. That is, right-handed speakers tend

to say more positive things while gesticulating with their right hand, and similarly for left-handed speakers. This analysis is done by manually examining the 2004 and 2008 U.S. Presidential Debates. Two linguists separate the transcripts into clauses and assign each a sentiment. They then watch the corresponding video and identifying handedness.

In this thesis, we will focus on a number of gesture attributes beyond handedness. As previously mentioned, we also examine the correlation between speaker gesticulation and parts of speech (but not sentiment) in Chapter 5.

2.3 Automatic Pose and Gesture Recognition

To realize our objective of automatically indexing gestures in videos, we examine the related work at various levels: on a frame-by-frame level, we review literature of human detection, pose estimation and body part recognition. At the gesture level, we review literature related to temporal segmentation of gestures and recognition of gestures.

While the proliferation of infrared-based cameras allowing for depth information to be captured along with imagery has led to an explosion of gesture-based applications and research, we restrict our literature review to those methods using only single-camera video information. This is because for our purpose of video analysis, this is usually the only information available.

2.3.1 Human Detection

[Dalal and Triggs, 2005] propose the use of grids of histograms of oriented gradients (HOGs) as features for pedestrian detection. The method works as follows. The detector image window is divided into small cells, each cell accumulating a local 1D histogram of edge orientations over the pixels of the cell. The combination of the histograms form the descriptor. Next, cell histograms are created where each pixel in a cell casts a weighted vote (usually the gradient value) for an orientation-based histogram channel based on the orientation of the gradient element centered on it. Next, local contrast normalization is applied to account for changes in illumination and contrast. The final descriptor is the vector of all components of normalized cell responses from all blocks. Classification is done with a linear SVM.

The end result was shown to be extremely robust against variations in backgrounds and foregrounds with near-perfect classification of the established MIT dataset and good performance on a introduced dataset ("INRIA"). HOGs have become one of the most widely used features for detection of pedestrians and other objects and is a key first step in pose estimation.

We examine similar work on human detection in video. [Niebles *et al.*, 2008] proposed a method for the automatic extraction of moving people in arbitrary videos . The algorithm is divided into two stages. First, a pedestrian detector (i.e., [Dalal and Triggs, 2005]) is applied to provide approximate information regarding the possible locations of persons. Then, a clustering algorithm is applied across time to group together detections and reject false positives and find false negatives based on appearance similarity. In the second step, given the bounding boxes of each person in the video, a pose is estimated for each person using a method such as those reviewed in Section 2.3.2. The algorithm is tested on YouTube videos and demonstrated high precision but low recall.

Similar to this work, several of our methods for extracting upper body-based gesture features are dependent on pictorial structures-based methods for pose estimation.

2.3.2 Pose Estimation

Toward the goal of estimating human body poses, [Ramanan, 2007] proposes an iterative parsing approach which works as follows. For initialization, an edge-based deformable model is used to poorly estimate body parts. In this model, the image is parsed into several body part regions with the aid of low level cues. Then, in a first iteration, a region-based deformable model is used to identify ten specific body parts and estimate body position to build new region models using foreground and background color histogram models. This step is iterated. This method is robust and allows for pose estimation of non-humans and returns a most likely pose. However, it is not robust against occlusion.

Work on pictorial structures have been applied to improve pose estimation. In pictorial structures, objects are represented as a collection of parts arranged in a configuration with pairs connected by "springs" that pull each other to be in a relative location with respect to the other. A challenge arises in that matching a structure to an image can be computa-

tionally intensive. [Felzenszwalb and Huttenlocher, 2000] propose a dynamic programming approach for structures restricted to trees (which are common, such as for humans). Since the structure is a tree, the approach works by starting at leaves and guessing the location of its parents. The approach runs in $O(mn)$ time where m is the number of possible locations for the parts and n is the number of vertices in the structure. This approach is optimal and fast. In their paper, only a simple template matching is used for a cost match function, although more complex possibilities are mentioned.

[Ferrari *et al.*, 2008] applies this approach to the task of estimating upper- and full-body poses in humans. The human (upper) body is modeled with 6 parts, where each part is parameterized by location, scale and orientation. A method is introduced for estimating these parameters. [Yang and Ramanan, 2011] uses a similar parts-based model but does not parameterize parts with orientation, resulting in significantly faster performance. In one interesting application of gesture analysis using pictorial structures-based posed estimation, [Buehler *et al.*, 2009] demonstrates a system whereby British Sign Language is automatically learned from captioned video.

As previously mentioned, we apply these results to our some of our methods to efficiently estimate poses for gesture recognition. The method by Ferrari et al. will be most widely used as the method by Yang et al. is, although faster, less accurate.

2.3.3 Hand and Arm Recognition

In addition to the entire body, a closer look at hand detection and pose recognition is needed.

Accurate hand detection in video can be achieved if augmented with full or upper body pose estimators, as shown by [Buehler *et al.*, 2008], who propose a method of hand detection and tracking via a part-based model which infers the position of the hands based largely on the position of the other body parts, starting with the head. However, when other body parts are not always clearly visible (e.g., in videos focusing on the upper bodies of speakers such as the US presidential debate videos used in this thesis), other features must be relied upon.

Hand detection can be done using a number of features, often combined: skin color,

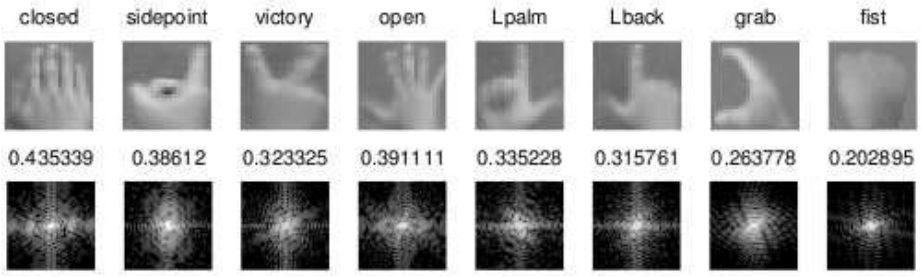


Figure 2.4: Hand pose classes in [Kolsch and Turk, 2004] and their corresponding Fourier transforms, indicating the level of grey level variation, which can be used to separate the classes.

Haar-like features, and HOGs. [Mittal *et al.*, 2011] combine skin color and shape features (as determined by HOGs) to produce hand detectors robust to different poses and backgrounds. However, the method is trained on images of entire hands. Similarly, [Kadir *et al.*, 2004] trained a boosted classifier using Haar-like features to detect entire hands. Such detection techniques would, in general, be unable to detect when one hand is occluded by another hand. Furthermore, our own attempts to train models using these features to recognize partial hands (e.g., just a thumb sticking out) resulted in poor performance. This capability is desirable for the purpose of propagating bounding boxes across frames, when hands may appear in and out of view. However, color and shape features are still useful for hand blob detection.

Once hands are detected, it may be possible to look further by examining the exact poses of the hands. [Kolsch and Turk, 2004] propose a method for hand pose recognition based the object recognition method of [Viola and Jones, 2002]. The main contribution is a technique which greatly reduces the training time than what is necessary for [Viola and Jones, 2002] by estimating class separability without need for training. This is done by using a Fourier transform to describe the amount of grey level variation in an image, for instance, as shown in Figure 2.4. The method was tested on 2300 images of hands in varying conditions (male, female, indoor, outdoor, etc.) and shown to be robust.

During our work in developing gesture attributes derived from hand motions, we sought to train hand detectors based on [Viola and Jones, 2002]. However, due to the high degree

of variability in hand appearance in natural gestures, we ultimately pursued a different approach, as described in Section 3.2.

Related research examines arm appearances. In a classroom-specific work, [Yao and Cooperstock, 2002] seek to detect raised arms. The system they describe assumes a single fixed-camera aimed at the students in a lecture hall with students' heads assumed to be at the same level. Motion is detected by subtracting frames and objects are detected based on an edge map. The edges are used to classify shapes and, combined with a skin color model, used to identify arms. Our own attempts to classify arms using shapes (e.g., parallel lines) often failed due to the low quality of the test video and the poor estimations of forearms. Nevertheless, skin color remains a good feature and is used for hand detection in Chapter 3.

2.3.4 Natural Gesture Recognition

[Wilson *et al.*, 1997] approaches the problem of temporal segmentation of natural gestures according to the gestural phases proposed by [Kendon, 1980] whereby gestures have temporal phases separated by periods of rest. The first step is to identify candidate rest states. This is done by representing video frames in low-dimensional space using eigenvector decomposition then creating a distance matrix to represent the difference between every pair of subsequences of a fixed length. Rest states would correspond to sequences that are repeated often. Next, a Markov state description is used to identify rest states from candidates. The states are rest, stroke and transition. The Viterbi algorithm is used to generate the best possible parse, and if a tested subsequence is indeed a rest state, then the parsed input should spend a significant amount of time in the rest state. The method was evaluated against manually labeled ground-truth.

While this work provided interesting results the authors concede that temporal segmentation of natural gestures is difficult to evaluate as there is currently no objective standard.

[Kettebekov *et al.*, 2003] presents a framework for natural gesture recognition that makes use of audio cues. The main contribution is the use of prosodic signal-level audio features. The authors argued that because gestures and speech are loosely coupled, the use of pure audio is not always helpful, whereas prosodic manifestations may be useful. Deictic gestures

in narrated weather reports are searched for as a benchmark. The authors proposed two approaches for classification: a *feature-based co-analysis* whereby the motion of the head, hand and relative distance in-between as well as the fundamental frequency contour are used as features in a hidden Markov model (HMM) for classification, and a *co-articulation framework* whereby periods of continuous speech or lack of continuous speech are identified and classified based on rises and falls in intonation. Finally, the combined method is tested.

That work presents us with two useful overall results: a description of a HMM-based approach for natural gesture recognition that can use visual and audio features (separately or combined) as well as an quantitative exploration of the multimodal nature of gestures similar to [Eisenstein and Davis, 2006]. Although our work also examines the relationship between gestures and speech (part of which is derived from audio features), low-level audio features are not necessary in our methods for extracting gesture features. Future work could investigate if they could be incorporated to improve accuracy in detection or recognition.

2.4 Measures of Audience Engagement

As our ultimate goal is to demonstrate a correlation between points of interest in videos and the speakers' gestures, then in order to build a ground-truth, we must be able to examine when the audience's attention is piqued. To this end, we briefly review the literature for different methods of measuring audience engagement—itself an unsolved and difficult problem.

2.4.1 Identifying Moments of Engagement in Speech

One way to try to identify when a person is paying attention is by examining the accompanying discourse. We have already reviewed in Section 2.2.3 related work by [Roth and Bowen, 1998] which demonstrated the relationship between students' attention and shifts in speech.

[Grosz and Sidner, 1986] presents a theory of discourse structure which describes attentional state as one of the components supplying key information on how speech is processed. In particular, the authors identify *cue phrases* and *interruptions* as signals for changes in

attentional state. Cue phrases include terms such as "now, next, anyway, etc." while interruptions are more general and can include interruptions in speech by another person (termed *true interruption*), *digressions*, (which can be identified by terms such as "by the way, incidentally") and *satisfaction-precedes* ("first, second, finally, moreover") among others.

The use of conjunctions to mark shifts in discourse is shown across languages, as shown for Japanese by [Watanabe *et al.*, 2007]. The authors examine the correspondence of pauses in discourse (to study prosody) and conjunctions at discourse boundaries. They conclude that both pauses and conjunctions correspond to boundary strengths, but conjunctions show a stronger correspondence.

The heavy use of conjunctions in cue phrases inspires our work in Section 5.1 which explores how different classes of conjunctions may co-occur with gesture with the goal of indirectly identifying these shifts in discourse through gesture. We restrict our domain to the English language.

2.4.2 Methods of Measuring Engagement

A number of different methods exist for capturing a person's engagement. The most straight-forward method is perhaps a direct user study. The field of event recognition in computer vision research uses human subjects to manually identify and label segments of video where an event of interest is present [Cao *et al.*, 2011; Revaud *et al.*, 2013]. Of course, the task we face in this thesis is far more subjective, as what catches a viewer's attention can be highly idiosyncratic. More related to this are user studies in the study of video summarization. [Ma *et al.*, 2002] propose a fully automatic multi-modal video summarization scheme and perform user studies to evaluate its efficacy. The authors recruit 20 subjects to first watch summarized video clips, followed by the original video before assigning scores for "enjoyment" and "informativeness". While this approach may be effective for determining which segments of video may be more relevant *after* their extraction, it would be difficult to apply this to gathering a ground-truth of relevant segments of video without prior candidates.

Another approach, eye tracking—the process by which the gaze is tracked—has found

numerous applications in determining web usability [Buscher *et al.*, 2009], assessing the efficacy of advertising media [Buscher *et al.*, 2010] and in the automotive industry to identify when drivers may be distracted [Strayer *et al.*, 2011]. [Nakano and Ishii, 2010] use eye tracking to estimate a person's engagement in a conversation. In their approach, the subject's gaze is classified as looking at the speaker's face, the speaker's body, or some object. They conclude that the subject is least likely to be engaged when their gaze is not on the speaker (face or body) and propose a system for automatically recognizing these moments.

A third approach uses signals from social networks, particularly Twitter. [Diakopoulos and Shamma, 2010] use Twitter to gauge audience sentiment during the 2008 U.S. Presidential Debates. Tweets are aggregated by searching for related hashtags (e.g., #current, #debate08, etc.) and then assigned a sentiment label manually using Amazon Mechanical Turk. As the Tweets can be time-synced to the debate, the authors are able to derive conclusions regarding audience sentiment with regard to the candidates and to the debate topics. Naturally, there is a lag between a moment in the debate and the corresponding Tweet.

Combining social media and neural activity, [Abelson, 2013] attempts to correlate subjects' neural activity against social media statistics. In the study, subjects are recruited to watch the television show "The Walking Dead" while being monitored using electroencephalography (EEG). The resulting data is averaged across subjects and processed using the method of [Dmochowski *et al.*, 2012] for identifying moments of subject attentiveness, which we will review in greater detail in Section 2.4.3. Abelson discovers a statistically significant correlation between these results and the corresponding number of Tweets found online, used as a proxy for measuring particularly engaging scenes.

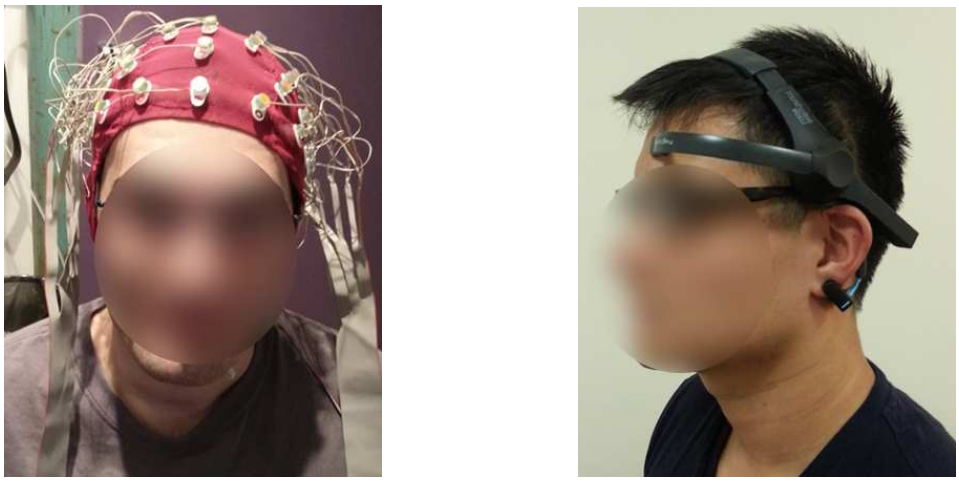
2.4.3 Measuring Audience Engagement from Neural Activity

The methods for measuring engagement reviewed in Section 2.4.2 are indirect. That is, we assume a person is paying attention if they indicate it in a questionnaire, or if their eyes follow a target, or if they Tweet about it. Here, we review methods which attempt to identify moments of engagement at the source: the human brain.

Functional magnetic resonance imaging (fMRI) has been shown to be effective at iden-

tifying attention [Hasson *et al.*, 2004; Hanson *et al.*, 2009]. However, they lack the time resolution we desire to identify exact moments of engagement.

EEG has long demonstrated that it can capture neurological information related to attention. For instance, EEG devices have been approved for use as tests for attention deficit hyperactivity disorder (ADHD) in children [Tavernise, 2013]. A number of devices exist for gathering EEG information. The wet-gel setup, as seen in Figure 2.5a, is most common and accurate method for research purposes. However, as the technology matures, companies are beginning to push simple inexpensive devices for public use (see Figure 2.5b), although these devices remain considerably less accurate.



(a) (b)

Figure 2.5: (a) Subject wearing 64-electrode wet-gel EEG cap. (b) Subject wearing NeuroSky MindWave Mobile dry-sensor EEG headset.

EEG records electrical activity along the scalp. [Dmochowski *et al.*, 2012] propose a method for interpreting this data and identifying moments of heightened engagement in subjects. Intuitively, the method works under the hypothesis that neural activity would be most similar across subjects at moments when something piques their interest. Therefore, the method identifies these moments by computing the points of maximum correlation of EEG measurements across subjects and comparing it against a significance derived from a permutation test.

This is the approach we ultimately apply to gather a ground-truth of when subjects are engaged when watching video. Unlike methods discussed in Section 2.4.2, it is direct, less sensitive to lag (i.e., the subject does not need to interrupt their viewing to give feedback) and less intrusive for the subject.

Chapter 3

Extracting Gesture Features from Video

The current body of work around gesture recognition tends to focus on the class of gesture, such as pointing, waving, etc. This is an important task and drives many gesture-based applications that have grown increasingly popular such as the Microsoft Kinect. Recognition and classification of poses and gestures is important for our goals as well. However, much less attention is given to the minute mid-level features or *attributes* of gestures: how emphatic is the point? How fast are the hands waving? How wide are the arms spread? These attributes are important since they appear to be related to attention, emphasis and impact.

In this chapter, we present our work toward the goal of characterizing gestures. To accomplish this, we:

1. Propose a system for recognizing poses salient to our intended domain in Section 3.1, where salience is determined through manual analysis described in Chapter 4.

2. Propose a method for recognizing and tracking hands even in the case where they often clasp together in Section 3.2. This is a necessary step toward characterizing gestures based on hand motions and positions of both hands.

3. Propose a number of gesture attributes based on the results above in Section 3.3. We use these attributes to find correlations to audience engagement in Chapters 5 and 6.

3.1 Classifying Upper Body Poses

We develop a system toward the goal of identifying gestures salient to teaching in lecture videos. The system allows for the identification of the stroke poses of the *point* and *spread* gestures discussed in Chapter 4. Much of this work was initially presented in [Zhang and Kender, 2011].

Given a lecture video as input, the frames are extracted at a constant rate. We can make some domain-specific assumptions which simplify the problem. We assume that the video focuses on a single person (i.e., the instructor) who tends to be near the center of the video. If multiple persons are detected, whether it is because a student was caught in the frame or because of a false detection, then we can assume that the one with a midpoint closest to the center of the image is that of the instructor. A second assumption is that the person of interest is always standing upright, that is, the torso is vertical (with a small range of flexibility). Anything exceeding this range can be assumed to be a poor estimation.

As an initial step, we apply a human detector [Dalal and Triggs, 2005] on all frames and disregard those with negative results.

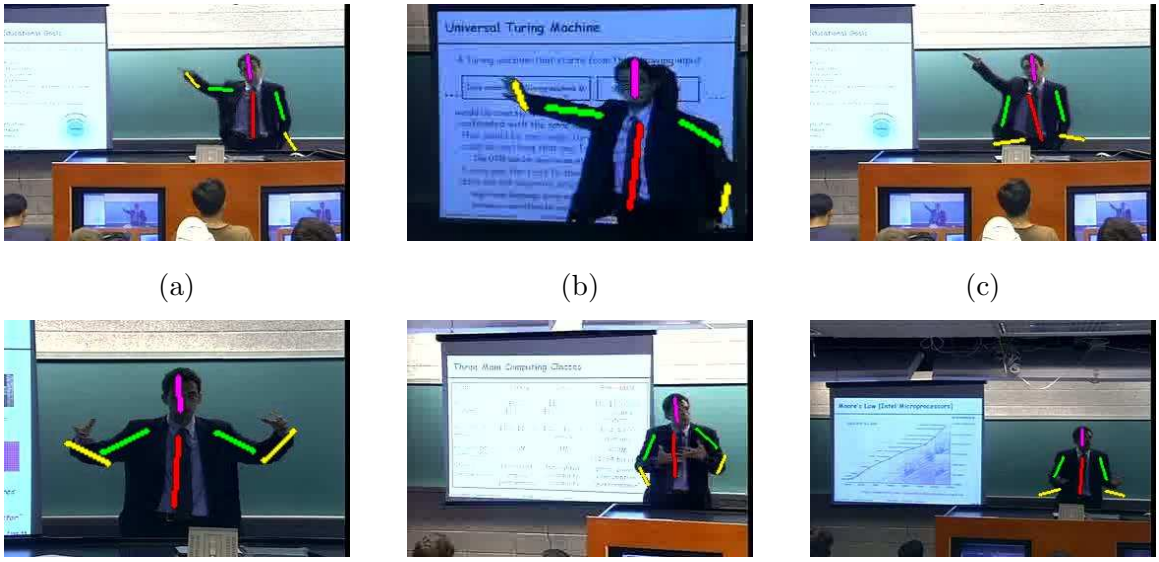
Next, for the actual task of pose identification, we aim to classify each input image as belonging a *point pose*, a *spread pose*, or neither. Point and spread (two arms "spread" open) poses were identified as being often present in gestures at semantically significant moments during manual analysis of videos. For more details regarding their usage, please refer to Chapter 4. We approach the problem by building binary classifiers for each class. We elaborate on the construction of these classifiers as follows.

3.1.1 Pose Estimation

The pose of the instructor in focus is the basis for our descriptors. As discussed in [Roth, 2001; Zhang *et al.*, 2010], the arms and torso play a significant role in gestures in teaching. Therefore, we focus on five body parts: the torso, lower and upper right arm, lower and upper left arm.

To extract the approximate positions and orientations of these body parts from a given image, we use the pose estimation framework introduced by [Ferrari *et al.*, 2008]. This

framework provides the endpoints of each body part in question as output (among other data). Figure 3.1 shows the output of the pose estimator overlaid on sample input images. Examples of imperfect pose estimations are shown in Figure 3.2.



(a) (b) (c)

(d) (e) (f)

Figure 3.1: Examples of point poses (a, b, c) and spread poses (d, e, f) with automatically estimated poses overlaid.

