FORM: AN EXPERIMENT IN THE ANNOTATION OF THE KINEMATICS OF GESTURE

Craig Martell

A DISSERTATION

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2005

Mitchell P. Marcus
Supervisor of Dissertation

Rajeev Alur
Graduate Group Chairperson

Acknowledgements

In many ways, writing the *Acknowledgments* section of a dissertation is almost harder than the work itself. So many people contribute in so many ways that it is usually hard to keep track of it all. This is not true in my case; the debts I owe to others for helping me bring this thesis to completion are very clear to me. There were those in the beginning who helped me formulate and shape my ideas; there were those in the middle who helped me design FORM and who performed the Herculean task of gathering and annotating the data that is the meat of what follows; and, finally, there were those who served to sharpen the writing and presentation of the work to make it much better than it was. To all of these, let me say that my thanks here does justice neither to your contribution nor to the depth of my appreciation. However, I hope it will serve. This thesis would have not have been the same without you. All of what is good in it has your stamp on it, and that which is not so good remains because I simply did not listen well enough.

Firstly, I would like to acknowledge my parents. In very different ways, they both served to make me the curious person I am today.

And to my wife, I would like to say thank you for still finding that curiosity worth while—even after all of this.

From the beginning, the members of my dissertation committee have served as mentors, sounding boards, sponsors, and sanity checks. I would like to thank my advisor, Mitch Marcus, for his friendship, guidance, intellectual conversation, and the latitude to explore the things that truly interested me. His support of my intellectual wanderings—and his pulling me back in when I wandered too far—made for a very positive (and very rare) graduate experience. I would like to thank Mark Liberman for showing me, in numerous ways, that one can be a scientist about things that most believe are too fuzzy. I would like to thank Martha Palmer for sharing my goal of understanding all aspects of communicative interaction, for saving the day with funding when needed, and not letting my behavior get in the way of her doing her job well. I would like to thank Brian MacWhinney not only for giving of himself—almost literally—in the way of data but also for his understanding why building the data was so important to me in the first place. I hope my work has a fraction of the influence his has had on the field. Last, and certainly not least, I would like to thank Norm Badler. This work would be far less interesting if he had not stuck to his guns while allowing me to stick to mine.

Building the data for this thesis could never have been done alone. The FORM team worked long and hard, and all of them deserve my deepest appreciation. So, to Chris Osborn, Joshua Kroll, Paul Howard, Kari Myers, Lisa Britt, Jesse Friedman, Alex McCauley, Ryan Johnson, Danielle Dylinski, Mike Weston, Wendy Laun, Anshuman Srivastava, Narayanan Mahesh, Maia Jachimowicz and Oleksiy Syrotkin, I thank you all for your dedication to getting it right—even when it was driving you nuts.

Additional thanks go to Chris Osborn and Joshua Kroll for being interested, for toughing it out, for talking it through, and for being there.

Thank you to Mark Liberman, Steven Bird, Chris Cierri, Xiaoyi Ma, Stephanie Strassel and all the other people at the Linguistic Data Consortium for the funding, space, and putting up with the hassles we caused.

Special thanks are due to Adam Kendon. His research, his time, his advice and his training of the annotators were crucial to getting this work done.

Throughout my time at Penn, I was privileged to be able to interact with some of the brightest people in their fields. Just having them all around to chat with in the hallways, or over dinner, contributed to my thinking and my work. In particular, though, I would like to thank Andrew Schein, Minsu Kang, Tom Schoenemann, Robin Clark, and Dan Bikel. Thanks for all, guys!

This work was supported in part by the International Standards in Language Engineering project, funded by the National Science Foundation and the European Commission (ISLE-9910603); the Talkbank Project, funded by the National Science Foundation (BCS-9980009 and BCS-9978056); and the National Science Foundation Discourse and Gesture Resources Grant (EIA98-09209). This support was crucial and is very much appreciated.

ABSTRACT

FORM: AN EXPERIMENT IN THE ANNOTATION OF THE KINEMATICS OF GESTURE

Craig Martell

Mitchell P. Marcus

The most obvious way that humans communicate is through speech, and, as such, there has been a great deal of work in Linguistics, Logic, and Computer Science aimed at understanding, formalizing, and automatically generating and analyzing all aspects of human speech. However, speech is not the only means of communication available to us; we are able to send complex and subtle messages to each other via a variety of other means as well.

Gesture is another important channel for conveying intent and meaning. However, unlike the current state of research on speech, gesture research has only very coarse-level categorizations covering the types of gestures and very little in the way of fine-grained techniques for analysis. The current state of the science has gestures divided into essentially four broad categories—beat, iconic, metaphoric, and deictic—and has each gesture decomposable into only four types of constituent phases: preparation, stroke, hold and retraction.

We accept this state of the field as our starting point. That is, we accept that there are at least some gestures that are classifiable as described above and that these gestures may be able to be broken down into their constituent phases. However, a coding scheme that only labels the gesture or phase as a whole runs the risk of missing important variations in meaning created by subtle changes in the components of a gesture in question. Slight differences in the make-up of a beat gesture, for example, may well express very different things concerning the mood or intention of the speaker.

Accordingly, we have developed a fine-grained, gesture coding scheme—FORM—that allows annotators to exhaustively capture the constituent parts of the gestures of video-recorded speakers. In this thesis we present the FORM annotation scheme, inter-annotator-agreement studies, and the results of some hidden-Markov-model experiments using FORM. Additionally, we compare FORM to more accepted methods of automatic data gathering.

The ultimate goal of this project is to develop something like a “phonetics” of gesture that will be useful for both building better HCI systems and doing fundamental scientific research into the communicative process.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The most obvious way that humans communicate is through speech, and, as such, there has been a great deal of work in Linguistics, Logic, and Computer Science aimed at understanding, formalizing, and automatically generating and analyzing all aspects of human speech[1]. However, speech is not the only means of communication available to us; we are able to send complex and subtle messages to each other via a variety of other means as well.

1.1 Facial Expression and FACS

Facial expression, for example, allows us to display joy at seeing someone, disgust for the food we just sampled, or the fear we feel at the movies as we imagine ourselves in the place of the protagonist in a dangerous situation. Examples such as these might be part of the catalog of any freshman psychology student when asked to give examples of how we communicate via facial expression; and, they may indeed be excellent examples. However, they only paint a very coarse-level picture.

[1] See [17], [18], [81], [82],[1], [70], [98], [50], and [69] as good starting points.

Research has demonstrated that facial-expression communication is very subtle and fine-grained. Small changes in a few muscles of the face can result in very different messages being sent. Paul Ekman and Wallace Friesen, in their research on the relationship between facial expression and deception [29] theorized that involuntary emotional expression might “leak” through and be detectable in a facial expression despite a conscious, voluntary attempt to mask it. They found that the clues used by the would-be deceived are the more subtle aspects of the facial expression, what they called *microexpressions*.

The important aspect of their research for our purposes here was the methodology they developed for studying these microexpressions. They found that none of the then existing techniques for measuring facial expression “would succeed in discriminating between the smiles of actual enjoyment and the smiles of feigned enjoyment masking