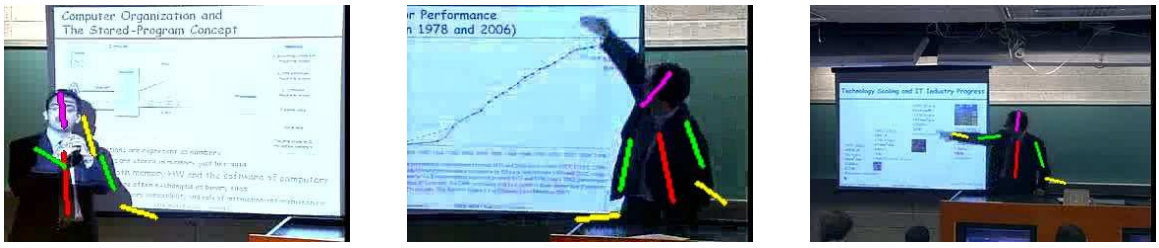


Figure 3.2: Examples of images where automatic pose estimation was imperfect.

The extracted pose is then represented as a set of 2D vectors pointing outwards from the neck (see Figure 3.3). To get the appropriate directions, we begin by assuming that the torso is more or less vertical, and then associate the related joints by finding the closest points. Given this vector representation of the five body parts, we can compute a descriptor comprised of joint angles.

3.1.2 Pose Descriptors

For each pose detected in an image, we can compute a 4-dimensional descriptor $D = [\alpha_0, \alpha_1, \alpha_2, \alpha_3]^T$. The values α_i represent the angles between the torso and left upper arm, torso and right upper arm, left upper arm and left lower arm, and right upper arm and right lower arm for $i = 0, \ldots, 3$ respectively (see Figure 3.3).

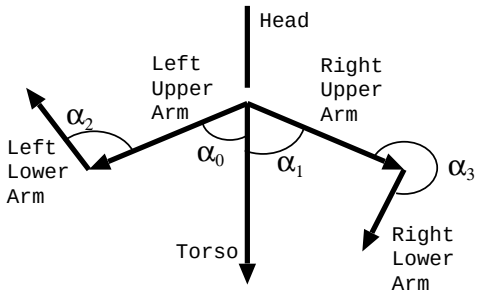


Figure 3.3: Model and descriptor values of estimated poses. Each body part is represented as a vector (depicted as arrows; the head is ignored).

Given vectors A, B representing a body part and the body part that is connected to it, we can compute the angles α_i as follows.

$$\alpha_i = a_i \cos^{-1}\left(b_i \frac{A \cdot B}{|A|\,|B|}\right)$$

where

$$a_i = \begin{cases} -1 & \text{if } i \in \{0, 2\} \text{ and } B \text{ is above } A \\ -1 & \text{if } i \in \{1, 3\} \text{ and } B \text{ is below } A \\ 1 & \text{otherwise} \end{cases} \tag{3.1}$$

$$b_i = \begin{cases} 1 & \text{if } i \in \{0, 1\} \\ -1 & \text{otherwise} \end{cases} \tag{3.2}$$

Note that for Equation 3.1, the determination of whether or not vector B is above or below A can be done in a variety of ways and is omitted here for conciseness. It may also be interesting to note that because this descriptor is attempting to model 3D poses using only 2D data (rotations about the shoulders are lost), the movement may not match the human joint model precisely. That is, α_2, α_3 may take on values that are physically impossible for human elbows.

3.1.3 Classifier

We use LibSVM [Chang and Lin, 2011] to train a classifier. We empirically determined that the RBF kernel performed best for both classes. Parameters were found using a simple grid search.

3.1.4 Evaluation

We evaluate our pose classification system by attempting to classify point and spread poses on manually labeled ground-truth. Approximately 20 minutes of a computer science video lecture was manually annotated. The video was sampled at 2 frames per second (as was done in [Zhang *et al.*, 2010]) resulting in 2500 frames. The video features a single instructor standing in front of both a blackboard and slide giving a lecture on computer architecture. Occasionally, student(s) sitting near the front of the classroom can be seen in the foreground. The videos were provided by the Columbia Video Network: cameras were human operated by lightly trained students in ambient light with no post processing. The video is provided at the low resolution of 352×240—audio and video quality are poor. The lighting condition varied throughout the video as the video faded in and out, or as the instructor adjusted the lights (to shift focus to and from the slides). The video does not focus solely on the instructor, as it shifts focus to a view of the slides from time to time.

We trained a binary classifier for each of the point and spread classes. For the automatic pose estimation, we use a pre-trained model provided by [Ferrari *et al.*, 2008] which we empirically found to produce satisfactory results.

From the 2500 frames collected for ground-truth, we identified 141 positive and 970 negative samples for the point class, and 125 positive and 986 negative samples for the spread class. The remaining frames were discarded as they did not have a visible person.

For evaluation, we performed 2-fold cross validation. The results, as measured by F-score are shown in Table 3.1. During training, we use all possible positive samples available in the partition but limit the number of negative samples to 7 times the number of positive samples (this factor was selected heuristically). This subset is randomly chosen from the entire pool of negative samples. During testing, all samples available in the testing partition are used.

Class	Precision	Recall	F-Score
Point	0.72	0.47	0.54
Spread	0.35	0.45	0.39

Table 3.1: Performance results of point and spread classifiers.

		Predicted		
		Neither	Spread	Point
Actual	Neither	744	77	24
	Spread	76	54	11
	Point	30	38	57

Table 3.2: Confusion matrix comparing classes.

Table 3.2 shows a confusion matrix constructed from these results.

3.1.5 Observations

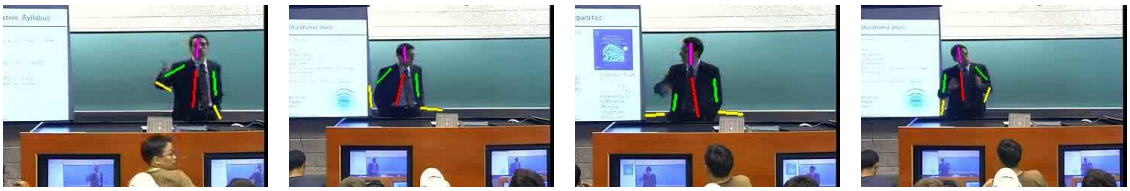
The performance of the classifier on identification of point poses is considerably superior than performance for identification of spread poses. This is not surprising as it is a rather distinctive pose. Misclassifications in this class tend to arise from poor pose estimations.

In this dataset, the point class is comprised entirely of pointing to the left. The reason for this is straighforward, as the instructor will tend to point to notes on the slides (he does not use the blackboard) which is always located to his right (when facing the camera). Presumably, pointing in the other direction could be trained in the same way.

Classification of spread poses was considerably more difficult. It encompasses a wider range of possible poses, some of which—e.g., standing with both arms outstretched—are difficult to distinguish in 2 dimensions. Two reasons were noticed which may account for the difficulty in training good models. One was the high number of incorrectly estimated poses, particularly of the arms, when the person is posing with arms outstretched (e.g., Figure 3.4d). Future work could explore the addition of hand detectors which may be used to aid pose estimation. A second reason was the granularity of the angles and the inherent

noisiness of spread poses. For instance, the person with both arms straight with hands resting on the table, and the person with slightly bent elbows with hands outstretched may both produce the same pose estimation.

Figure 3.4 shows examples of misclassifications for both point and spread classes.



(a) (b) (c) (d)

Figure 3.4: Examples of misclassifications. Figure (a) is a false positive point, (b) is a false positive spread, (c) is a false negative point, (d) is a false negative spread. As can be seen here, poor pose-estimation results are at least partly the cause of misclassifications.

3.2 Recognizing and Tracking Hands

In attempting to better discriminate poses, we shifted our focus from the upper body to focus on the hands specifically. Tracking hands in video is an important step in many applications of computer vision, particularly in gesture recognition for computer-human interfaces and gesture analysis. While the growing popularity of the Microsoft Kinect and other depth-sensing devices have greatly improved the efficacy of body part tracking and pose estimation, there is still a need for video-based methods for analysis applications. To this end, one frequently occurring challenge is the tracking of two hands simultaneously, particularly in natural communicative gestures when the hands may clasp together—a state which can confuse blob trackers.

We propose an algorithm for video processing which can track hands separately even if they clasp together. Intuitively, this is achieved by propagating bounding boxes from frames with many disjoint hand blobs to nearby frames with fewer distinct blobs. This is done iteratively until boxes have been propagated to all frames. Much of this work was initially presented in [Zhang and Kender, 2013].

The proposed method can be divided into three stages. We assume an input video containing a frontal shot of a single gesticulating person decomposed into individual frames. First, skin color detection and skin pixel clustering produce candidate hand blobs. As the primary contribution is not simply hand detection, we developed this method to accommodate mutual hand occlusion on input videos with clear hand blobs. Second, tracking is performed in a non-temporal manner by pushing frames (with associated blobs) into a priority queue which prioritizes frames with the most disjoint, separate blobs. Tracking is performed in priority order and blobs are associated across frames using a disjoint set data structure. Finally, a post-processing step identifies blobs which are hands. Note that this method cannot be applied in real-time except with a fixed-time delay. The overall method is illustrated in Figure 3.5.

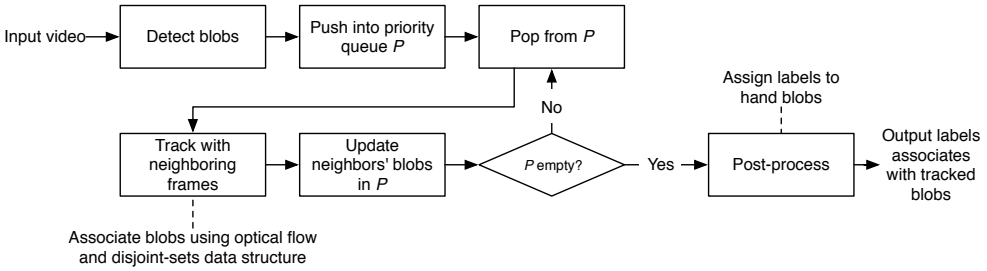


Figure 3.5: Overview of the proposed tracking algorithm.

3.2.1 Detecting Hand Blobs

This stage may be replaced with any method that produces candidate hand blobs such as [Mittal *et al.*, 2011]. For our purposes, we use the following method.

Given an input video V decomposed into frames $F_i, 0 \leq i < n$, we extract skin blobs from each frame as follows.

First, we train a simple skin color model using a subset of all frames (in our work, we trained using 10% of all frames in each input video):

1. Apply face detection [Viola and Jones, 2002].

2. Apply a general skin color model such as [Gomez and Morales, 2002] to identify skin pixels, and take a convex hull of these pixels in the image. Use the normalized red

and blue components of the encompassed pixels as samples.

3. Fit the sample pixels to a normal distribution.

To determine if a pixel is skin-colored or not, compute its probability based on the distribution estimated above, as shown in Figure 3.6b for the original image shown in Figure 3.6a. Automatically determine a threshold based on the sample pixels from the face and threshold the image as shown in Figure 3.6c (the lower 10th percentile of skin likelihoods worked well in our case).

The reason for training a new video-specific skin color model is to account for weaknesses in general color models such as [Gomez and Morales, 2002], which is less precise for darker skin tones. In [Gomez and Morales, 2002], a pixel with color (r, g, b) is a skin pixel if it satisfies all of the following conditions:

$$r/g > 1.185 \tag{3.3}$$

$$r \times b/(r+g+b)^2 > 0.107 \tag{3.4}$$

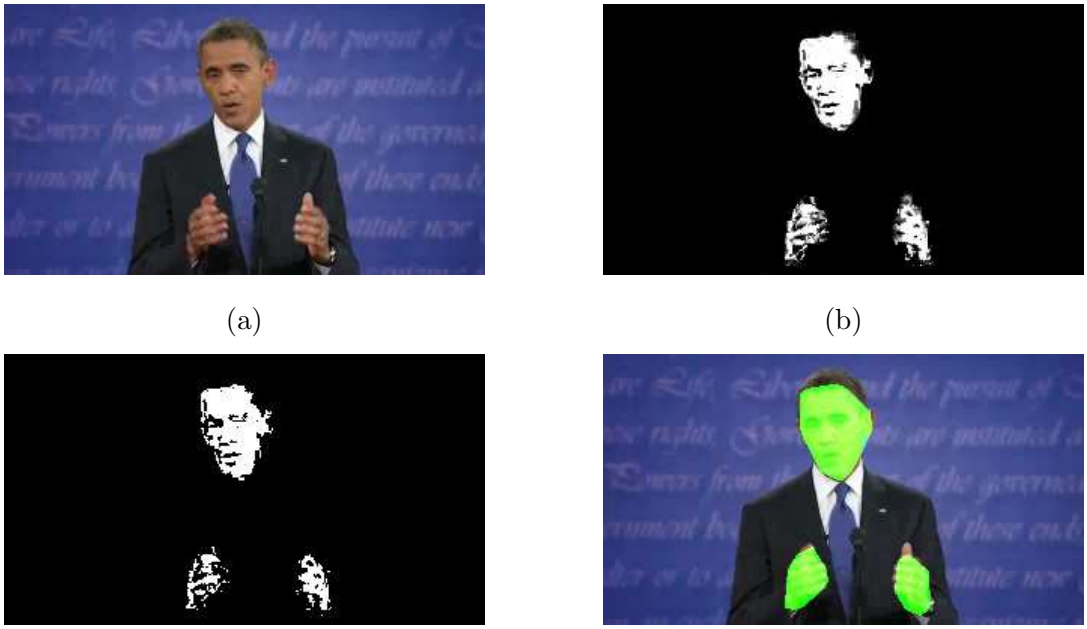
$$r \times g/(r+g+b)^2 > 0.112 \tag{3.5}$$

Next, we cluster the skin pixels into blobs according to spatial image distance, as shown in Figure 3.6d. Note that in this step, two clasped hands would usually be clustered together into a single blob. As a simplification, we represent each blob using the convex hull of its constituent skin pixels.

Each video can now be treated as a collection of blobs $b_{i,j}, 0 \leq j < m_i$ for each frame F_i.

3.2.2 Tracking

In this stage, we seek to track blobs in such a way that single blobs enclosing two hands (e.g., two hands clasped together) are automatically split into two blobs. We also seek to assign labels to blobs in such a way that blobs tracked across time would have the same label. To achieve this, we propose the following greedy algorithm which gives priority to frames with well separated blobs.



(a) (b)

(c) (d)

Figure 3.6: Blob detection stage: (a) is the original image, (b) shows the skin color probabilities, (c) shows the skin color probabilities thresholded and (d) is the result of the thresholded skin pixels clustered into blobs.

1. Given blobs $B_i = \{b_{i,j}\}$ and corresponding labels $L_i = \{l_{i,j}\}$ for each frame i (initialized to a unique ID for each blob), we push all B_i into a priority queue P, which prioritizes first by maximizing $|B_i|$ (since more blobs means greater likelihood of having 2 separate hands) then by maximizing the total pairwise spatial distances of blob centroids.

2. Pop the top frame from the priority queue: $F_i = \text{top}(P)$, where i is the position of the frame in the video.

3. Process frames F_i with F_{i-1} and F_i with F_{i+1} (i.e., propagate blobs from F_i). Without loss of generality, we describe the algorithm for processing frame F_i with F_{i+1}:

   (a) If frame F_{i+1} is already flagged (i.e., processed), then continue to the next frame.

   (b) Otherwise, for each blob $b_{i,j} \in B_i$, compute the optical flow of its pixels from

frames F_i to F_{i+1} using a method such as [Farnebäck, 2003], which gives preference to matching pixels within blobs of B_{i+1}.

(c) Retain only these pixels, with their convex hull forming a new blob $b'_{i,j}$. Insert this new blob into the updated blob set B'_{i+1}. Set $B'_i = B_i$. Also, associate the label $l_{i,j}$ with label $l'_{i+1,j}$ (initialized to be unique) in a disjoint-set data structure via union-find, that is, find($l_{i,j}$) = find($l'_{i,j}$), and set $l'_{i,j} = l_{i,j}$. Add labels $l'_{i,j}, l'_{i+1,j}$ to L'_i, L'_{i+1}, respectively.

(d) Repeat from step (a) until all blobs in frame F_i are processed. Then flag F_i as processed.

4. Repeat step 3 until P is empty.

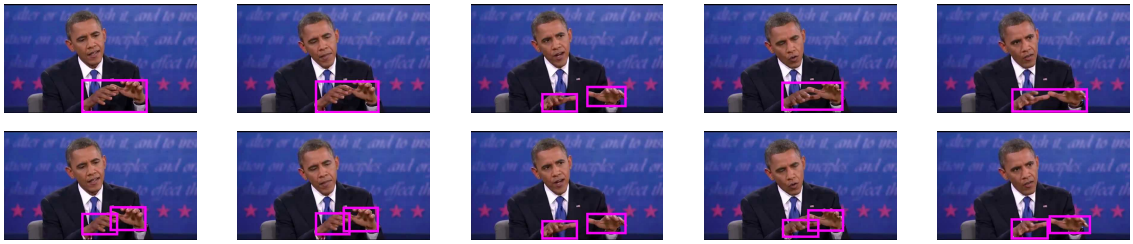
   By processing frames in this manner the bounding boxes of larger blobs may be overwritten with more smaller blobs. It will never be the case that two smaller blobs need to join into a single, larger blob. Also, scenes where the hands are initially clasped but come apart are therefore processed correctly.

5. Output updated blobs B'_i and labels L'_i for each frame.

Note that by processing frames in this order—unlike existing approaches which process frames sequentially—it becomes possible to infer when one blob contains two hands based on their neighboring frames. Of course, this approach only works if the hands are separate blobs at one point in the video sequence. That is, it will not be effective for a video sequence where the person's hands are constantly clasped together. As previously mentioned, one disadvantage of processing frames out of order is that tracking cannot be applied in real-time except with a fixed-time delay.

Examples of a sequence of frames with hand blobs propagated through tracking compared against hand blobs detected independently are shown in Figure 3.7.

Some statistics were collected from processing 47 video clips totaling 28 minutes and sampled at 15 Hz. It was observed that forward propagation and backward propagation tend to occur approximately equally (49.6% forward vs. 50.4% backward), and that there are few long runs, where a run length is defined as the total number of frames affected by



(a) (b) (c) (d) (e)

Figure 3.7: The top row shows the original blobs produced by clustering skin-colored pixels. Note every column except for (c) the two touching hands are clustered together into one blob. The bottom row shows the blobs after applying the proposed tracking algorithm.

a seed frame, propagating both forward and backward. That is, while the longest detected run is 140 frames, the average run length is only 4 frames. A visualization of the propagation of bounding boxes across frames and processing time for a single video clip totaling 71.5 seconds is shown in Figure 3.8. It can be seen that seed frames tend to be random.

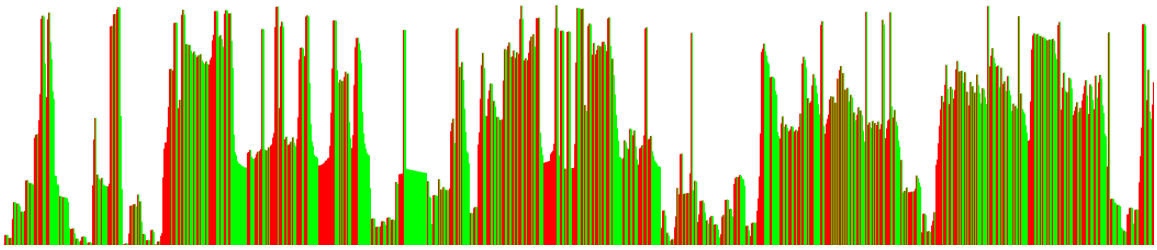


Figure 3.8: Propagation of blob bounding boxes across frames (horizontal axis) and processing time (vertical axis, starting at the top). Green and red indicate forward and backward propagation, respectively. Note the near equality of red and green.

3.2.3 Post-Processing

3.2.3.1 Associate Overlapping Blobs

In some cases, due to the imprecisions in computing optical flow and the proposed method, single blobs may be split into multiple overlapping small blobs within a frame. Also, blobs in roughly the same positions in neighboring frames may also not get associated with the

same label. To reduce these errors, we can simply apply a post-processing stage whereby labels for blobs with significant overlap (i.e., over 75% of area) with other blobs within the same frame or with neighboring frames are associated.

3.2.3.2 Labeling Blobs as Hands

To finally associate blobs with hands as in Figure 3.9, we need to apply a hand detection algorithm to "seed" the labels at points throughout the video. These seed bounding boxes may be determined automatically (e.g., using [Mittal *et al.*, 2011]) or possibly even provided manually. One naive heuristic for the single-speaker case is simply to use the detected blobs (minus the face) and label hands as left or right based on their relative position to the face. In our experiments described in Section 3.2.5, we use this naive heuristic as a baseline. To take advantage of tracking, we can improve on results by assigning labels based on the most frequently occurring label for a tracked blob across time.

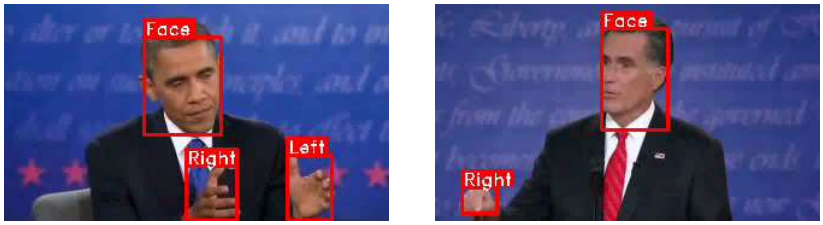


Figure 3.9: Results of blob tracking algorithm which recognizes the face and left and right hands.

3.2.4 Ground-Truth Data

For our experiments, we used recorded video of the first and third 2012 US Presidential debates, which are widely available online. The videos are very high quality with a resolution of 640 × 360 pixels, a stationary camera, and tend to focus on the upper bodies of the speakers, who gesticulate frequently while they speak. Their hands appear in and out of view.

To select these shots, we automatically segmented the videos using HSV color histogram comparison with 3D histograms with 16 bins for each dimension and compared using the L2

distance and a threshold of 0.1 similar to [Smeaton *et al.*, 2010]. We manually selected 322 frontal shots of a single speaker, totaling 142.8 minutes. We applied our baseline hand blob detection technique to extract 4163 images of hand blobs at 1 Hz, which were presented to Amazon Mechanical Turk raters. From this, 155 images of hand blobs were identified to contain two hands. Using this set as seeds, we took 4 additional surrounding frames at 2 Hz (2 frames before and after) to generate a new dataset of 698 images (some frames overlapped) spanning 47 video segments. Our proposed method processed each of these segments at 15 Hz. This final dataset was then manually annotated using the Mechanical turk box labeling tool of [Xiao *et al.*, 2010] for each hand present, producing 1301 instances of hands. Each hand was also labeled as a left or right hand (at most two hands are present in each shot).

3.2.5 Evaluation

We evaluate two different aspects of our proposed method. First, we evaluate its ability to identify when a single hand-blob actually contains two hands. This step can be done without the post-processing step. For this, the baseline method is simply the hand blob detection step without tracking (i.e., only Section 3.2.1).

Second, we evaluate its ability to track hands across time by augmenting a naive heuristic and assigning left and right hand labels to detected hands, as described in Section 3.2.3. For this, the baseline method is simply the naive heuristic applied to each frame independently without using the tracking information.

We detected bounding boxes of hand blobs using our baseline and proposed methods. Normally, a bounding box around the head is detected as well but they were removed automatically.

To evaluate our bounding boxes, we adopted the metric of [Mittal *et al.*, 2011] and [Everingham *et al.*, 2010]. That is, an overlap between a candidate bounding box B_c and a ground truth bounding box B_g is calculated according to: $O(B_c, B_g) = \frac{\text{area}(B_c \cap B_g)}{\text{area}(B_c \cup B_g)}$, and the boxes are considered equivalent if $O(B_c, B_g) > 0.5$. Using this measure, we computed precision P and recall R for the baseline and the proposed methods, where $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$, where TP, TN, FP, FN refer to true positive, true negative, false positive

Method	Precision	Recall	F-Score
Baseline	0.65	0.53	0.59
Propagation Tracking Method	0.65	0.66	0.65

Table 3.3: Performance of baseline and presented propagation tracking (without post-processing) methods on evaluation dataset of 698 images with 1301 instances of hands, regardless of labels.

Method	Precision	Recall	F-Score
Baseline	0.53	0.43	0.48
Propagation Tracking Method	0.63	0.56	0.59

Table 3.4: Performance of baseline and presented propagation tracking (with post-processing) methods on evaluation dataset using left / right hand labels.

and false negative, respectively.

The results of the hand detection evaluation are shown in Table 3.3. It can be seen that while precision of the proposed method is no better than baseline, recall is significantly higher as it successfully detects when a single blob actually contains two hands.

The results of the tracking evaluation are shown in Table 3.4. Because the left / right hand labels are shared across frames through tracking, it is more accurate.

An examination of the results indicates some possible reasons for failure including: the subjective manual annotations in cases where the hands are clasped (Figure 3.10a). Note that while the proposed method successfully identified that two hands are present, the inaccuracy of the exact bounding box contributed to it being marked as a mismatch. Another case is when only the thumb and forefingers of a hand are visible, resulting in two separate blobs from the proposed method but usually resulting in only one bounding box in the ground truth (Figure 3.10b). Finally, in some cases, a single large blob may be incorrectly clustered into several small blobs, which register as false positives (Figure 3.10c).

(a) (b) (c)

Figure 3.10: Green and red boxes indicate insufficient overlap / mismatch for ground-truth and detected boxes, respectively. Magenta indicates successful match for both ground-truth and detected boxes.

3.3 Gesture Attributes

In Section 3.1 we described a method for recognizing semantically relevant poses. In Section 3.2 we described a robust method for tracking hands in video. In this section, we will apply some of these results toward extracting certain characteristics of gestures which we will correlate with audience engagement in Chapters 5 and 6.

3.3.1 Arm Angular Variance

Given an input video V, let I_f, I_{f+1} represent two consecutive image frames of V. We automatically estimate the pose of the speaker according to a 6-part upper body model by applying the method described in [Ferrari *et al.*, 2008] to image I_f. This outputs for each body part p and pixel i an L1-normalized likelihood $b_{p,i}$, that pixel i belongs to part p where:

$$p \in \langle \text{Head}, \text{Torso}, \text{UL}, \text{UR}, \text{LL}, \text{LR} \rangle$$

UL, UR, LL, LR represent upper and lower left and right arms, respectively. Next, dense optical flow between I_f, I_{f+1} is found using [Farnebäck, 2003], producing an image flow vector $\mathbf{F}_i$ for each pixel $i \in I_f$. Then, a weighted mean vector $\bar{\mathbf{F}}$ of the flows can be computed for each part p within frame I_f so $\bar{\mathbf{F}}_{p,f} = \sum_i b_{p,i}\mathbf{F}_i$. Let $\bar{\mathbf{F}}_p = \{\bar{\mathbf{F}}_{p,f} : I_f \in V\}$, i.e., the set of all flow vectors in the video.

The mean image flow of the torso is subtracted from those of the arm parts (i.e., UL, UR, LL, LR) to account for any body locomotion, and the resulting motion-compensated vectors

are collected into their corresponding four arms sets across all frames in the sequence V. We suspect that information in the lower arms (or possibly even the lower dominant arm) may be sufficient. However, currently, too much noise is introduced by the pose estimation method for lower arms.

Finally, for a each arm set (i.e., the motion vector of the arm across frames in V), we compute their circular variance to produce a 4D feature vector $X(V)$, as follows:

$$X(V) = [\text{Var}(\bar{\mathbf{F}}_p) : p \in \langle \text{UL}, \text{UR}, \text{LL}, \text{LR} \rangle] \tag{3.6}$$

Where Var computes the circular variance which is defined in Equation 3.7 [Fisher, 1996]. Intuitively, the circular variance is a statistic for computing variance for angles, so unlike the usual definition of sample variance, it allows for values which "wrap around".

$$\text{Var}(\bar{\mathbf{F}}_p) = 1 - \left| \frac{1}{|V|} \sum_{f=1}^{|V|} \bar{\mathbf{F}}_{p,f} \right| \tag{3.7}$$

Feature generation is summarized in Figure 3.11 and examples of estimated poses and computed flows are shown in Figure 3.12.

Intuitively, this method for feature generation will represent the amount of "arm-waving" in a given segment, revealed by sudden and drastic changes in movement orientation. While the use of movement magnitude information could also be useful, the imperfect performance of the pose estimator introduces too much noise. Our justification for using circular variance is similar. While the use of circular variance discards potentially useful temporal information, it also mitigates the effects of false negatives (i.e., frames mistakenly believed to not contain a speaker) and poorly estimated poses as these vectors are averaged with the other flow vectors for the part in the video.

3.3.2 Velocity and Direction Change

Here, we describe two strongly related gesture attributes: velocity and direction change. While the measure of "arm waving" described in Section 3.3.1 seeks to capture how "frantically" an arm is being waved as a single scalar, by using both velocity and direction we hope to distinguish between more nuanced versions of the movements.

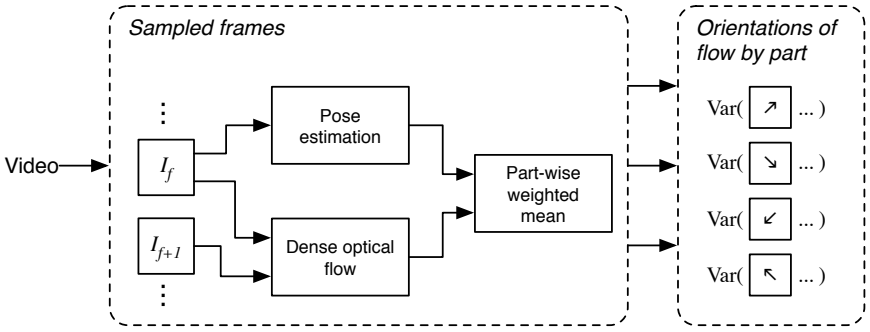


Figure 3.11: Overview of feature generation. For each video sequence, we sample frames uniformly. For each pair of consecutive frames, the pose is estimated and dense optical flow computed and averaged for each part. The final feature vector is the circular variance of the orientations of average optical flow across frames separated by part.

We begin by applying the hand detection and tracking algorithm described in Section 3.2 to distinguish and track left and right hands, as well as the face, as shown in Figure 3.7. From this we can derive our attributes, anchored at each frame, computed solely from the positions of the bounding boxes of detected hands and faces.

The attributes we initially explored include: distance from hands to each other (in cases where there are two hands), distance from each hand to the face, velocity of motion for each hand, and direction change for each hand. Preliminary analysis using PCA on the set of all attributes suggested that velocity and direction change were responsible for a significant part of the variation in the features, therefore, we focus on these two attributes here.

3.3.2.1 Velocity

For each frame f, we can compute the velocity of a hand's motion simply by summing up the distances that a point on the hand (e.g., the upper left corner of the bounding box) moves within a neighborhood of the anchor frame, and dividing by the number of frames elapsed. In our case, the velocity can also be normalized to the range of $[0, 1]$ by dividing by the maximum velocity across all videos. This is acceptable since the speakers are centered in the shots in roughly the same way. We denote the normalized velocity at frame f by $\mathbf{v}_f$.

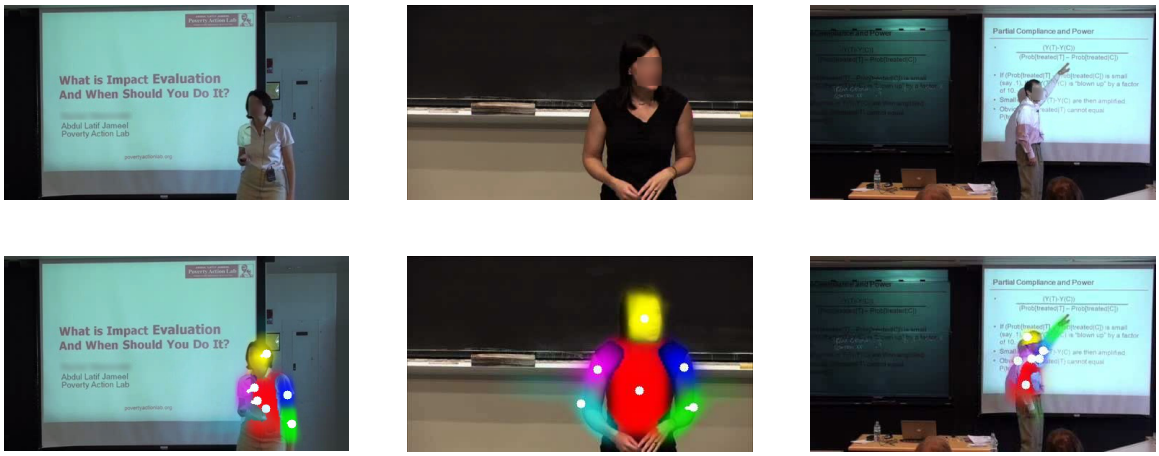


Figure 3.12: Examples of pose estimations (i.e., estimating the position and orientations of the head, torso, upper and lower left and right arms, which are shaded) and part-based optical flows (visualized by the white arrows from the centroids of each part). The originals are shown in the top row while the parts and flows are visualized in the bottom row. The rightmost column shows an example of a poor pose estimation.

3.3.2.2 Direction Change

For each frame f, we can also compute the degree of change of direction in the gestures. First, for each hand, we concatenate the (x, y) positions of the centroids of each hand bounding box within the neighborhood of each frame g as a row matrix M_g. We subtract the mean of these centroids from each row of the matrix and apply PCA to recover the principal components and the amount of variance associated with each. If a principal component is responsible for a majority of the variance (in our work, at least 75%) then that component is retained as a measure of the direction of the hand movement within the window of frames. We denote this vector at frame g by $\vec{\mathbf{p}}_g$

Next, to compute the degree of direction change in the neighborhood of frame f, we find the maximum angle between principal components in the set of frames. Since the angle is between $[0, \pi]$, it can also be normalized to the range $[0, 1]$. We can compute the normalized angle by:

$$\mathbf{d}_f = \max_{i,j} \left\{ \arccos \left(\frac{\vec{\mathbf{p}}_i \cdot \vec{\mathbf{p}}_j}{\|\vec{\mathbf{p}}_i\| \cdot \|\vec{\mathbf{p}}_j\|} \right) \right\} / \pi \tag{3.8}$$

for all i, j in the neighborhood of f.

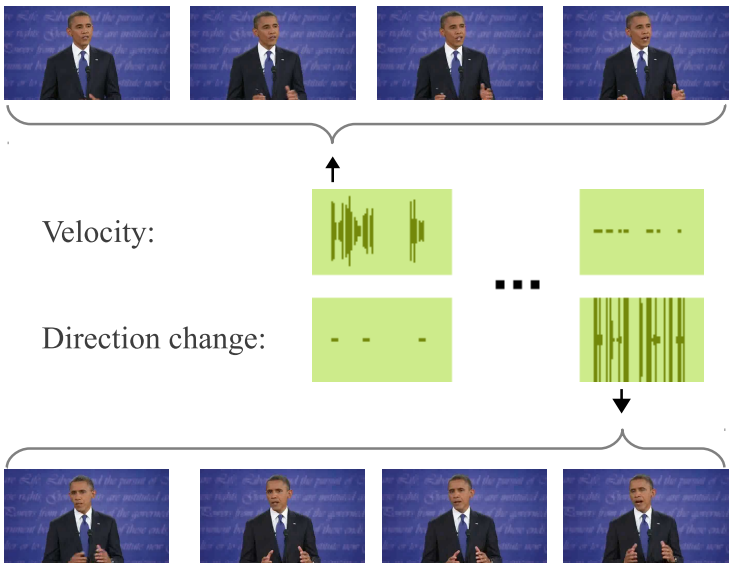


Figure 3.13: Examples of gestures with high velocity and a low amount of direction change (e.g., a "swipe" motion) and low velocity and high amount of direction change (e.g., a "jittery" beat motion). The graphs of the velocity and direction change are center-aligned.

3.3.2.3 Combining Attributes Across Hands

The attributes are computed for each hand, but can be combined simply by taking the maximum (across hands) for each frame.

Based on both the velocity and direction change attributes, it is possible to identify certain types of motions, such as long "swipe" motions or short and quick "jittery" motions, as shown in Figure 3.13.

3.3.3 Extremal Poses

One frequent observation across the related literature and our own research is that gestures are highly idiosyncratic—that is, they vary greatly from person to person. Even so, each individual often forms their own habits (this is often heavily influenced by their cultural background, but that is beyond the scope of this thesis). This means that given sufficient information about a person's gestural habits during speech across time, we can therefore identify the moments when they use unusual or extremal poses. We hypothesize that these

moments when their body language deviates from their regular patterns may correlate to significant events in the corresponding speech.

In the following, we describe our approach for building a measure to identify these moments for two speakers in one of our domains of interest.

3.3.3.1 Ground-Truth Data

In our domains of interest, it is possible to get a significant amount of information regarding a speaker's gestural habits. We focus here on the Presidential debates and the gestural habits of the two U.S. presidential candidates: (President Barack) Obama and (former Massachusetts Governor Mitt) Romney.

We use video of the first and third (i.e., final) debates (not counting the Vice Presidential debates). In their original form, the first debate video is 1:31:50 hours long while the third is 1:36:20 long. Both have a resolution of 640×360 pixels and are recorded at 24.58 fps using stationary cameras, and tend to focus on the upper bodies of the speakers, who gesticulate frequently while they speak. Their hands appear in and out of view.. These are available through a number of sources, including the Internet Archive[1].

The full footage includes a variety of shots, including shots of the moderator and wide-angle shots of the candidates together. For simplicity, we only wish to use frontal shots of the each candidate alone. To automatically extract the desired shots we used a combination of automatic and manual techniques (fully automating this task is non-trivial and beyond the scope of this thesis). First, we automatically segmented the videos using HSV color histogram comparison with 3D histograms with 16 bins for each dimension and compared using the L2 distance and a threshold of 0.1 similar to [Smeaton *et al.*, 2010], resulting in 322 shots. Next, we manually discard the shots which do not contain a frontal view of a single candidate, leaving us with 291 shots of interest.

Finally, we apply the method for hand tracking and labeling described in Section 3.2 to the each of the desired shots at 12 Hz to obtain bounding boxes and associated labels (i.e., left or right hand). We use the centroid of the bounding boxes for each hand in each frame. A summary of all the data used and a breakdown of the number of frames for each

[1]The Internet Archive, *http://www.archive.org*

	First debate	Third debate
Desired shots featuring Obama		
Number of shots	59	81
Total number of frames	29046	22430
Desired shots featuring Romney		
Number of shots	72	79
Total number of frames	23577	23121
No. frames where Obama's hand(s) are detected		
Only right	1949	907
Only left	6532	12343
Both	11023	4377
No. frames where Romney's hand(s) are detected		
Only right	2662	3517
Only left	4774	786
Both	6228	1023

Table 3.5: Summary of video data used to learn extremal pose models for Obama and Romney. It is interesting to note that Obama's left-handedness is very apparent.

speaker-video-hand grouping is shown in Table 3.5.

3.3.3.2 Model

We propose to model the likelihood of a pose as defined by the position of centroids of the bounding boxes of the visible hand(s) in a frame. To do this, we first stratify the data according to speaker and video. Throughout each of our videos, the speakers maintain a roughly constant pose, simplifying our task. However, the poses cannot be combined across videos as their poses vary, that is, in the first debate they stand in front of a podium while in the second, they sit a table.

The model will also depend on how many hands are present—whether it is a left-handed gesture, a right-handed gesture, or a gesture using both hands. The hands in our videos

may be partially occluded or simply out of view, but since we are focusing on how gestures affect the audience who sees it, this is nevertheless valid.

We examine the position of the hands for each speaker, across each video, as seen in the heatmaps in Figures 3.14 and 3.15. We model the position of each hand as a 2D gaussian distribution. Estimating the gaussians in the cases where only one hand is visible is straightforward. In the case where both hands are visible, we model the positions as a mixture of 2 gaussians, estimated using the expectation-maximization (EM) algorithm.

Finally, as a measure of the likelihood of a given pose, we use a function of the Mahalanobis distance. Since the distance is bounded in our case (i.e., the size of the image is finite), this also provides us a way to normalize the measure to the range $[0, 1]$ consistent with the velocity and direction change attributes described in Section 3.3.2. In the case where there are 2 hands, we simply take the maximum Mahalanobis distance of the two (i.e., the most unlikely position).

Formally, for each frame of video, we can compute the likelihood that a pose is unusual as follows. Let $H = \{\text{left}, \text{right}, \text{both}\}$ be the set of hand(s) which may be detected in the frame. Suppose we have trained gaussian mixture models G_i for each $i \in H$ representing the likelihood of the hand position. Let $X = \{(x_i, y_i), \ldots | i \in H\}$ be the set of hand positions in the frame and $h \in H$ be the hand(s) detected in the frame, then the likelihood measure M is simply:

$$M(X|h) = \max_{x \in X} \left\{ \frac{D_{G_h}(x)}{D_{\max}} \right\} \tag{3.9}$$

where D is the Mahalanobis distance and $D_{\max}$ is the maximum Mahalanobis distance given the size of the image, which we use as a normalization factor.

We have now presented a number of gesture attributes and techniques for extracting them. In the following chapters, we will examine how they correlate with different measures of audience engagement and identify those that drive correlation.

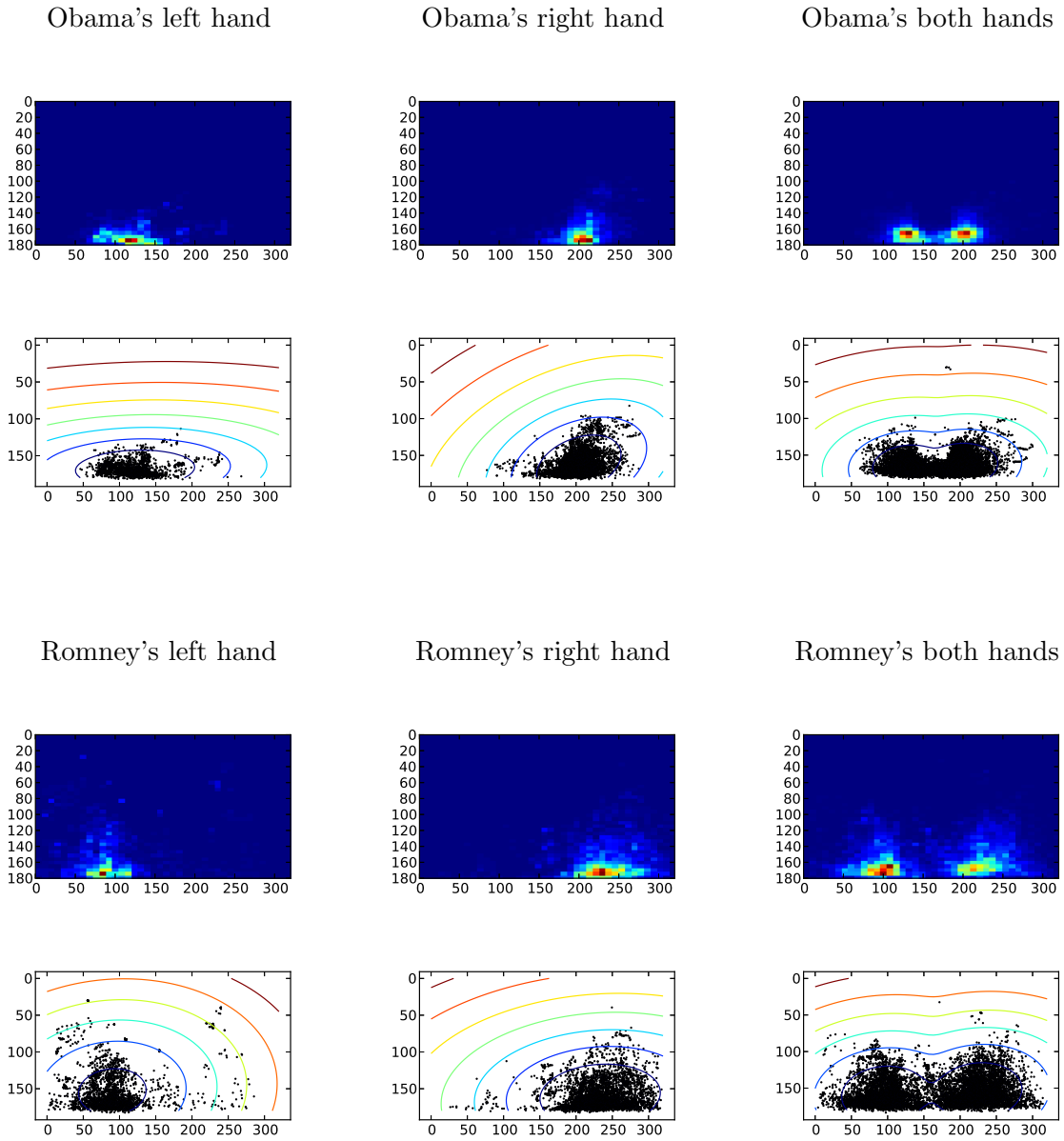


Figure 3.14: Hand positions and estimated mixtures of Gaussians for the first debate. Contours represent Mahalanobis distances.

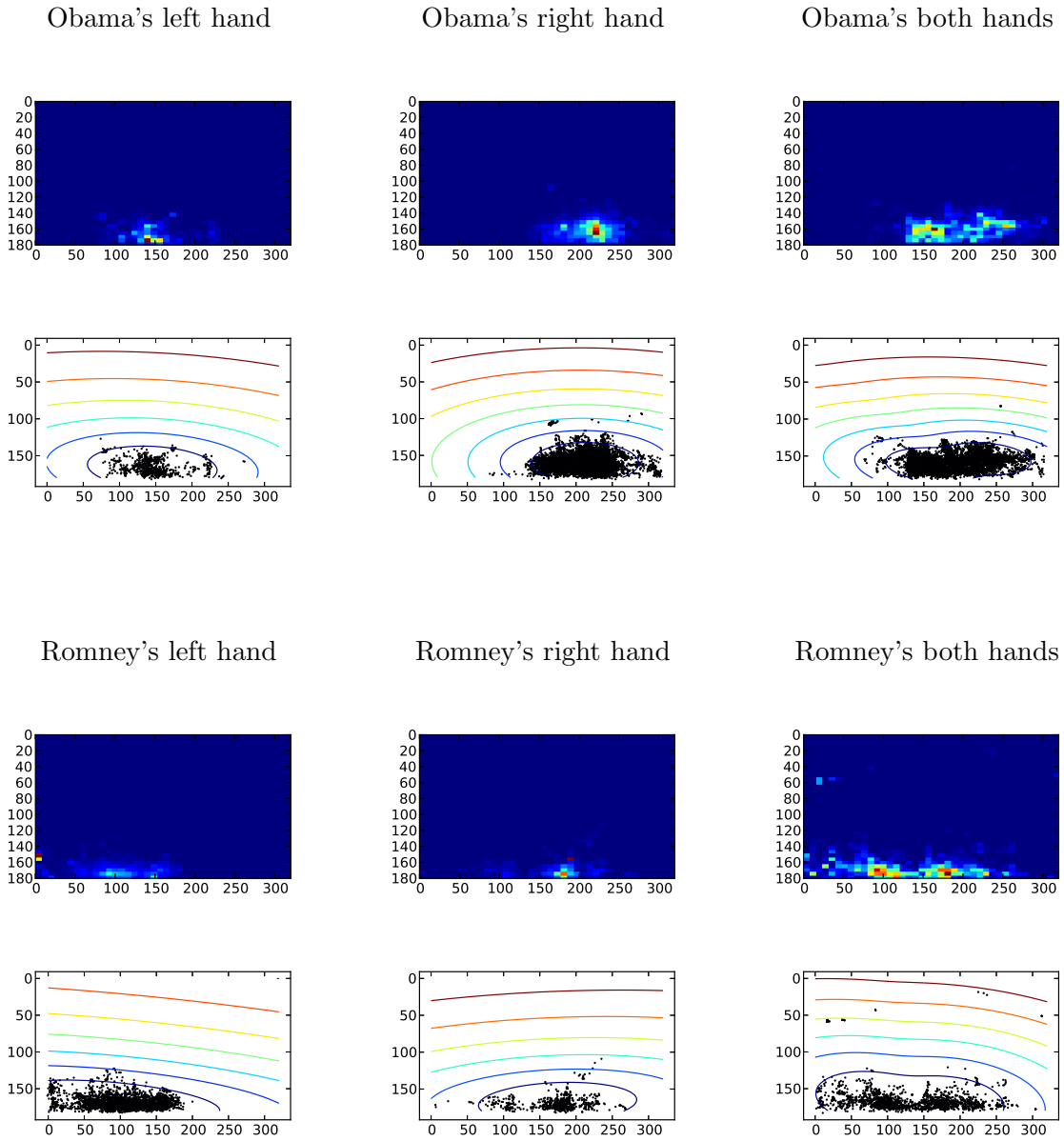


Figure 3.15: Hand positions and estimated mixtures of Gaussians for the third debate. Contours represent Mahalanobis distances.

Chapter 4

Annotation and Taxonomy of Gestures in Videos

As part of our goal to identify and categorize salient gestures in lecture videos, we examine over 2 hours of recorded lectures, totaling over 14,000 frames. We collected the gestures by developing a novel annotation tool, manually labeling gestures in videos, and analyzing the results. We also propose a new taxonomy which encapsulates the gestures which we argue to correlate to pedagogic significance and worthy of further investigation. This work was intially presented in [Zhang *et al.*, 2010].

4.1 Gesture Annotation Tool

We introduce a novel tool designed for annotation of gestures in video. In this section, we focus on a discussion of the tool's usage and user interface design.

The tool takes as input a sequence of still images, an optional audio file, as well as an index file stored in a directory. The audio and still images are usually extracted from a video. This was done mainly to increase the ease of integration between the annotation tool and many implementations of computer vision algorithms, which often process still images or sequences of still images rather than video files directly. This has the added benefit that the tool becomes less concerned with video formats. Producing the requisite files from a video is simplified through the use of a script (available as part of the tool). Video frames

are usually stored at a rate of 30 frames per second but we find they may be extracted at a rate as low as 2 per second without loss of significant gestural information, for memory efficiency.

Once the appropriate files are available, the user can create a new project in the annotator tool, specify generic metadata (e.g., project author, comments) as well as the index to the video, and begin the process of annotation. The annotations and associated metadata can be exported to XML.

Gestures in the tool are represented as a collection of keyframes within a subsequence of the images where the poses are specified in detail. As we generally follow the three-phase (or multi-phase) model of gestures as described in [Kendon, 1980; Wilson *et al.*, 1997], the use of keyframes allows us to roughly identify the phases in addition to the distinguishing poses of the gesture and their temporal relationships. The representation was inspired by existing work, but modified to acknowledge their restriction on upper body gestures, and to gestures that preferentially occur in one-sided communications (teacher monologues).

4.1.1 User Interface

The main user interface (Figure 4.1) is divided into two sections: the video player, and the gesture editor. The video player gives users the ability to watch the sequence of images in rapid succession as a video, and optionally provides audio if an audio stream is available and the operating system is capable of supporting the codec. The user is capable of jumping to specific frames, speed up and slow down playback, and other common features.

The gesture editor itself is divided into two tabs: video frames and a list of gestures. The video frames tab is visible in Figure 4.1 and shows a sequence of the video frames in a timeline format. This timeline feature was developed after we observed that it facilitated the identification of the various phases of a gesture, as well as the exact frames those phases occur as the user can see "across" time. We also observed that at least two gestures may sometimes overlap. Specifically, out of 372 annotated gestures in our collected data, 26 of them were overlapping with another gesture (for example, the lecturer simultaneously shrugged while making hand/arm gestures). Therefore, the user is capable of specifying sequences of frames for different gestures, which are shown as different gestural tracks. The

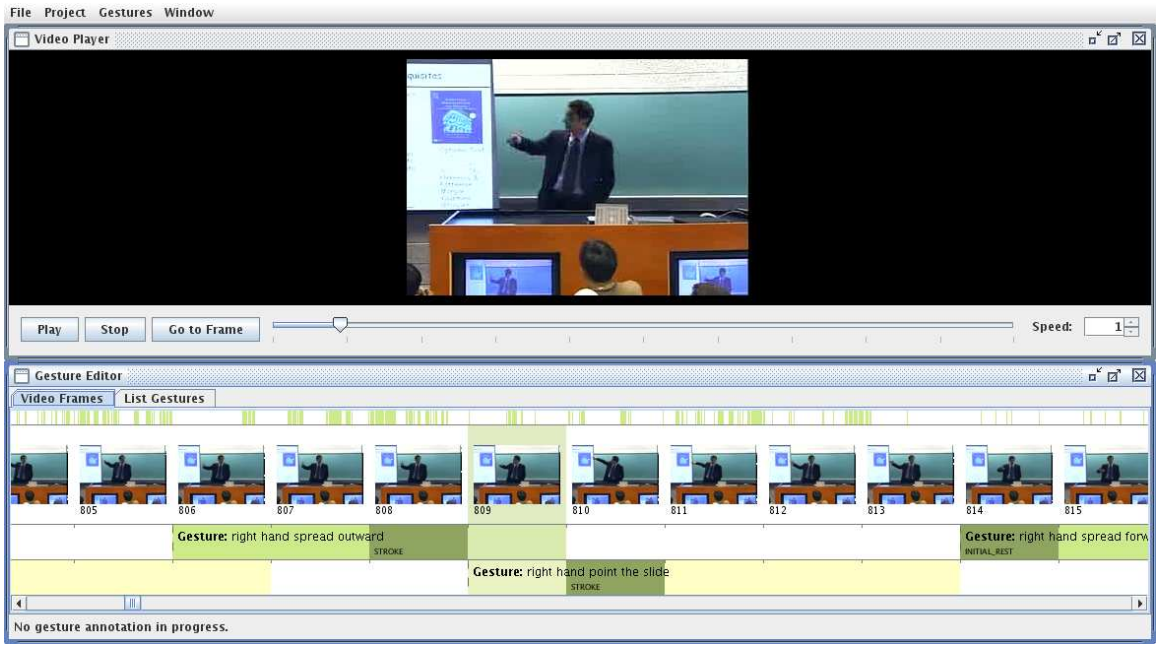


Figure 4.1: The main user interface of the gesture annotator tool.

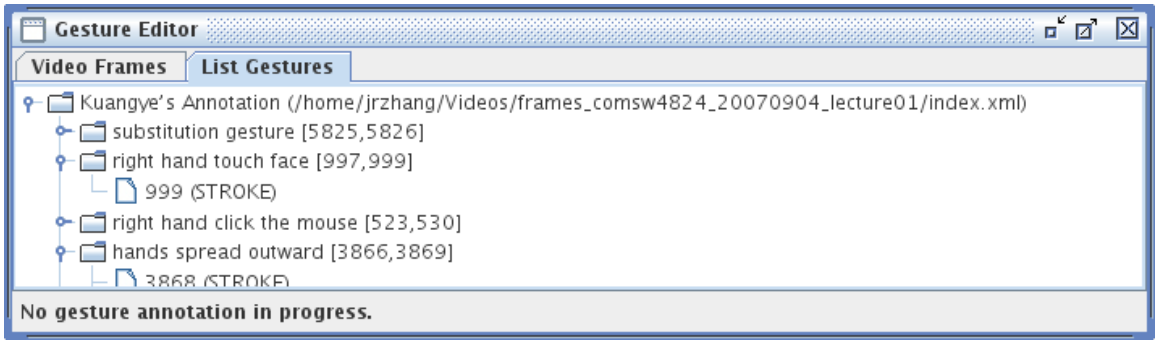


Figure 4.2: The tree-view tab of the gesture editor internal window, which lists the existing annotations in a project in a hierarchical format.

list of gestures tab is shown in Figure 4.2 and contains a tree UI structure which displays hierarchical data and provides the user with a textual overview of the current annotations in the project.

To mark a sequence of frames as belonging to a gesture, the user can select the sequence and use the pop-up menu that appears. The user is then asked to provide a description of the gesture. This highlights the sequence and makes other options available, particularly the ability to mark individual frames (within the newly marked sequence) as a keyframe, which are highlighted as a darker color in the gesture sequence (see the bottom of Figure 4.1).

An alternate way to mark the start and end timestamps of a gesture is to play the video and mark the endpoints with hotkeys.

A third interface is shown when the user identifies a keyframe and wishes to specify the pose of the instructor. This interface allows the user to choose the best way to describe the pose, according to their judgment. The user may choose to use the avatar poser (as seen in Figure 4.3), provide a textual description, or specify that there is no human visible in the frame. The user may also specify the phase of the keyframe (i.e., in deference to the three-phase gestural model) as well as provide an optional comment.

4.1.2 Annotating Poses By Avatar

Once a user has identified a keyframe and wishes to further illustrate the pose of the lecturer, the graphical poser can be used.

In our preliminary findings, we observed that most significant gestures in teaching can be represented using simple upper body, arm and head movements. We chose this as a starting point which is reflected in the granularity of our poser. The state of the poser can be represented in 14 bits, with all possible selections shown in Figure 4.3. Some examples of gesture and their approximate avatar representations are shown in Figure 4.4. A discussion on the appropriate level of granularity is given in Section 4.2.5.

The user interface is defined to balance the user's ability to describe the pose both accurately and quickly. The radio buttons in the graphical UI are positioned in a way as to correspond to the parts of the body and also to minimize the distance between one another,

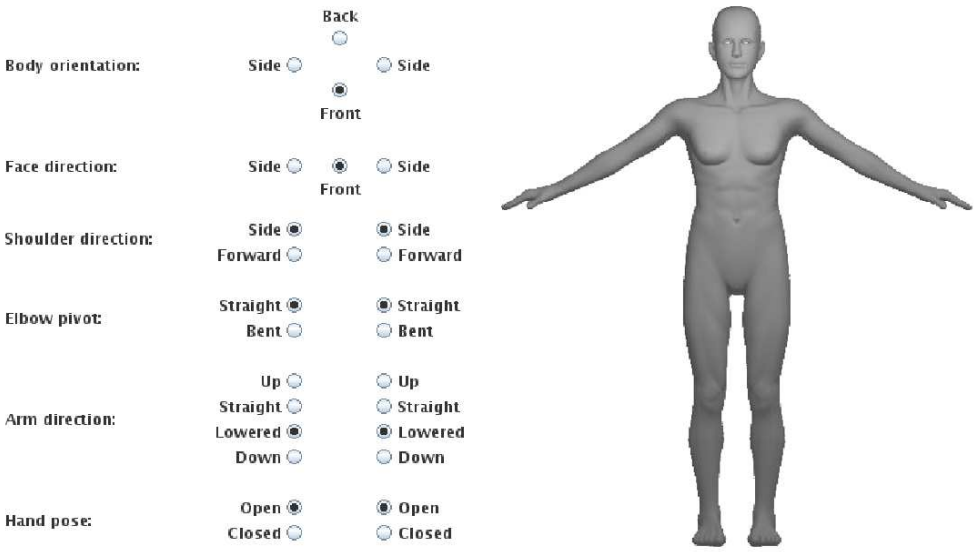


Figure 4.3: The avatar poser controls in the default configuration, along with the corresponding avatar preview image.

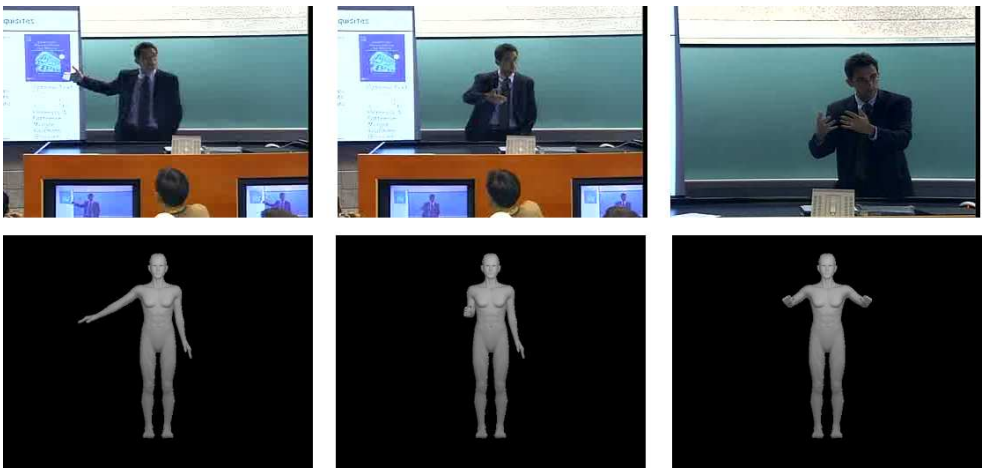


Figure 4.4: Examples of gestures and their avatar representations below.

so users may select them faster.

The avatar control radio buttons are placed beside a preview window, which changes to reflect the latest pose selected by the user. The avatar in the preview window will always face forward regardless of body orientation, as we noticed it was easy for annotators to mirror the lecturer's pose, even when they are turned around. We also considered other avatar representations, including the possibility of using two separated avatars to represent the lecturer from different perspectives; our present version seems sufficient.

4.2 Annotation and Analysis

Two 75-minute computer science video lectures have been manually annotated for gestures using the proposed tool. In following with Martell's observation of strong intra-annotator but weak inter-annotator consistency [Martell, 2002], both videos were annotated by the same person. Each video captures a different instructor from different cultural backgrounds, presenting topics from different areas of computer science (one lecture is on machine learning, the other is on computer architecture). During preprocessing, the video frames were extracted and collected as a sequence of still images at a rate of 2 frames per second. The videos were provided by the Columbia Video Network and have characteristics described in Section 1.2. The videos do not focus solely on the instructor but sometimes switch to a view of the slides presented for a period of time (for the computer architecture video and the machine learning video, 24% and 41% of the frames extracted were marked as belonging to a gesture, respectively).

Part of one of the videos was also annotated by a second person to explore inter-annotator consistency; see Section 4.2.5.

Finally, observations were collected from both annotators regarding the level of granularity for the avatar poser, the frame rates of the extracted video, and high-level patterns noticed in the gestures.

The first lecture video (video A by instructor A) presents an introduction to computer architecture, an outline of the course, and an overview of the material without elaborating on the theory.

The second video (video B by instructor B) provides an introduction to machine learning but goes directly into a detailed explanation of linear regression, presenting a lot of mathematics.

4.2.1 Proposed Taxonomy

During annotation, gestures were assigned a textual label according to the template "*body part, semantic class, orientation*." For example, a gesture where the instructor points with his right hand would be labeled as *right hand point right*, where *right hand* is the body part, *point* is the semantic class and *right* is the orientation (i.e., the direction in which he is pointing). We identified 126 unique labels falling into nine semantic classes. We defined a new semantic class whenever we noticed that the gesture was frequently repeated or that the gesture was semantically relevant to the lecture content.

The nine semantic classes were identified as follows. We note that some of them do not cleanly fall into the four or five classes commonly assumed in the literature. We introduce the class of "pedagogic" gestures to label those gestures whose purpose seems to be to structure the lecture or to encourage or remind the students. This category has not been documented in the prior literature, but is apparent in this context, since much teaching depends on developing and maintaining a supportive but asymmetric relationship with the students.

- *Put.* These can be iconic or metaphoric gestures, where the instructor "puts" abstract concepts or objects somewhere into the visible space to help describe their relationships to one another.

- *Spread.* These are gestures where both hands and arms are extended in front of the body and spread outward in a circular fashion. Spread gestures may be iconic or metaphoric, and often correspond to an important point in the discourse. However, they often serve as pedagogical commentary, independent of lecture content, indicating the difficulty of the content.

- *Swipe.* These occur when one or both arms are moved simultaneously in one direction. These tend to be metaphoric gestures, e.g., an instructor makes a swipe gesture to

indicate that an abstract object has moved.

- *Close & Open.* These encompass a set of gestures that are visually similar to spread gestures, i.e., hands and arms are spread outward or inward in a circular motion, however, arms are generally not extended and therefore they form a much smaller spread. They are considered a separate class since they are less semantically relevant than spreads and are best considered as beats.

- *Flip & Swing.* These are gestures where one or both hands are flipped in a small circle. These pedagogical gestures indicate the continuation of a theme in the discourse. These gestures can also be considered as a beat (two phase) form of a cohesive gesture, a kind of pedagogic punctuation or backward reference.

- *Touch.* These are simple beat gestures where the instructor touches an object (usually the table, glasses, etc.) as a beat or as a pedagogic "timeout".

- *Pointing.* These are clearly deictic gestures and accounted for the majority of gestures in both videos (see Table 4.1). When an instructor points, it generally means that they wish the students to pay attention to a specific region of the slide or blackboard.

- *Hold.* In between gestures, instructors are sometimes noticed to stay relatively motionless. Some of the existing literature may consider this non-gesture to be a phase separating the preparation, stroke and retraction phases. Holds usually indicate that the discourse is focused on a specific point, and it can often be a deliberate pedagogical gesture.

- *Others.* A number of gestures were observed but held no noticeable semantic significance or did not occur frequently enough to merit their own class. These gestures were assigned the "others" class.

4.2.2 Granularity of Avatar Poser Tool

One of the lecture videos was used to examine and improve the completeness of the gesture grammar. If a pose could not be expressed by the current grammar, the annotator verbally described possible additions to the grammar that would enable it to express that pose. From

the video, 183 poses were encoded using the current tool, whereas 91 poses could not be expressed by the grammar. From analyzing the necessary additions for these 91 poses, we explored five additions that significantly increased the expressiveness of the grammar. Extra precision on shoulder direction and elbow angle helped encode 51 of the poses; 22 poses needed shoulder joint rotation; and 44 needed forearm pronation/supination. Otherwise, the grammar appeared well-matched to what was observed.

We also found several ambiguities when proposing additions to increase the expressiveness of the gesture grammar, since different joint configurations can lead to almost the same overall pose. The main source of ambiguities occur when two rotation axes coincide, such as the forearm and shoulder when the arm is straight.

4.2.3 Dimensionality Reduction

We applied Principal Component Analysis (PCA) to the pose data to gain additional data which can help us refine the tool and pose representation, as well as provide insights regarding the pattern of poses in gestures.

Examining the entire corpus of poses for one instructor (instructor A), we compressed each pose using the annotation tool, into a ten-dimensional vector whose components encoded the quantized positions of: "*body, face, left hand, right hand, left arm, left shoulder, left elbow, right arm, right shoulder, right elbow.*" We map each component of the pose to a value either between -1 to 1 or 0 to 1, divided into equal intervals. We used PCA for dimensionality reduction, and found that the first two principal components account for more than half of the variance of poses (51%), and the first five account for nearly all (81%). These eigengestures can be roughly interpreted as:

- Right arm raised with elbow straightened versus right arm lowered with elbow bent, which is basically a point versus a rest gesture (33%, see Figure 4.5).
- Both arms used symmetrically from the shoulder, either both to the side or both forward, which is basically a spread versus a rest gesture (18%).
- Right elbow used anti-symmetrically from the left elbow in a "Mr. Roboto dance"-like chop (12%).

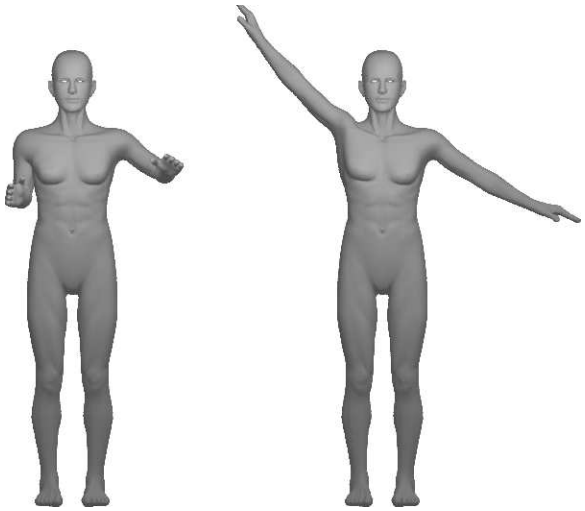


Figure 4.5: Example of an eigengesture. The left and right poses correspond to the maximum and minimum values and basically represent a point versus a rest.

- Both hands opened or closed symmetrically (9%).
- Right arm raised, but with bent elbow (9%).

We note that the position of body and face did not contribute much to the gesture variance, which is expected, since the body of the lecturer is usually turned towards the class. Also, due to low granularity in the hand annotation, independent hand information also did not significantly contribute to the variance.

4.2.4 Inter-Annotator Analysis

Approximately 60% of video A was annotated by two independent, novice annotators. We attempted to compare these results. As previously stated, there is no standardized method for comparing gesture annotations, so we approached this intuitively.

As a rough metric, we compared the work of the two annotations in terms of segmentation. A visualization of the comparison is shown in Figure 4.6. Colored regions represent frames that are marked as belonging to a gesture. It can be seen from the figure that, using this metric, inter-annotator agreement is strong: roughly 74% agreement, not too far from reports in the existing literature.

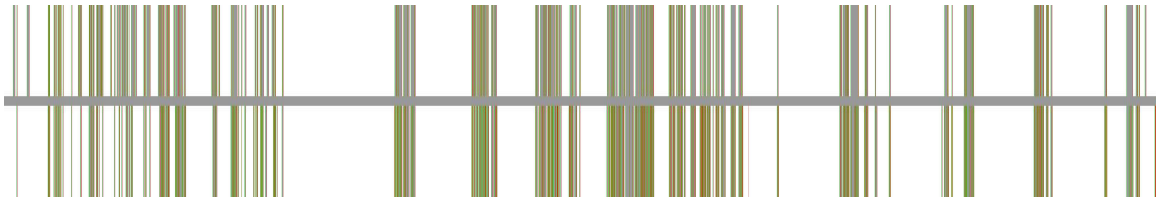


Figure 4.6: Inter-annotator comparison. The colored regions indicate parts of (roughly half) of video A that have been marked as a frame belonging to a gesture. The line in the middle separates the work of the two independent annotators: one on top, one below. Red and green ticks mark the boundaries of gestures: green ticks indicate the beginning of a gesture, and red ticks indicate the end.

More precise segmentation however is notably more difficult. In Figure 4.6, green tick marks indicate the start of gestures, and red ticks mark the end of gestures. From this perspective, inter-annotator agreement is very low and is difficult to compare. As previously mentioned, what one annotator may mark as one long gesture, another may break into several smaller gestures.

4.2.5 Observations and Evaluation

We observed 372 and 494 gestures from videos A and B respectively. These gestures were broken down into the nine classes as summarized in Table 4.1. We noticed in these lecture videos three observations about which the literature is basically silent.

First, we noticed that gestures are highly idiosyncratic. For instance, instructor B seldom does the spread gesture and tends to do more point and hold gestures than the instructor A. The lecture content clearly impacts the gesture distribution. For example, instructor B uses two hands to point at slides to explain details of matrices, while instructor A points with just one hand since discourse was mostly about theoretical topics. Nevertheless, habits of each instructor clearly exist. In video B, the instructor relies on slides more, so deictic gestures occur more frequently. In video A, the instructor refers to the slides less, and so relies on iconic or metaphoric gestures more.

Second, we observed that the gestures are often pedagogic and are correlated to the

Semantic Class		A	A (%)	B	B (%)
Close & Open	B	42	11.29	49	9.91
Flip & Swing	B, C, P	21	5.65	4	0.81
Hold	P	33	8.87	71	14.37
Point	D	123	33.06	292	59.11
Put	I, M	16	4.30	10	2.02
Spread	I, C, P	81	21.77	24	4.86
Swipe	M	8	2.15	5	1.01
Touch	B, P	5	1.34	7	1.41
Others		43	11.56	32	6.48
Total		372		494	

Table 4.1: Counts and distributions of gestures according to the nine semantic classes for videos A and B. The abbreviations I, M, B, C, D, P stand for *iconic*, *metaphoric*, *beat*, *cohesive*, *deictic* and *pedagogic* respectively. Four of the gesture classes (spread, flip & swing, touch, hold) appear to be pedagogic.

difficulty and pacing of the lecture material. Explanatory gestures, such as swinging or spreading, suggested that key points were being told. More intense gestures indicated that the material was more difficult or an important concept, while slower gestures seemed to indicate content that was less important.

Third, we noticed that successive gestures tend to overlap on their ends, and do not completely follow the three-phase model of gestures. This has made it difficult to tag adjacent gestures, because there is no hard boundary between when one gesture ends and the next gesture begins. Our tool was modified to allow overlapping gestures, shown as separate layers.

The observations made here influence much of the work in this thesis, such as the importance of the point and spread gestures we seek build classifiers for in Section 3.1, and the idiosyncrasy of gestures lead to investing across a greater variation of speakers in Section 5.1 as well as to our exploration of extremal poses as defined by speaker-dependent models in Section 3.3.3 which turn out to be significantly correlated with audience neural activity in Chapter 6.

Chapter 5

Gestures and Indirect Measures of Engagement

In Chapter 3 we introduced methods for extracting gesture attributes from video. In Chapter 4 we manually examined video to find correlations between gestures and semantically relevant segments. In this chapter, we build on these results to demonstrate the correlation between gestures and semantics quantitatively. We approach this in three ways:

First, we hypothesize that parts-of-speech can be used as indicators of semantically relevant segments of video. There is precedence for this as reviewed in Section 2.2.2. We look for correlations between gestures and conjunctions.

Next, we perform user studies which allow us to explore different methods of presenting gesture data to users in a video browser in a useful way and to receive feedback on the types of gestures that may be engaging their attention.

Finally, to demonstrate the relationship between speaker gestures and audience engagement we perform an experiment whereby volunteers watch videos under observation by electroencephalography, allowing us to gather detailed engagement data at the neurological level. This data is then correlated to gesture attributes. Significance tests are performed to demonstrate the effect of speaker gestures on audience attentiveness, and specifically to identify the type of gestures which may pique the audience's interest.

5.1 Correlating Gestures with Conjunctions Indicating Contrast

In this section, we explore the correlation of arm gesture velocity with semantically significant moments in speech, as indicated by the use of certain conjunctions which indicate shifts in semantic weight between clauses. To demonstrate this, we produce a fully automatic classifier which labels a segment of text according to whether or not it contains a conjunction of a certain functional class using only visual features as input. We restrict ourselves to the domain of educational lectures, each with a single English speaker presenting a variety of subject matter. Our proposed system automatically extracts features based on arm motions of the lecturer. The features are extracted from segments of video corresponding to predetermined segments of text, provided in the form of subtitles which may be either generated through automatic speech recognition (ASR) or manual labeling. These results were initially presented in [Zhang and Kender, 2012].

5.1.1 Classifier

We frame this task as a binary classification problem. For each segment of video, we regularly sample every nth frame into a sequence of frames and associate with it the set of words of its corresponding natural language subtitle. In our experiments, a sampling rate of 3 frames per second was sufficient. We assign a binary label Y to the video and subtitle segment, where Y is $+1$ if its subtitles intersects with a set of conjunctions of interest, and -1 otherwise. In Section 5.1.3, we will discuss experiments with different sets of conjunctions

For classification, we use an AdaBoost-based classifier [Freund and Schapire, 1996] with decision trees as weak learners. We train using samples (X, Y) with arm angular variance as visual features X derived from only the video frames, as described in Section 3.3.1. The classifier and training system is depicted in Figure 5.1.

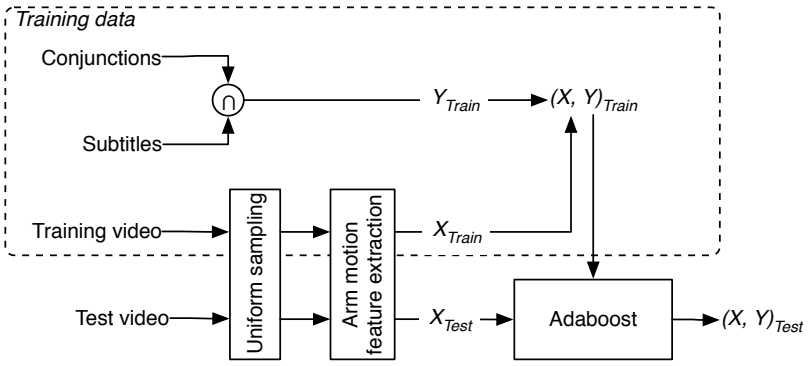


Figure 5.1: Overview of the classification and training system. For each video we compute gestural features through pose and flow estimation on sampled frames. Training labels are assigned based on the presence of certain conjunctions in segments of subtitles accompanying the video. The classifier attempts to assign labels to test samples based on gestural features alone.

5.1.2 Video and Subtitle Data

For our experiments we use data gathered from MIT OpenCourseWare which is free for public use. We extract clips from 10 videotaped lectures from three courses, each from a different field of study (MIT course numbers in parentheses): *Abdul Latif Jameel Poverty Action Lab Executive Training: Evaluating Social Programs 2009 (RES.14-001)*, *Principles of Chemical Science (5.111)* and *Multicore Programming Primer (6.189)*. Over 12 different speakers of different genders and ethnic backgrounds are present in these videos, mainly from RES.14-001 which is a seminar-based course. Each video features an individual English speaker lecturing in front of a classroom, usually standing in front of a slide or blackboard. The lighting condition varies, as do the appearance of the speakers due to clothes and other attributes. The videos are low-resolution (480×270 for RES.14-001 and 5.111, and 478×360 for 6.189) with a frame rate of 15 frames per second.

In addition to the video data, time-synced subtitles are also available for each lecture. For RES.14-001, the subtitles are produced by ASR from YouTube. Subtitles for 5.111 and 6.189 are manually produced by MIT.

We apply an automatic upper-body detector to the lecture videos and apply a simple

greedy heuristic to produce video segments at least 10 seconds long (subtitles are segmented correspondingly). Further, we manually remove videos which do not fit our criteria such as those containing multiple speakers in one frame (i.e., guest lecturers), the speaker writing on a blackboard, shots of the classroom, shots of the slides without a full shot of the speaker. Finally, we segment the remaining video according to each individual time-synced subtitle, producing the best association we have between word and gesture. This leaves us with 4243 video clips and corresponding subtitles totaling 3.83 hours, averaging 3.25 seconds in duration and 7.89 words each. We sample these videos at 3 frames per second.

Our reasons for using publicly available subtitle data rather than purely ASR-generated data are three-fold. Primarily, the quality of the manually produced subtitles are impossible to surpass. This is closely followed by the YouTube-generated subtitles which has quality significantly surpassing that of our own using open source software such as Sphinx[1]. Finally, we argue that the short duration of each subtitle and the intuitive need to search for gestures within a radius of a spoken word justifies the use of these subtitles even if they do not produce the finest grain word-to-gesture association.

5.1.3 Classes of Conjunctions

Conjunctions are parts of speech that connect sentences of clauses together. It is commonly classified into three categories, with differing semantic connotations: *coordinating*, *correlative* and *subordinating*.

We are particularly interested in conjunctions as they can indicate semantically important points in speech, such as points of contrast, emphasis or complexity. Being able to successfully identify these points could be applicable for video summarization or as cues in non-linear semantic video browsers. Furthermore, as common particles, conjunctions can be found in discourse regardless of subject matter. Another advantage in their being so common is their relatively high rate of recognition in ASR.

In this paper, we experiment with subsets of 45 commonly used English conjunctions denoted C. They are listed with their frequencies of occurrence in our dataset in Figure 5.2.

[1]Available at *http://cmusphinx.sourceforge.net/*.

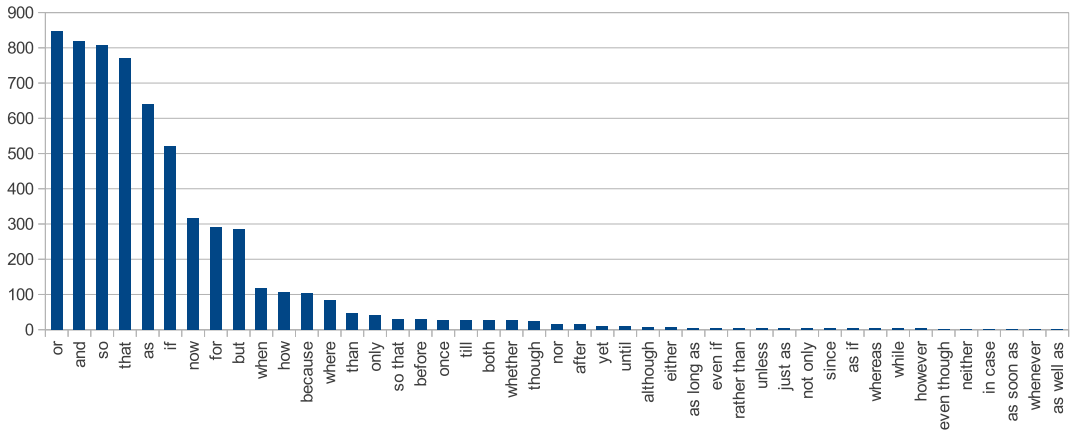


Figure 5.2: List of conjunctions C and their frequencies of occurrence in the dataset.

We experiment with the following subsets of C, which have different semantic implications. We test the common classes of coordinating C_{Coor}, correlative C_{Corr} and subordinating C_{Sub} conjunctions. We also introduce a class of "contrasting" conjunctions C_{Cont}, with words and phrases which indicate contrasts between the joined sentences or clauses. Examples of the members of these sets are given in Table 5.1. These sets are not necessarily mutually exclusive. The current classification of conjunctions (except for "contrasting") is derived from standard linguistics research [Scharton and Neuleib, 2001].

5.1.4 Observations

Using the 4243 samples described in Section 5.1.2 and sets of conjunctions described in Section 5.1.3, we perform experiments using 4-fold cross validation. Three folds are used for training, while the remaining fold is balanced (i.e., the same number of positive and negative samples are taken) and used for testing. The classifier selects from the testing fold those samples it believes, on the basis of gestures, must contain a conjunction of the given class. We measure performance of this selection according to precision ($P = \frac{TP}{TP+FP}$), recall ($R = \frac{TP}{TP+FN}$), F1 score ($\frac{2PR}{P+R}$), as well as overall classification accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$). Recall that TP, TN, FP, FN refer to true positive, true negative, false positive and false negative, respectively.

Class	Members	Pos (%)
C	*See Figure 5.2*	72.2
C_{Coor}	and, but, for, nor, or, so, yet	49.6
C_{Corr}	both, either, just as, neither, nor, not only, or, whether	21.0
C_{Sub}	after, although, as, as far as, as if, as long as, as soon as, as though, because, before, if, in order that, since, so, so that, than, though, unless, until, when, whenever, where, whereas, wherever, while	44.8
C_{Cont}	although, but, for, however, if, neither, nor, or, so, though, yet	46.2

Table 5.1: Subsets of C and examples of their members, as well as the percentage of the dataset which contains those conjunctions.

	C	C_{Coor}	C_{Corr}	C_{Sub}	C_{Cont}
Accuracy	0.508	0.556	0.500	0.530	0.549

Table 5.2: Classification accuracies. That is, the percentage of the balanced test set (both positive and negative samples) that is correctly classified.

We begin by computing the overall classification accuracy (i.e., number of positive and negative samples correctly classified) by training and testing on the sets of conjunctions. The results are summarized in Table 5.2. In the case of C_{Corr}, the output classifier was simply classifying all samples as negative. One possible reason for this is the dearth of positive training samples which appears to prevent learning. As such, we will exclude it from further analysis.

We evaluate the classifiers trained and tested on all classes except for C_{Corr} according to precision/recall/F-score, as shown in Figure 5.3. Perhaps unsurprisingly, C results in a classifier that produces extremely high recall, but at a precision of 0.508 on a balanced test set, it is not better than chance. Not all conjunctions, e.g., "and", appear to have strongly associated gestures. That is, "and" is truly a stopword.

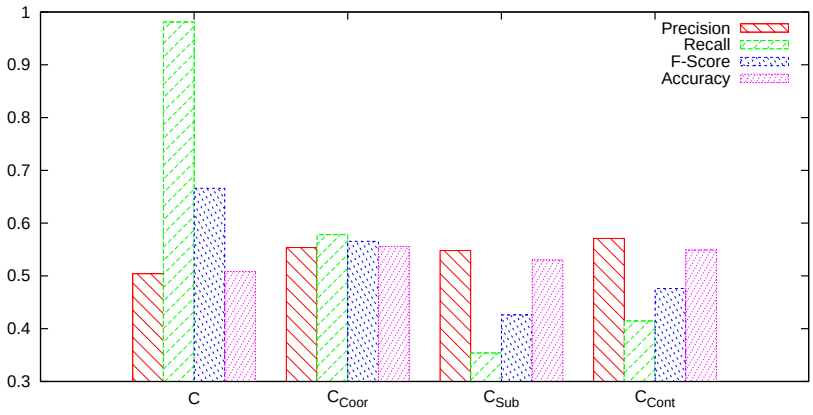


Figure 5.3: Overview of classification performance for different conjunction classes.

Classifiers trained on C_{Sub} offers slightly higher classification accuracy (0.530) and higher precision (0.548) but low recall (0.354). We speculate that the reason for this is the lack within C_{Sub} of very common conjunctions such as "and" and "or" (both with very high frequencies in the dataset, as seen in Figure 5.2). In fact, it can be seen in Figure 5.3 that "and" results in greater recall in C_{Corr} versus C_{Cont}, followed by "or" which results in greater recall in C_{Cont} versus C_{Sub}.

Classifiers trained on C_{Coor} and C_{Cont} offer perhaps the strongest argument for the correlation between arm gesture motion and some conjunctions. These classifiers performed similarly, although C_{Coor} resulted in significantly higher recall, but slightly less precision. The reason for the similar performance is likely due to the overlap between conjunctions in these sets, with the exception of "and", which is likely the cause for the difference in recall, as it is an extremely common conjunction. Set C_{Cont} contains conjunctions which are usually used to denote significant changes in meaning during discourse, and its high classification accuracy and precision offer slight but existent evidence of a correlation with arm gesture motion.

5.2 Gestures as Indicators of Segments of Interest for Video Browsing

To investigate how other gesture attributes may correlate with engagement, we perform a user study with two goals: first, to explore ways to present gesture data to viewers in a useful way through a video browser, and second, to gain additional feedback from viewers as to which speaker gestures may be particularly interesting. In this section, we will propose two user interface elements, describe the user study, and summarize some of the observations gathered based on subjects' feedback (both quantitative and qualitative).

5.2.1 User Interface

We propose two possible ways to present gesture-based features to users to aid them in video browsing. They are presented to users underneath a standard video player as shown in Figure 5.4.

5.2.1.1 Gesture Attributes Graph

The most straightforward approach is to simply present the results of the gesture attribute computations (particularly velocity and direction change as described in Section 3.3) visually in the form of a center-aligned graph. The graph is center-aligned so that small signals are more easily distinguished from no signal. A graph for each attribute is displayed on a separate row. A red line indicates the playback time of the video in the gesture attributes graph, as shown in Figure 5.4 highlighted by ③. The gesture attributes graph is illustrated in Figure 5.4 under ①. Users can click anywhere on the gesture graph to jump to the corresponding time in the video. Intuitively, we hope to observe that users would skip sections with low gestural activity, and pay more attention to segments with high gestural activity.

5.2.1.2 Emphasized Subtitles

Another way to highlight segments of interest is through emphasized subtitles, which would help users to quickly scroll through the transcribed content of a video, with key words

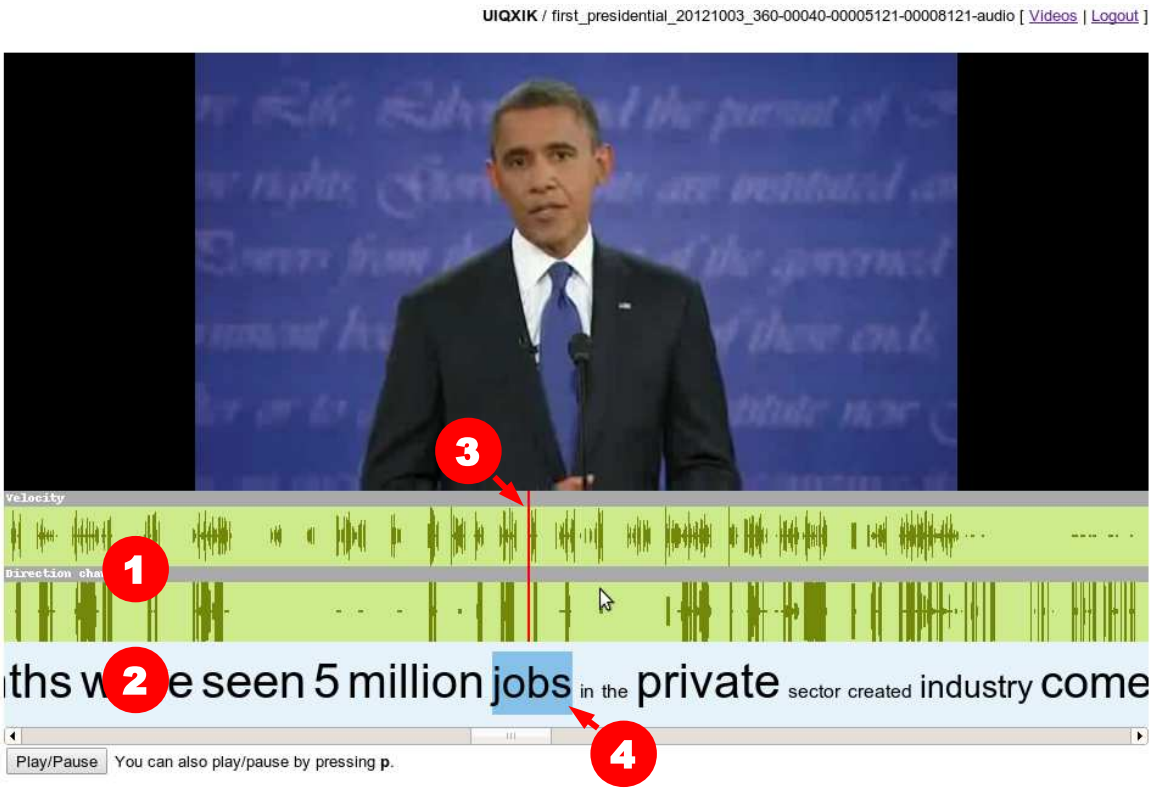


Figure 5.4: The user interface presented to subjects in our user study implementing the ① *gesture attributes graph* (Section 5.2.1.1) which indicates the velocity and direction change gesture attributes for each frame, and the ② *emphasized subtitles* (Section 5.2.1.2) which highlights subtitles based on associated gestures. A time cursor (the red bar highlighted by ③ and blue box highlighted by ④) marks the position in the video.

enlarged.

We obtain transcripts of the political debates and apply the Sphinx Long Audio Aligner[2] to temporally align individual words to their position in audio so that each word has a starting frame w_s and end frame w_e. Due to inaccuracies, many words are dropped, so the ability to watch the video is still important to viewers, particularly during the user study.

Each word, minus stopwords, is then resized according based on the associated gesture attributes. Since both velocity $\mathbf{v}_f$ and direction change $\mathbf{d}_f$ are normalized, one simple way to compute a "weight" w_w for each word is to take the maximum attribute within the word time frame, that is:

$$w_w = \max_{w_s \leq f < w_e} \{\max\{\mathbf{v}_f, \mathbf{d}_f\}\}$$

Intuitively, this can be though of as the "most emphatic" gesture within the span of each word. Different methods for combining weights remains for future work.

As a baseline in our user study (Section 5.2.3), we use named entity detection to determine which words to highlight. The named entities are automatically identified using a publicly available service[3], with the weights determined by the relevance of the detected named entities. Most of these are the names of people, places, events and government programs.

5.2.2 User Study and Ground-Truth Data

For our user study, we used recorded video of the first and third 2012 US Presidential debates, which are publicly available. The videos are very high quality with a resolution of 640×360 pixels at 24.58 fps, a stationary camera, and tend to focus on the upper bodies of the speakers, who gesticulate frequently while they speak. Their hands appear in and out of view.

To select these shots, we automatically segmented the videos using HSV color histogram comparison, with 3D histograms with 16 bins for each dimension, compared using the L2 distance and a threshold of 0.1 similar to [Smeaton *et al.*, 2010]. This resulted in 322 shots. From these, we randomly selected 36 shots with durations between 60 to 100 seconds,

[2]CMU Sphinx Long Audio Aligner, *http://cmusphinx.sourceforge.net/wiki/longaudioalignment*

[3]TextRazor, *http://www.textrazor.com/*

totaling about 48 minutes. This duration was intentionally selected so that subjects in the user study could complete the task in under one hour.

To compute gesture attributes, the hand tracking algorithm was applied to each video clip at a frame rate of 12 Hz.

We presented each of these 36 videos to 3 different human raters via Amazon Mechanical Turk[4], who were asked to watch the complete video and each provided up to 10 keywords to summarize the spoken content in the videos, without time limitations. A total of 7 different raters contributed unevenly to the set of all keywords. The keywords were then corrected for spelling, split into words (tokenized by whitespace or punctuation) and stemmed. For each video, we retain only the keyword stems which were selected by at least 2 raters. From 792 unique stems, 415 were retained for evaluation.

5.2.3 User Study

We recruited 12 university students and observed their usage of different configurations of our combined video browsing tool in an IRB-approved user study. These subjects were observed in a proctored setting and are different from the raters that provided the ground-truth keywords. The results of this user study are discussed here.

Each subject was given a brief training session and then asked to provide keywords for the spoken content of the 36 videos, by typing the keywords into a textbox. The entire session takes approximately 45 minutes. As a restriction, for each video, they were only given *half* the time of the video duration to complete the task (i.e., to both skim the video and input keywords). This is done to force subjects to skim through the video and make it impossible to watch the entire video. Each video is presented to 3 different subjects in each of the following configurations:

1. Subtitles emphasized according to *named entity* relevance *without* the gesture attributes graph.

2. Subtitles emphasized according to *gesture attributes without* the gesture attributes graph.

[4]Amazon Mechanical Turk, *http://www.mturk.com*

Precision

Graph \ Emphasis	NE	GA
No	0.52	0.52
Yes	0.50	**0.55**

Recall

Graph \ Emphasis	NE	GA
No	0.33	0.33
Yes	0.33	0.34

Table 5.3: Precision and recall of subjects for the task of selecting keywords for videos. The columns indicate which type of emphasis was used (NE: named entity, GA: gesture attributes) and the rows indicate whether the gesture attributes graph was visible.

3. Subtitles emphasized according to *named entity* relevance *with* the gesture attributes graph.

4. Subtitles emphasized according to *gesture attributes with* the gesture attributes graph.

No video was viewed twice by the same person and each person performed the task using a variety of configurations.

Using the ground-truth keywords, we can evaluate how well subjects performed using precision ($P = \frac{TP}{TP+FP}$) and recall ($R = \frac{TP}{TP+FN}$). The results of our user study are shown in Table 5.3, where the scores are averaged across subjects. We note that the highest performance was attained through the use of gestures alone, without reference to any external sources of knowledge such as the catalog of named entities.

As a comparison, we find the average precision and recall of each ground-truth rater (weighted by the number of videos labeled) to be 0.72 and 0.81, respectively. It is interesting to note that even with half the time of the video, subjects using browsing aids achieved a relatively high precision as compared to the ground-truth raters who had unlimited time and no pressure.

5.2.4 Observations

We also asked subjects to complete a questionnaire upon completion of their task. From this, we are able to make qualitative observations about our interface.

Subjects were asked to rate the "helpfulness" of each interface configuration from 1 (very unhelpful) to 5 (very helpful). Their average responses are shown in Table 5.4. Interestingly,

Helpfulness

Graph \ Emphasis	NE	GA
No	3.6	3.3
Yes	**4.2**	3.6

Table 5.4: Average of user study subjects' "helpfulness" ratings for each configuration which range from 1 (very unhelpful) to 5 (very helpful). The column and row headings are the same as Table 5.3.

the most popular configuration (subtitles emphasized according to named entities with the gesture attributes graph visible) also corresponds to the *lowest* precision in Table 5.3. A larger user study would be needed to verify significance, although comments seem to indicate that the gesture attributes graph itself was popular whereas the named entities emphasis allowed users to see more subtitles at a given time.

First, we observed that named entities emphasis was preferred, but mostly because it means more words were visible in the subtitles box. Subjects commented that having the subtitles with certain words enlarged were extremely helpful in identifying keywords but, as emphasis by gesture led to more words being enlarged, fewer words were visible at any given time. Future iterations would quantize the gesture attributes (e.g., make it binary so users can be more decisive) or explore different methods of indicating emphasis while controlling for the number of words visible.

Second, sudden gestures seemed to indicate greater semantic value. When asked if they noticed any particular types of gestures that seemed to correlate with semantic importance, subjects noted that pointing and "sudden" gestures were more likely to catch their attention. Further work on gesture attributes could capture these better than the velocity attribute does.

Finally, velocity and direction change were of similar value. When asked if they tended to favor one attribute over the other when presented with gesture information, subjects did not consciously have a preference. From a user's perspective, a unified feature would probably be less confusing.

The studies in this chapter were done to help us gain more insight into what attributes of gestures may pique audiences' attention. For instance, the development of the extremal poses gesture attribute (Section 3.3.3) was inspired by user feedback from the study. The use of indirect measures such as accompanying speech and user studies allowed us to do so inexpensively. In Chapter 6, we will study how gestures correlate to engagement with greater rigor using EEG—a more expensive but direct method of measuring engagement which functions by recording and interpreting neural activity.

Chapter 6

Correlating Gestures with EEG

In Chapter 5 we examined how gestures affect audience engagement, measured indirectly through parts of speech or user studies. Based on the results and feedback from users in those studies, we now seek to more rigorously measure audience engagement and demonstrate its correlation against gestures quantitatively, as well as to identify specific attributes—both neural and gesture—which drive correlation and show statistical significance.

Unlike the transcripts or software user studies used in Chapter 5, recruiting subjects and performing EEG experiments is considerably more expensive (in terms of compensating participants financially) and time consuming (the wet-gel setup we use takes approximately an hour to set up and needs to be washed out of hair). Given this, we focus on a single domain for simplicity: the 2012 U.S. presidential debates. This simplifies the parameters of our experiment as the American political landscape is dominated by two parties: the Democratic Party and the Republican Party. Although voters who identify as Independent have seen a resurgence in popularity in recent times according to Gallup's [Jones, 2012] from 33% in 1988 to 40% in 2011, the vast majority of voters voted for one of these two parties while only 1.6% voted for others in the 2012 election. Each party was also represented by a single leader: U.S. President Barack Obama for the Democratic Party and former Massachusetts Governor Mitt Romney for the Republican Party.

Despite restricting the experiment to such a narrow domain, the results of the study could nonetheless have significant impact due to the importance of the domain. Political consulting is big business in America, where even the slightest edge in capturing audiences'

attention could command lucrative fees, and ultimately change the outcome of an election.

6.1 Experiment

We showed 6 videos, each approximately 10 minutes long totaling 61 minutes, to each of 20 subjects in an electrostatically shielded room[1]. Due to artifacts and noise during data collection, approximately 14 minutes of data was ultimately discarded, leaving us with 47 minutes worth of EEG data for analysis. This experiment was approved by the Institutional Review Board of Columbia University and all subjects gave written informed consent prior to the experiment. Subjects were instructed to sit comfortably and watch attentively to the debates and refrain from overt movements. The order of the video clips were kept the same across subjects. In between each of the 6 videos, subjects were given a break and the experiment did not resume until the subject indicated they were comfortable.

Subjects were fitted with a standard 64-electrode cap following the international 10/10 system, as shown in Figure 6.1. EEG data was recorded using a BioSemi Active Two AD Box ADC-12[2] system at 2048 Hz which was downsampled to 512 Hz for processing.

The details of how the videos were created from clips of the presidential debates are described in Section 6.1.1, and details of the subjects we selected are given in Section 6.1.2.

6.1.1 Video Stimuli

We use clips of the 2012 U.S. Presidential Debates as video stimuli to present to the subjects. We use only the first and third debates where the speakers stand at a podium or sit at a table, and discard the second presidential debate (a town-hall style debate where speakers are allowed to move around freely and thus making automatic recognition and analysis of gestures extremely difficult) and the vice-presidential debate. This allows us to focus on two people: Obama and Romney. Similar to Section 5.2.2, the videos have a resolution of 640×360 pixels recorded at 24.58 fps using a stationary camera, and tend to focus on the upper bodies of the speakers, who gesticulate frequently while they speak.

[1] ETS-Lindgren, Glendale Heights, IL, USA.

[2] BioSemi, The Netherlands.

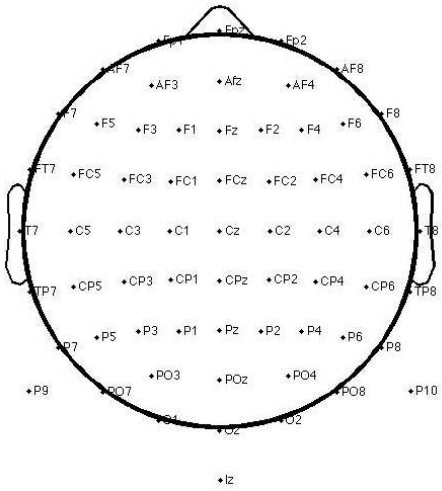


Figure 6.1: Approximate positioning of the electrodes on our EEG skull cap.

First, we select clips that have a greater likelihood of resulting in reliable gesture attribute extraction. To achieve this, we automatically segmented the videos using HSV color histogram comparison with 3D histograms with 16 bins for each dimension and compared using the L2 distance and a threshold of 0.1 similar to [Smeaton *et al.*, 2010]. We then manually discard all clips that do not contain a continuous frontal shot of a single speaker. Given the high quality and structured nature of the recording, most clips satisfy this constraint. We also manually identify the speaker in each clip.

Next, to keep this experiment as fair as possible, we manually select clips such that Obama and Romney are given equal "airtime". The debates also cover a range of topics, as defined by the moderators of the respective debates and listed in Figure 6.2, sorted by increasing level of subjects' interest. We also attempt to balance the time each speaker spends discussing each of these topics, i.e., each speaker spends approximately the same amount of time discussing each topic, but the time discussing each topic may differ.

Ultimately, 28 clips were selected (8 from the first debate and 6 from the third debate featuring Obama, and 8 from the first debate and 6 from the third debate featuring Romney). The total duration of the clips is approximately 30.5 minutes, so each clip is 65 seconds on average.

Finally, to produce the video stimuli to present to subjects, we randomly shuffled the 28

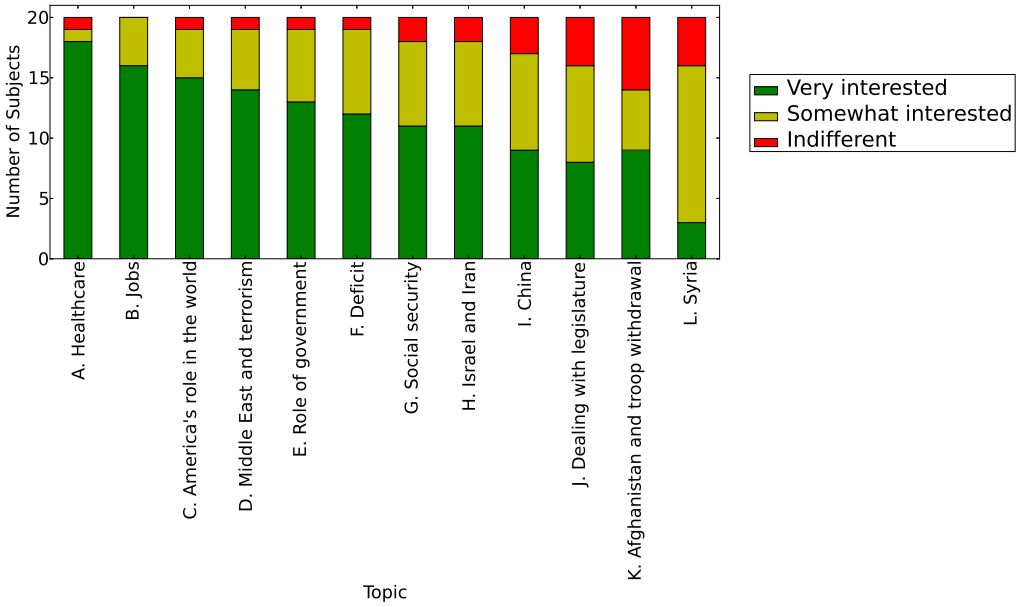


Figure 6.2: List of topics covered in the video stimuli as defined by the debate moderators, along with subjects' level of interest. Topics are listed in order of increasing weighted subject interest.

clips, and then made silent versions of each by removing the audio track. In total, we show 61 minutes of video to each subject in total. In order to allow for a break approximately every 10 minutes, we separate this into 6 videos. We create each of these 6 "long videos" by concatenating the smaller clips (both audible and silent versions) together, as shown in Figure 6.3. In each case, the audible clip is always played after its silent version, so that subjects do not associate visual scenes with the spoken content on a second viewing of each clip. Each of the long videos begins with a 5-second video of a countdown, followed by the clips, each separated with a 3-second blank (i.e., black screen). The silent and audible versions of each clip are always placed in the same long video. The final durations of the 6 videos are 9.32 minutes, 10.39 minutes, 10.72 minutes, 11.93 minutes, 9.82 minutes and 8.94 minutes.

However, noise and artifacts such as interruptions to the electrodes (non-injurious to the subjects) which occurred during the EEG data collection process led us to discard data recorded during both the audible and silent versions of 7 video clips for all subjects. The

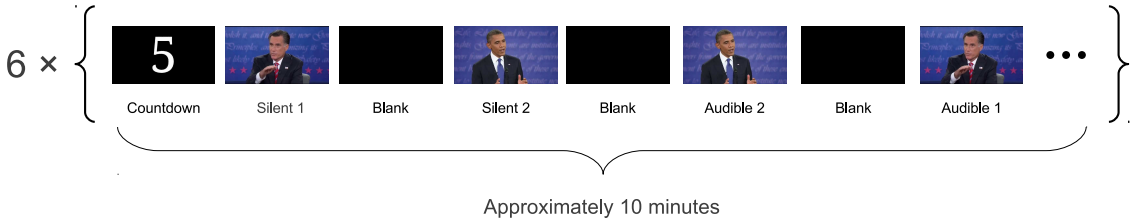


Figure 6.3: Each of the 6 videos shown to each subject begins with a 5-second countdown, followed by silent then audible versions of clips of the debates, each separated with a 3-second blank screen.

audible and silent versions of these 7 clips totaled 14 minutes in duration, therefore leaving us with 47 minutes of EEG data for analysis.

6.1.2 Subjects

For this study, we recruited 20 college students and young professionals of various ethnic backgrounds and nationalities. In addition to their age and gender, they were also asked to identify their political leanings on a scale from 1 to 5, where 1 is Strongly Democrat and 5 is Strongly Republican. Only those who were not neutral (i.e., 3) were considered for this experiment. Those who were not American and therefore did not vote but nonetheless followed American politics and thus identified with a political party were included in this experiment.

We stratify subjects according to their political affiliation (Democrat or Republican) and their gender (male or female). The 20 subjects are divided into 10 Republicans and 10 Democrats, or 10 females and 10 males. The breakdown of the subjects by political and gender groups is shown in Table 6.1. The subjects' ages range from 18 to 31 with a median age of 21. A breakdown of ages by group is available in Table 6.2.

To account for possible topic bias (i.e., what if one group was overwhelmingly interested in candidates' opinions on healthcare while another was only interested in their thoughts on the Middle East), subjects were also asked to indicate their level of interest in each of the topics from Figure 6.2 as one of: *indifferent*, *somewhat interested*, or *very interested*. Subjects were selected randomly restricted only by age, gender and political affiliation

	Female	Male	Total
Democrats	5	5	10
Republicans	5	5	10
Total	10	10	20

Table 6.1: Number of subjects in each political and gender group.

Group	No. Subjects	Min. Age	Med. Age	Max. Age
All	20	18	21	31
Democrats	10	18	21	29
Republicans	10	18	22	31
Females	10	18	21	29
Males	10	18	21.5	31

Table 6.2: Range and medians of ages of subjects divided by group.

independent of their interests in the topics. Their responses are aggregated by group and shown in Figure 6.4. It can be seen that Democrats and Republicans have roughly equal numbers of subjects that are at least somewhat interested in each of the topics. However, more males expressed indifference in each topic than females, therefore we do not believe personal interest in topics is a confounding factor in our correlation analysis of Democrats versus Republicans, but may be one when comparing females versus males. One possible confounding factor when comparing Democrats versus Republicans is the fact that this experiment took place beginning roughly 6 months *after* the 2012 U.S. presidential election, so the outcome was already known. Nevertheless, we believe this should not significantly affect our main hypothesis, that is, how audiences react to speakers' gestures, but it is something to keep in mind when making observations across party lines.

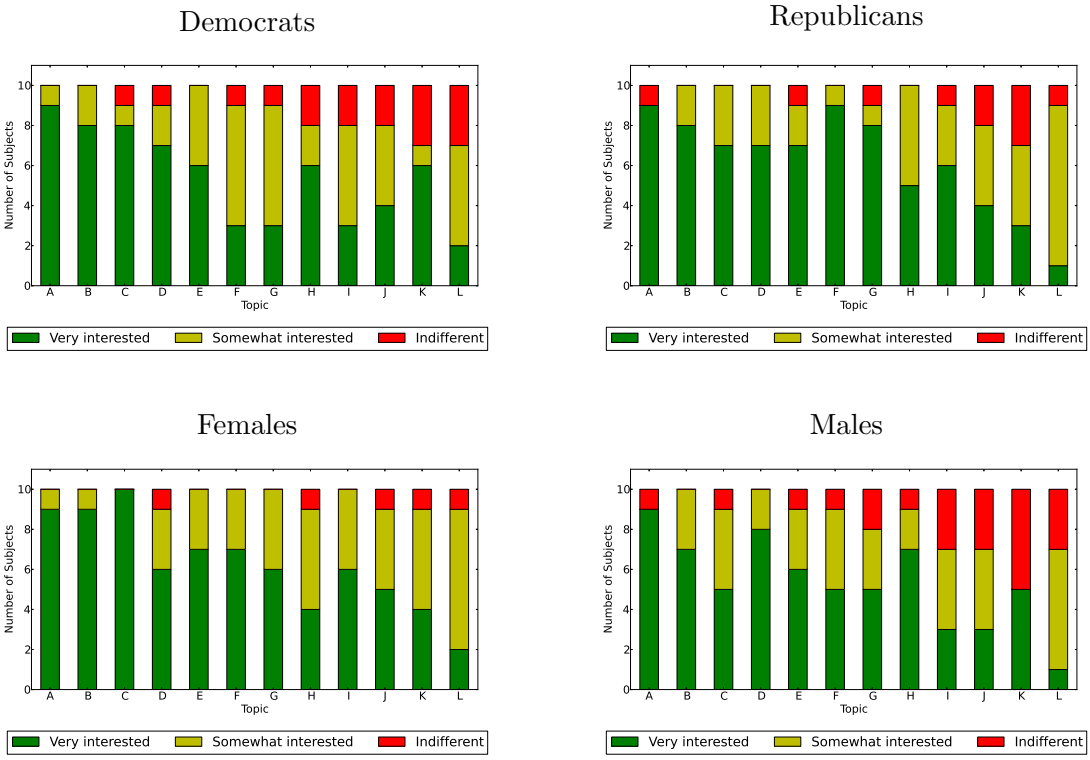


Figure 6.4: Subjects' interests in each topic, separated by group. Topics correspond to Figure 6.2. Democrats and Republicans have roughly equal numbers of subjects that are at least somewhat interested in each of the topics. However, more males expressed indifference in each topic than females.

6.2 Correlations

6.2.1 Deriving Engagement from EEG

Once EEG data was collected from all subjects for each clip, we apply standard post-processing. A software-based 0.5 Hz high pass filter was used to remove DC drifts, and 60 Hz and 120 Hz notch filters were used to minimize line noise. Eye blink artifacts were removed by independent component analysis (ICA) through EEGLAB [Delorme and Makeig, 2004]. EEG samples whose squared magnitude falls above four standard deviations of the mean power of their respective channels were replaced with zero.

To identify moments of engagement, we apply the method of [Dmochowski *et al.*, 2012] on EEG data across all subjects for each clip. Intuitively, this method works on the hy-

pothesis that only moments of stimulus which engage audiences will elicit a common neural response from most audience members, and therefore seeks to find maximal correlations between subjects' EEG data.

In this method, for a given clip, let $\mathbf{X}^{(i)} \in \mathbb{R}^{D \times T}$ be the EEG data of subject i where D is the number of channels (electrodes, so $D = 64$) and T is the number of time samples (on average for each clip, $T = 33280$). To derive engagement from a group of N subjects, form aggregated matrices $\mathbf{X}_1, \mathbf{X}_2$ as follows:

$$\mathbf{X}_1 = \left[\mathbf{X}^{(1)} \ \mathbf{X}^{(1)} \ \ldots \ \mathbf{X}^{(1)} \ \mathbf{X}^{(2)} \ \mathbf{X}^{(2)} \ \ldots \ \mathbf{X}^{(N-2)} \ \mathbf{X}^{(N-2)} \ \mathbf{X}^{(N-1)} \right] \quad (6.1)$$

$$\mathbf{X}_2 = \left[\mathbf{X}^{(2)} \ \mathbf{X}^{(3)} \ \ldots \ \mathbf{X}^{(N)} \ \mathbf{X}^{(3)} \ \mathbf{X}^{(4)} \ \ldots \ \mathbf{X}^{(N-1)} \ \mathbf{X}^{(N)} \ \mathbf{X}^{(N)} \right] \quad (6.2)$$

That is, columns of $\mathbf{X}_1, \mathbf{X}_2$ correspond to $\binom{N}{2}$ combinations of pairs of subjects. Therefore, $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{D \times \binom{N}{2} T}$. We wish to find a weight vector $\mathbf{w} \in \mathbb{R}^D$ such that the Pearson product-moment correlation coefficient between $\mathbf{Y}_1 = \mathbf{X}_1^T \mathbf{w}$ and $\mathbf{Y}_2 = \mathbf{X}_2^T \mathbf{w}$ is highest, that is:

$$\mathbf{w} = \arg\max_{\mathbf{w}} \frac{\mathbf{Y}_1^T \mathbf{Y}_2}{\|\mathbf{Y}_1\| \|\mathbf{Y}_2\|} \quad (6.3)$$

After differentiating, $\mathbf{w}$ can be found by solving a generalized eigenvalue problem, following the full derivation in [Dmochowski *et al.*, 2012]. As such, there are multiple solutions. The weight vector $\mathbf{w}$ that maximizes the correlation coefficient between $\mathbf{Y}_1$ and $\mathbf{Y}_2$ corresponds to the largest eigenvalue, and similarly for the second, third, etc. We take the three weight vectors corresponding to the three largest eigenvalues (i.e., the three strongest correlations) and compute correlations between EEG data after applying these weight vectors. We refer to the results (i.e., correlations) computed from the weight vectors as components 1, 2, 3 (corresponding to the largest eigenvalues and descending).

We can compute correlation at a finer time granularity and determine significance by computing Pearson correlation between corresponding 5-second EEG segments (across subjects for a given clip) at 1-second intervals. To determine significance, a permutation test [Fisher, 1935] is applied by splitting the EEG data into 5-second non-overlapping windows, randomly shuffling and taking the correlation at $p = 0.05$ (Bonferroni corrected). An example of inter-subject correlations for a video clip for 3 components and its significance threshold is shown in Figure 6.5.

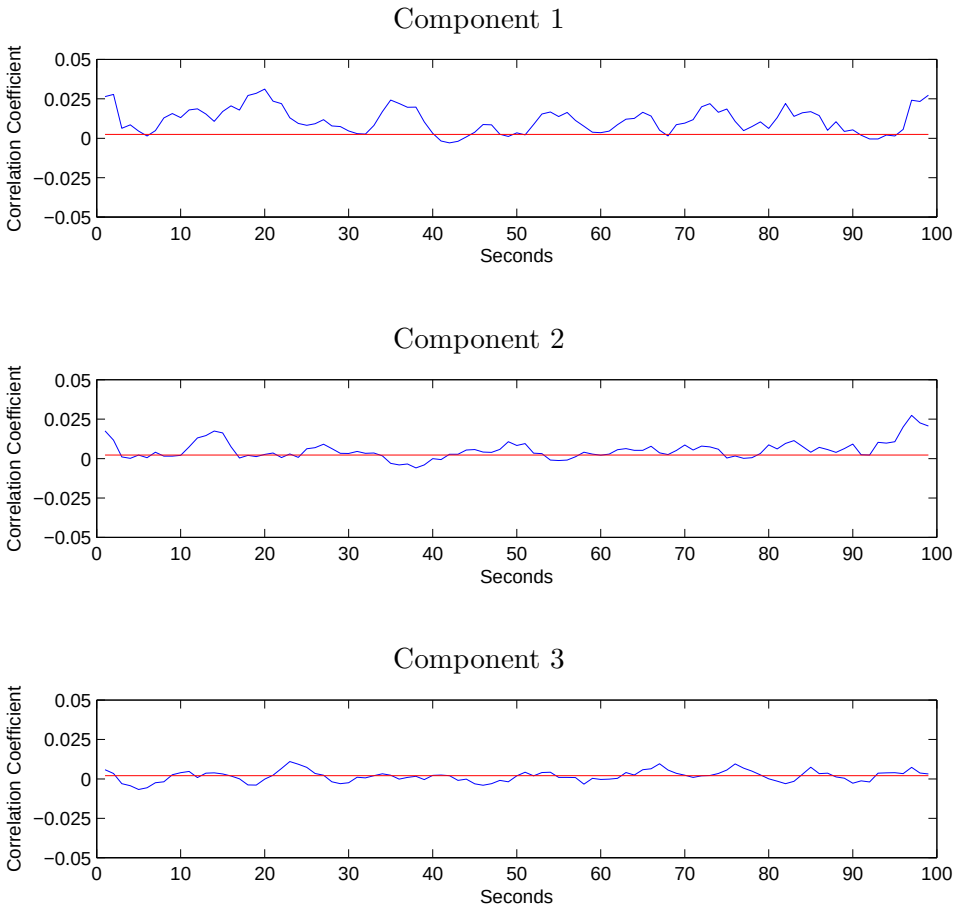


Figure 6.5: Inter-subject correlation for a video clip for three components (blue line), as computed according to [Dmochowski *et al.*, 2012]. The red line is the significance threshold.

To summarize, for each clip, the method of [Dmochowski *et al.*, 2012] gives us a correlation coefficient $r_{t,c} \in [-1, 1]$ at time sample t and the corresponding correlation coefficient $R_{t,c}$ from a permutation test at $p = 0.05$ (a threshold of significance) for each of the top three components c. Time samples which have significant correlations are taken as those times where audiences' attention is engaged. We transform the coefficients $r_{t,c}$ into a feature vector of engagement as follows:

$$e_{t,c} = \max(r_{t,c} - R_{t,c}, 0) \tag{6.4}$$

Because insignificant correlations all equate to inattentiveness (as far as we are concerned), we set all values below the significance threshold to zero. Also, as a result of this

transformation $e_{t,c} \in [0, 1]$, much like our visual features described in Section 3.3. An example of engagement features derived from the corresponding EEG components shown in Figure 6.5 is shown in Figure 6.6.

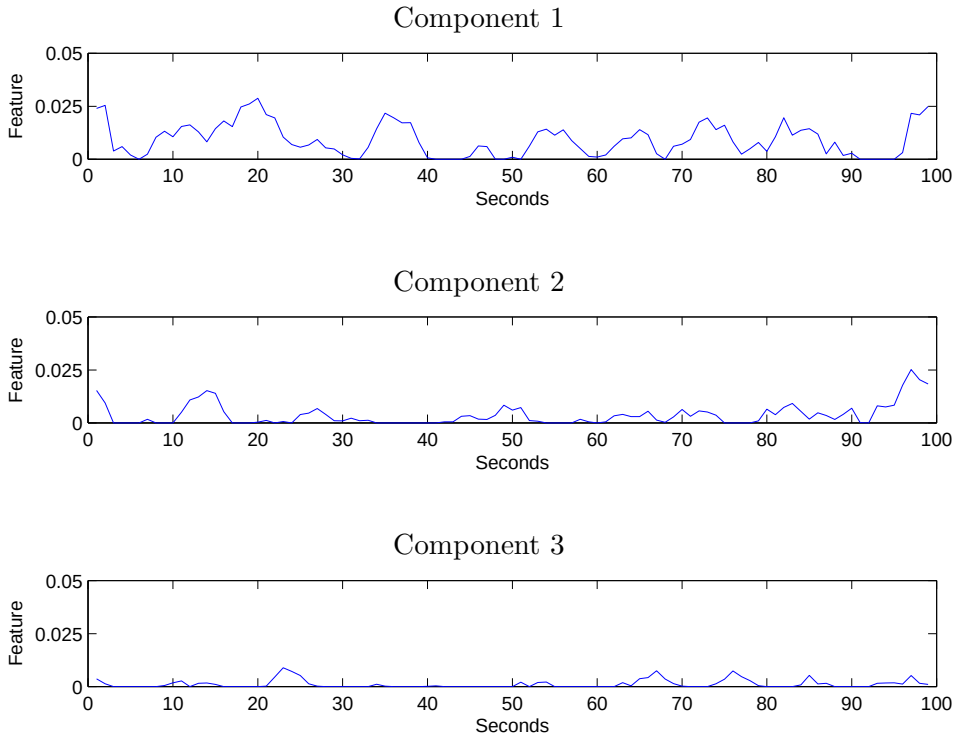


Figure 6.6: Corresponding engagement feature values for each component in Figure 6.5 derived from EEG.

From the weight vector $\mathbf{w}$, we can also compute the scalp projections of the synchronized activity since each weight corresponds to an electrode. This is computed on a per-clip basis across all subjects. We show representative scalp projections for EEG data collected during both audible and silent playbacks of three clips in Figure 6.7. In the audible group, recordings near the visual cortex are given more weight in the first component. This simply implies that subjects are attentively watching the videos. In the second component, it appears that more weight is given in regions near the prefrontal cortex. As this region of the brain is normally associated with executive function and decision making, one possible implication is that subjects are judging the speakers. The scalp projections for the corresponding clips during silent playback are more difficult to interpret as they are more

varied. One possible explanation for this is that in the absence of audio, subjects did not pay attention. This is consistent with feedback from subjects after the experiment, who stated that they were "confused" about what to do during the silent videos clips. After all, the main appeal of a political debate are the words of the candidates. Interpretations for the third component are unclear at this time and we leave that for future work.

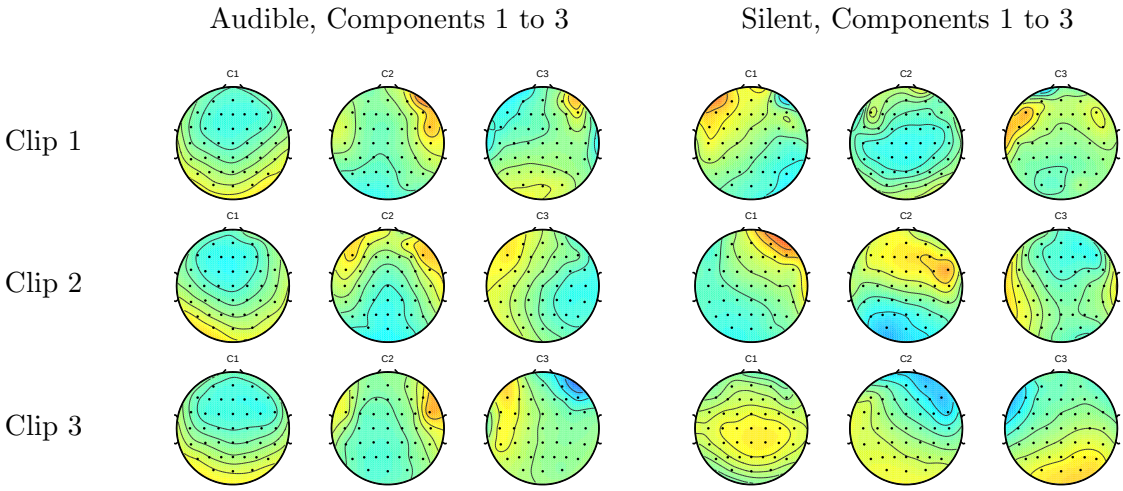


Figure 6.7: Scalp projections showing weights given to EEG channels (electrodes) in order to produce the maximally correlated components. The weights are for all subjects for each of three video clips—one on each row—for each audio mode.

6.2.2 Correlating Gestures Against Engagement

For both of the silent and audible versions of each clip and each one of the five subject groups, we now have three components of engagement features derived from EEG data. We can also compute three types of gesture attributes for the clip: velocity, direction change, extremal pose, each computed according to Section 3.3. All of these features are time-synced. The visual features are computed at 12 Hz, so the engagement features (EEG) which were originally derived at 1 Hz must also be upsampled to 12 Hz by linear interpolation. To summarize, for each of the 28 video clips originally selected, we have the following feature vectors:

- **3 gesture attributes**: velocity, direction change, extremal pose.

- **30 engagement features**: 5 subject groups (all, Democrats, Republicans, females, males) $\times$ 2 audio modes (silent, audible) $\times$ 3 components of EEG-derived engagement features.

We now wish to find the correlations between the gesture attributes and the engagement features. These are usually computed on a per-clip basis, but to compute the correlations, we concatenate the vectors for all clips.

Because we do not hypothesize that there is a direct correlation between the exact magnitude of the gesture attributes and the "level" of engagement (i.e., magnitude of the engagement feature)—that is, there is no reason a slightly more emphatic gesture should result in slightly more engagement—we choose a correlation measure that is less dependent on magnitude. Therefore, instead of Pearson product-moment correlation, we use Spearman rank correlation.

Suppose have two sets of values which we wish to correlate: $x_i \in X, y_i \in Y$ with means $\bar{x}, \bar{y}$, and r_i, s_i are the ranks of x_i, y_i in X, Y respectively. We give the formulas for Pearson correlation r and Spearman correlation ρ in Equations 6.5 and 6.6 respectively. Since Spearman uses ranks and not the exact values, the effects of extremely large or small values on the final correlation is mitigated.

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \tag{6.5}$$

$$\rho = \frac{\sum_{i=1}^{n} (r_i - \bar{r})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{n} (r_i - \bar{r})^2} \sqrt{\sum_{i=1}^{n} (s_i - \bar{s})^2}} \tag{6.6}$$

In addition to stratifying subjects according to political and gender groups, and audio mode (i.e., whether or not the video clip was audible or silent), we can also stratify the videos according to speaker (Obama or Romney), and debate (first or third).

Tables 6.3 to 6.11 show all cases where a statistically significant correlation was found between a visual feature and an EEG component for the various stratifications by speaker, debate and subject group. The corresponding correlation coefficient at $p = 0.05$ (Bonferroni corrected, assuming 30 hypotheses) is also shown in each table. Observations gained from these results are discussed in Section 6.3.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
All	Audible	0.098	Dir Change	1	0.095
	Silent	0.083	Velocity	2	0.073
Democrats	Audible	0.135	Extremal	1	0.129

Table 6.3: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for all videos showing Romney or Obama from both the first and third debate. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
Democrats	Audible	0.077	Velocity	1	0.075
	Silent	0.167	Extremal	1	0.129

Table 6.4: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for all videos showing Romney or Obama from only the first debate. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
All	Audible	0.111	Dir Change	1	0.095
	Silent	0.105	Velocity	2	0.073
		0.103	Dir Change	2	0.09
		0.186	Extremal	1	0.123
Republicans	Audible	0.12	Dir Change	1	0.095
	Silent	0.229	Extremal	2	0.129
Democrats	Audible	0.172	Extremal	2	0.129
	Silent	0.159	Extremal	3	0.129
Females	Audible	0.097	Dir Change	1	0.095
		0.144	Dir Change	2	0.09
	Silent	0.108	Dir Change	2	0.09
		0.165	Extremal	2	0.129
Males	Audible	0.177	Extremal	1	0.134
	Silent	0.189	Extremal	1	0.129
		0.277	Extremal	2	0.129
		0.147	Extremal	3	0.13

Table 6.5: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for all videos showing Romney or Obama from only the third debate. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
All	Audible	0.112	Dir Change	1	0.095
	Silent	0.105	Velocity	2	0.073
Democrats	Silent	0.131	Extremal	2	0.129
Males	Silent	0.08	Velocity	3	0.071

Table 6.6: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for only videos showing Obama from both debates. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
Democrats	Silent	0.136	Extremal	1	0.129

Table 6.7: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for only videos showing Obama from only the first debate. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
All	Audible	0.183	Dir Change	1	0.095
	Silent	0.101	Velocity	2	0.073
		0.131	Dir Change	2	0.09
		0.213	Extremal	1	0.129
		0.139	Extremal	3	0.13
Republicans	Audible	0.153	Dir Change	1	0.095
		0.11	Dir Change	2	0.09
		0.098	Dir Change	3	0.088
	Silent	0.296	Extremal	2	0.129
Democrats	Audible	0.188	Extremal	3	0.129
	Silent	0.112	Dir Change	1	0.095
		0.232	Extremal	1	0.129
		0.295	Extremal	2	0.129
Females	Audible	0.137	Dir Change	1	0.095
		0.186	Dir Change	2	0.09
		0.095	Dir Change	3	0.088
	Silent	0.31	Extremal	2	0.129
		0.164	Extremal	3	0.13
Males	Silent	0.267	Extremal	1	0.129
		0.241	Extremal	2	0.129
		0.21	Extremal	3	0.13

Table 6.8: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for only videos showing Obama from only the third debate. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
All	Audible	0.149	Extremal	1	0.134
	Silent	0.081	Velocity	1	0.075
		0.159	Dir Change	1	0.095
		0.144	Extremal	1	0.129
		0.159	Extremal	2	0.129
Republicans	Audible	0.097	Dir Change	2	0.09
	Silent	0.078	Velocity	1	0.075
		0.104	Dir Change	1	0.095
Democrats	Silent	0.089	Velocity	1	0.075
		0.185	Dir Change	1	0.095
		0.161	Extremal	1	0.129
Females	Silent	0.137	Extremal	2	0.129
Males	Audible	0.136	Extremal	1	0.134
	Silent	0.163	Extremal	2	0.129

Table 6.9: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for only videos showing Romney from both debates. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
All	Silent	0.086	Velocity	1	0.075
		0.18	Dir Change	1	0.095
		0.152	Extremal	1	0.129
		0.137	Extremal	2	0.129
Republicans	Audible	0.116	Dir Change	2	0.09
		0.147	Extremal	1	0.134
	Silent	0.125	Dir Change	1	0.095
Democrats	Silent	0.094	Velocity	1	0.075
		0.218	Dir Change	1	0.095
		0.222	Extremal	1	0.129
Females	Silent	0.145	Extremal	2	0.129
		0.16	Extremal	3	0.13
Males	Audible	0.109	Dir Change	1	0.095
		0.222	Extremal	2	0.129
	Silent	0.075	Velocity	2	0.073
		0.119	Dir Change	2	0.09
		0.193	Extremal	1	0.129

Table 6.10: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for only videos showing Romney from only the first debate. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

Subjects	Audio Mode	ρ	Gesture	EEG Comp	ρ at $p = 0.05$
All	Audible	0.118	Velocity	3	0.077
		0.228	Dir Change	3	0.088
		0.19	Extremal	1	0.134
	Silent	0.107	Velocity	2	0.073
		0.098	Dir Change	1	0.095
Republicans	Silent	0.242	Extremal	2	0.129
Democrats	Audible	0.094	Dir Change	3	0.088
		0.139	Extremal	2	0.129
	Silent	0.128	Dir Change	1	0.095
		0.218	Extremal	3	0.13
Females	Silent	0.085	Velocity	3	0.071
		0.174	Dir Change	2	0.09
Males	Audible	0.178	Velocity	2	0.076
		0.208	Extremal	1	0.134
	Silent	0.297	Extremal	2	0.129

Table 6.11: Statistically significant Spearman rank correlations ρ between gesture attributes and components of engagement features with $p < 0.05$ (Bonferroni corrected) for only videos showing Romney from only the third debate. The threshold for statistical significance, i.e., the correlation coefficient ρ at $p = 0.05$, is given in the last column for the gesture/engagement pairing.

To determine statistical significance, we take an approach similar to [Dmochowski *et al.*, 2012]. We perform a permutation test whereby random samples are created by shuffling non-overlapping 5-second windows of the EEG data and correlating against the visual features to form the null distribution. An example of the null distribution when the extremal pose gesture attribute is correlated against randomly shuffled samples of the first component of engagement features is shown in Figure 6.8.

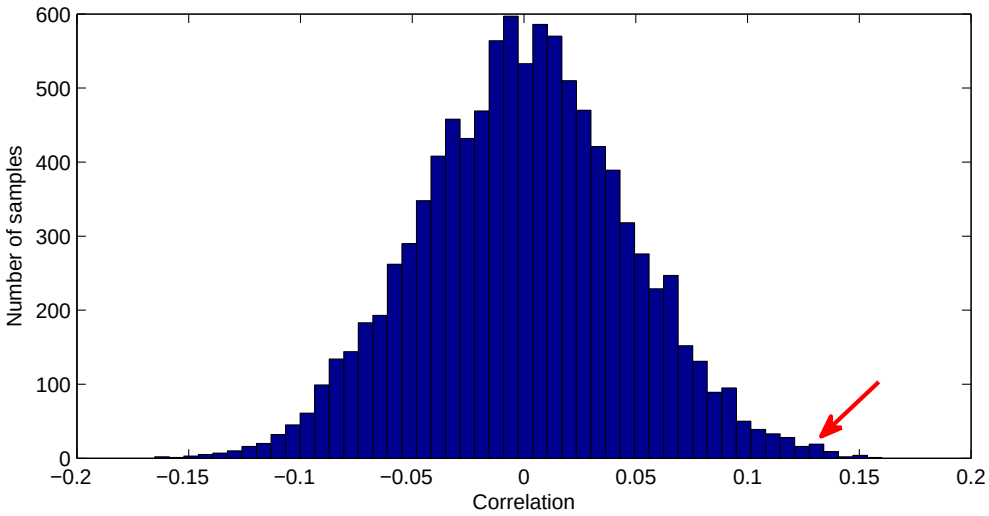


Figure 6.8: Distribution of Spearman correlations of the extremal pose gesture attribute and randomly shuffled engagement features derived from the first component of EEG data from all subjects from permutation test (i.e., the null distribution). The red arrow points to the correlation coefficient where $p = 0.05$, Bonferroni corrected.

6.3 Observations

From these results, we can make the following observations. Despite our best efforts, we could not account for many of the nuances in political science (e.g., some Republicans may actually be Libertarian and thus not extremely supportive of Romney), or the inherent biases in the population from which we selected our subjects, therefore we refrain from making observations regarding differences between groups of subjects.

Gestures are significant and may augment speech. Consistent with the main hypothesis of this thesis, when we examine the data for all subjects and all videos, there exists a statistically significant correlation between gesture attributes (specifically, direction change) (from Table 6.3) and audience engagement at $\rho = 0.098$. Statistically significant correlations occur between gesture attributes and engagement in other stratifications during audible playback as well. This demonstrates that gestures may be used to pique audiences' attention during speech.

Extremal poses and direction change are the most significant gesture attributes we can find. In the majority of the stratifications we examined, the extremal pose gesture attribute resulted in the highest correlations with engagement followed by direction change, as seen in Table 6.12. In general, one result we can gain from this is advice to public speakers: it doesn't matter what kind of gesture you do, but one way to recapture your audience's attention is to "break a pattern", that is, gesture in a way that is different than your norm.

Gesture \ EEG Comp	1	2	3	Total
Velocity	6	7	3	16
Dir Change	18	10	4	32
Extremal	19	18	8	45
Total	43	35	15	

Table 6.12: Frequency at which a gesture/engagement correlation was statistically significant across all stratifications.

The first and second components of the engagement features drives correlation. In the majority of the stratifications we examined, engagement features derived from the first component of EEG data resulted in the strongest correlations, followed by the second component, as seen in Table 6.12. Since the first and second components should correspond to highest and second-highest amounts of variance in the EEG data respectively, this result is consistent. This is also consistent with the results of [Dmochowski *et al.*, 2012].

People don't pay attention to political debates when there is no sound. The original goal of having subjects watch clips without sound was to use those results as a control. However, it appears that without sound, subjects simply stopped paying attention. As previously mentioned in Section 6.2.1 and Figure 6.7, the first component of the scalp projections for subjects watching audible clips shows higher weights around the visual cortex, while the scalp projections for the silent clips appear to be less focused. It is certainly encouraging to see that, for the most part, people watch debates to listen to what the candidates have to say.

Obama did not engage the audience through gesture in the first debate, whereas Romney did so in both debates. This is consistent with bipartisan reports that Obama "lost" the first debate and lacked energy [Landler and Baker, 2012]. As can be seen in Table 6.7, only a single stratification showed statistically significant correlations between gesture and audience engagement. However, Obama recovered in his third debate, with many significant correlations, as seen in Table 6.8. Romney, on the other hand, hand many significant correlations in both the first and third debates (Tables 6.10 and 6.11, respectively).

Chapter 7

Conclusion

Psychologists and linguists have long observed the importance of gestures in communication. We studied this relationship by taking a more quantitative approach and sought to find an application by building tools for automatically extracting gesture attributes and using them as an index for video browsing to aid users in finding interesting segments faster. We examined the relationship between speaker gestures and audience in greater detail by examining specific gesture attributes and correlating against several different measures of audience engagement.

7.1 Contributions

In this thesis we have built an argument for using speaker gestures as an index for semantic video browsers to help audiences locate points of interest in video. We have explored the feasibility of this in the domains of educational lectures and political debates, and introduced computer vision methods and user interface elements for integrating gesture features into semantic video browsers.

We have introduced a number of novel methods for recognizing and extracting poses of interest and gesture attributes from video. We presented a joint-angle descriptor derived from an automatic upper body pose estimation framework to train an SVM in order to classify extracted video frames in the educational lectures domain. Cross validation on the ground-truth data showed classifier F-scores of 0.54 and 0.39 for point and spread poses.

We also extended this work into a gesture attribute with measured arm gesture variance.

We have also presented a method for tracking hands which can distinguish when left and right hands are clasping, and tracks their positions by propagating tracking information from anchor frames in video. The method performs better than baseline on recall (0.66 vs. 0.53) without sacrificing precision (0.65 for both) when recognizing clasping hands. Its tracking efficacy also shows an improvement over baseline (F-score of 0.59 vs. 0.48 baseline). We applied this method toward the extraction of gesture attributes such as velocity, direction change and extremal pose.

We have developed a tool for the manual annotation of gestures, a taxonomy of gestures in lecture videos which introduces a new class of pedagogic gestures which appear to correspond to semantically significant segments of a lecture, and presented a manual analysis of gestures in lecture videos. These inspired later work such as focusing on gesture attributes (e.g., how unusual a gesture is) instead of specific gestures (e.g., is it wave).

Toward the goal of correlating gestures with engagement, we began by attempting to find correlations between gesture attributes and indirect measures of engagement as derived from parts of speech. We demonstrated this by building an AdaBoost-based binary classifier which uses decision trees as weak learners. It classifies videos according to whether its speech content contains conjunctions of interest using the angular variance of arm movements as a feature. We show that training on the set of all conjunctions produces a classifier that performs no better than chance, but that training on sets of conjunctions indicating contrast are capable of achieving 55% accuracy on a balanced test set.

We also performed user studies in order to experiment with interface elements for presenting gesture attribute information, as well as to gather feedback on what types of gestures pique subjects' interest. Subjects indicated that the interface elements we proposed—gesture attribute graph and emphasized subtitles—were helpful for the task of providing keyword summaries under time constraints. Subjects' summary keywords are also compared to an independent ground-truth, resulting in precisions from 0.50–0.55 even when given less than half real time to view the video.

Finally, we built on the results and feedback of earlier work and conclude our argument by correlating gesture attributes extracted from speakers in the domain of political debates

against engagement features derived from EEG recorded from 20 subjects watching clips of the 2012 U.S. Presidential Debates. We identified a gesture attribute, namely extremal pose, which seems to drive correlation against engagement in a majority of cases. We also identified statistically significant correlations between gesture attributes and engagement for all subjects watching all videos with sound (Spearman correlation $\rho = 0.098$ with $p < 0.05$ with Bonferroni correction) and significant correlations for some subgroups both with and without sound, with correlations going as high as 0.297 ($p < 0.05$, Bonferroni corrected). From these results, we conclude the importance of gesture in engaging audiences, and its feasibility as an index for video browsing.

7.2 Future Work

We identify a number of areas for future work, which include the development of better visual features and the identification of other multimedia features which correlate to moments of engagement (such as those determined through EEG). Some specific ideas for possible future work as listed as follows.

1. **Alternate user interface elements.** In Chapter 5, we explored two methods for presenting gesture information in video to viewers. The first, a gesture attribute graph, is fairly straightforward. The second, emphasized subtitles, augments subtitles with gesture attribute information. A plethora of other options exist which remain to be explored. In the most basic sense, a quantized form of the gesture attribute graph (i.e., instead of showing a graph of all values, we first quantize, perhaps binarize, the value) may reveal itself to be more helpful to users by reducing opportunities for confusion. More complex interface options include letting users search for poses or gestures (e.g., search for where an instructor is pointing to the board).

2. **Better gesture representation.** In this thesis, we assumed the multi-phasic gesture model of [Kendon, 1980] which allowed us to represent a gesture by its stroke pose. A more complex representation could capture motions and other temporal motion and lead to other gesture attributes. However, the challenge of segmenting gestures temporally still remains an unsolved problem, due to its open-ended nature. That

is, even human experts cannot agree when a gesture begins and ends. Furthermore, most of our representations here are in 2D (e.g., the poses estimated using [Ferrari *et al.*, 2008]). Being able to represent poses in 3D could also lead to more descriptive attributes. To accomplish this, a 3D dataset collected through tools such as the Microsoft Kinect[1] or the Thalmic Labs Myo[2] would be most helpful.

3. **More gesture attributes.** In this thesis, we closely examined three automatically extracted gesture attributes against audience engagement. Specifically: velocity, direction change and extremal pose. While we found interesting results, particularly with regard to extremal pose, we do not claim that these span the breadth of human gesture. Indeed our attributes were limited to those which could be derived solely from hand positions. Other possible attributes include: hand poses (e.g., the shape of the hand—is it pointing, in a fist, an open palm, etc.), body orientation, and head movements (e.g., head shaking, nodding, etc.).

4. **Study how words co-occur with gesture and engagement.** Some words may naturally co-occur with gesture and catch audiences' attention. We attempted a preliminary study of this but we were limited by the dearth of data, and therefore observations are far from conclusive. We examined temporally-aligned words in the presidential debates and their corresponding segments (within a 5-second window) of Spearman correlations between extremal pose and engagement, and we plotted these values in histograms for words with at least 15 occurrences. The results are shown in Figure 7.1. Clearly the amount of data is insufficient for any conclusions, but do suggest potentially interesting results. Particularly, the words *business* and *(interest) rate*, which are very politically charged, appear to have a skewed distribution of correlation coefficients—and interestingly, both are negative. An interpretation remains for future work.

5. **How other modalities correlate with engagement.** We have collected an interesting dataset of audience engagement as measured through neural activity. Nat-

[1]Microsoft Kinect, *http://www.xbox.com/KINECT*.

[2]Thalmic Labs Myo, *https://www.thalmic.com/myo/*.

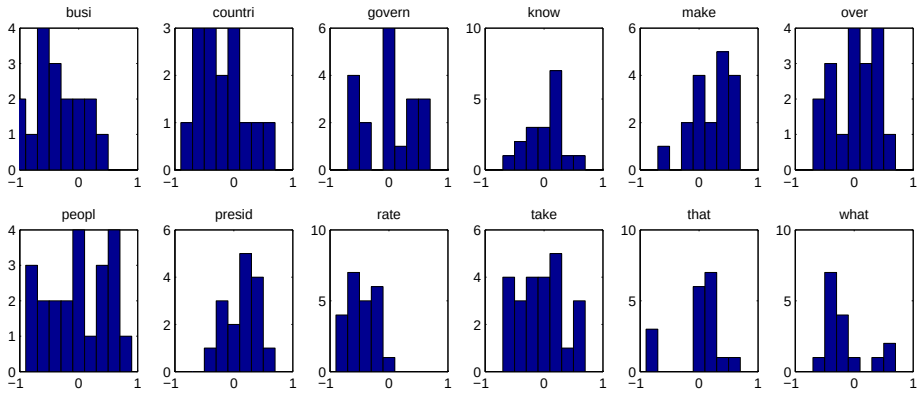


Figure 7.1: Co-occurrences of specific words and speaker gestures and audience engagement. The histograms show the correlation coefficients between extremal pose and engagement features which co-occurred within a 5-second window of a specific word (the word stem is shown).

urally, it would be interesting to see what other features beyond gesture may elicit audiences' attention. Our related work chapter discussed previous work which demonstrated the multi-modal nature of gestures and how speech may affect engagement (e.g., [Eisenstein and Davis, 2006; Kettebekov *et al.*, 2003; Grosz and Sidner, 1986; Watanabe *et al.*, 2007]). Our dataset offers an opportunity to study other features against engagement in a quantitative way. Examining how low-level audio features may correlate with engagement is one obvious direction for future work.

6. **Application to other domains.** For simplicity, we restricted our work to two domains: educational lectures and political debates. It would be interesting to expand this work to other domains, or even the general case. Most of the challenges in doing so lie in computer vision: i.e., how to recognize 3D poses and gestures in people under difficult conditions such as heavy occlusion, poor lighting, high movement, and large variation in appearance.

Bibliography

[Abelson, 2013] B. Abelson. Zombies, brains, and tweets: The neural and emotional correlates of social media. `http://brianabelson.com/assets/zombies_brains_tweets.pdf`, 2013.

[Bavelas *et al.*, 1995] J. Bavelas, N. Chovil, L. Coates, and L. Roe. Gestures specialized for dialogue. *Personality and Social Psychology Bulletin*, 21(4):394–405, 1995.

[Buehler *et al.*, 2008] P. Buehler, M. Everingham, D. P. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. British Machine Vision Conference*, 2008.

[Buehler *et al.*, 2009] P. Buehler, M. Everingham, and A. Zisserman. Learning sign language by watching TV (using weakly aligned subtitles). In *Proc. Computer Vision and Pattern Recognition*, 2009.

[Bull and Connelly, 1985] P. E. Bull and G. Connelly. Body movement and emphasis in speech. *Journal of Nonverbal Behavior*, pages 169–187, 1985.

[Bull, 1986] P. Bull. The use of hand gesture in political speeches: Some case studies. *Journal of Language and Social Psychology*, 5(2):103–118, 1986.

[Buscher *et al.*, 2009] G. Buscher, E. Cutrell, and M. R. Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pages 21–30. ACM, 2009.

[Buscher *et al.*, 2010] G. Buscher, S. T. Dumais, and E. Cutrell. The good, the bad, and the random: an eye-tracking study of ad quality in web search. In *Proc. SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 2010.

[Cao *et al.*, 2011] L. Cao, S-F. Chang, N. Codella, C. Cotton, D. Ellis, L. Gong, M. Hill, G. Hua, J. R. Kender, and M. Merler. Ibm research and columbia university trecvid-2011 multimedia event detection (med) system. In *NIST TRECVID Workshop*, 2011.

[Casasanto and Jasmin, 2010] D. Casasanto and K. Jasmin. Good and bad in the hands of politicians: Spontaneous gestures during positive and negative speech. *PLoS One*, 5(7):e11805, 2010.

[Chang and Lin, 2011] C-C. Chang and C-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Dalal and Triggs, 2005] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 886–893, June 2005.

[Delorme and Makeig, 2004] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[Diakopoulos and Shamma, 2010] N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1195–1198. ACM, 2010.

[Dmochowski *et al.*, 2012] J. P. Dmochowski, P. Sajda, J. Dias, and L. C. Parra. Correlated components of ongoing eeg point to emotionally laden attention–a possible marker of engagement? *Frontiers in Human Neuroscience*, 6, 2012.

[Donadio, 2013] R. Donadio. When italians chat, hands and fingers do the talking. `http://www.nytimes.com/2013/07/01/world/europe/when-italians-chat-hands-and-fingers-do-the-talking.html`, 2013.

[Eisenstein and Davis, 2006] J. Eisenstein and R. Davis. Visual and linguistic information in gesture classification. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Courses*, page 30, New York, NY, USA, 2006. ACM.

[Everingham *et al.*, 2010] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *Proc. International journal of computer vision*, 88(2):303–338, 2010.

[Farnebäck, 2003] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *Image Analysis*, pages 363–370, 2003.

[Felzenszwalb and Huttenlocher, 2000] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Proc. Computer Vision and Pattern Recognition*, volume 2, pages 66–73, 2000.

[Ferrari *et al.*, 2008] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. Computer Vision and Pattern Recognition*, June 2008.

[Fisher, 1935] Ronald Aylmer Fisher. The design of experiments. 1935.

[Fisher, 1996] N. I. Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, 1996.

[Freund and Schapire, 1996] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proc. International Conference on Machine Learning*, 1996.

[Gomez and Morales, 2002] G. Gomez and E. Morales. Automatic feature construction and a simple rule induction algorithm for skin detection. In *Proc. ICML workshop on Machine Learning in Computer Vision*, pages 31–38, 2002.

[Grosz and Sidner, 1986] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

[Gut *et al.*, 1993] U. Gut, K. Looks, A. Thies, T. Trippel, and D. Gibbon. Cogest conversational gesture transcription system. Technical report, University of Bielefeld, 1993.

[Hanson *et al.*, 2009] S. J. Hanson, A. D. Gagliardi, and C. Hanson. Solving the brain synchrony eigenvalue problem: conservation of temporal dynamics (fmri) over subjects doing the same task. *Journal of computational neuroscience*, 27(1):103–114, 2009.

[Hasson *et al.*, 2004] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, and R. Malach. Intersubject synchronization of cortical activity during natural vision. *science*, 303(5664):1634–1640, 2004.

[Haubold and Kender, 2007] A. Haubold and J. R. Kender. Vast mm: multimedia browser for presentation video. In *Proc. international conference on Image and video retrieval*, pages 41–48, New York, NY, USA, 2007. ACM.

[Jones, 2012] J. M. Jones. Record-high 40% of americans identify as independents in '11. `http://www.gallup.com/poll/151943/record-high-americans-identify-independents.aspx`, 2012.

[Kadir *et al.*, 2004] T. Kadir, R. Bowden, E.J. Ong, and A. Zisserman. Minimal training, large lexicon, unconstrained sign language recognition. In *Proc. British Machine Vision Conference*, volume 1, 2004.

[Kendon, 1980] A. Kendon. Gesticulation and speech: Two aspects of the process of utterance. In *The relationship of verbal and nonverbal communication*, pages 207–227. Mouton Publishers, 1980.

[Kettebekov *et al.*, 2003] S. Kettebekov, M. Yeasin, and R. Sharma. Improving continuous gesture recognition with spoken prosody. In *Proc. Computer Vision and Pattern Recognition*, volume 1, pages 565–570, June 2003.

[Kipp *et al.*, 2007] M. Kipp, M. Neff, and I. Albrecht. An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation*, 41(3-4):325–339, 2007.

[Kipp, 2001] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proc. EUROSPEECH-2001*, pages 1367–1370, 2001.

[Kolsch and Turk, 2004] M. Kolsch and M. Turk. Robust hand detection. In *Proc. Automatic Face and Gesture Recognition*, pages 614–619, May 2004.

[Landler and Baker, 2012] M. Landler and P. Baker. After debate, obama team tries to regain its footing. `http://www.nytimes.com/2012/10/05/us/politics/obama-team-tries-to-change-course-after-debate-disappoints.html`, 2012.

[Ma *et al.*, 2002] Y-F. Ma, L. Lu, H-J. Zhang, and M. Li. A user attention model for video summarization. In *Proc. Multimedia*, pages 533–542. ACM, 2002.

[Martell, 2002] C. Martell. Form: An extensible, kinematically-based gesture annotation scheme. In *Proc. International Conference on Language Resources and Evaluation*, 2002.

[McNeill, 1992] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University Of Chicago Press, 1992.

[Merler and Kender, 2009] M. Merler and J. R. Kender. Semantic keyword extraction via adaptive text binarization of unstructured unsourced video. In *Proc. International Conference on Image Processing*, pages 261–264, November 2009.

[Mittal *et al.*, 2011] A. Mittal, A. Zisserman, and P. Torr. Hand detection using multiple proposals. In *Proc. British Machine Vision Conference*, 2011.

[Morris and Kender, 2011] M. J. Morris and J. R. Kender. Vastmm-tag: a semantic tagging browser for unstructured videos. In *Proc. International Conference on Multimedia*. ACM, 2011.

[Nakano and Ishii, 2010] Y. I. Nakano and R. Ishii. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proc. international conference on Intelligent user interfaces*, pages 139–148. ACM, 2010.

[Niebles *et al.*, 2008] J. Niebles, B. Han, A. Ferencz, and L. Fei-Fei. Extracting moving people from internet videos. In *Proc. European Conference on Computer Vision*, 2008.

[Pozzer-Ardenghi and Roth, 2004] L. Pozzer-Ardenghi and W. Roth. Photographs in lectures: Gestures as meaning-making resources. *Linguistics and Education*, 15(3):275–293, 2004.

[Ramanan, 2007] D. Ramanan. Learning to parse images of articulated bodies. *Advances in neural information processing systems*, 19:1129, 2007.

[Revaud *et al.*, 2013] J. Revaud, M. Douze, C. Schmid, and H. Jégou. Event retrieval in large video collections with circulant temporal encoding. In *Proc. Computer Vision and Pattern Recognition*, 2013.

[Roth and Bowen, 1998] W. Roth and G. Bowen. Decalages in talk and gesture: Visual and verbal semiotics of ecology lectures. *Linguistics and Education*, 10(3):335–358, 1998.

[Roth and Lawless, 2002] W. Roth and D. Lawless. When up is down and down is up: Body orientation, proximity, and gestures as resources. *Language in Society*, 31(01):1–28, 2002.

[Roth, 2001] W. Roth. Gestures: Their role in teaching and learning. *Review of Educational Research*, 71(3):365–392, 2001.

[Scharton and Neuleib, 2001] M. Scharton and J. Neuleib. *Things your grammar never told you.* Longman, 2nd edition, 2001.

[Smeaton *et al.*, 2010] A.F. Smeaton, P. Over, and A.R. Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, 114(4):411–418, 2010.

[Strayer *et al.*, 2011] D. L. Strayer, J. M. Watson, and F. A. Drews. Cognitive distraction while multitasking in the automobile. *Psychology of Learning and Motivation-Advances in Research and Theory*, 54:29, 2011.

[Tavernise, 2013] S. Tavernise. Brain test to diagnose a.d.h.d. is approved. `http://www.nytimes.com/2013/07/16/health/brain-test-to-diagnose-adhd-is-approved.html`, 2013.

[Viola and Jones, 2002] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2002.

[Watanabe *et al.*, 2007] M. Watanabe, Y. Den, K. Hirose, S. Miwa, and N. Minematsu. Features of pauses and conjunctions at syntactic and discourse boundaries in japanese monologues. In *INTERSPEECH*, pages 118–121, 2007.

[Wilson *et al.*, 1997] A. Wilson, A. Bobick, and J. Cassell. Temporal classification of natural gesture and application to video coding. In *Proc. Computer Vision and Pattern Recognition*, page 948, 1997.

[Xaquin *et al.*, 2012] G. V. Xaquin, A. McLean, A. Tse, and S. Peçanha. What romney and obamas body language says to voters. `http://www.nytimes.com/interactive/2012/10/02/us/politics/what-romney-and-obamas-body-language-says-to-voters.html`, 2012.

[Xiao *et al.*, 2010] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proc. Computer Vision and Pattern Recognition*. IEEE, 2010.

[Yang and Ramanan, 2011] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. Computer Vision and Pattern Recognition*. IEEE, 2011.

[Yao and Cooperstock, 2002] J. Yao and J. R. Cooperstock. Arm gesture detection in a classroom environment. In *Proc. Workshop on Applications of Computer Vision*, pages 153–157, 2002.

[Zhang and Kender, 2011] J. R. Zhang and J. R. Kender. Identifying salient poses in lecture videos. In *Proc. International Conference on Image Processing*, Sept 2011.

[Zhang and Kender, 2012] J. R. Zhang and J. R. Kender. Arm gesture variations during presentations are correlated with conjunctions indicating contrast. In *Proc. ACM workshop on User experience in e-learning and augmented technologies in education*, pages 13–18. ACM, 2012.

[Zhang and Kender, 2013] J. R. Zhang and J. R. Kender. Recognizing and tracking clasping and occluded hands. In *Proc. International Conference on Image Processing*, Sept 2013.

[Zhang *et al.*, 2010] J. R. Zhang, K. Guo, C. Herwana, and J. R. Kender. Annotation and taxonomy of gestures in lecture videos. In *Proc. CVPR Workshop on Human Communicative Behavior Analysis*, June 2010.

[Zhang *et al.*, 2013] J. R. Zhang, J. R. Kender, and X. Ma. Dramatic speaker gestures as indicators of segments of interest for video browsing. In *preparation*, 2013.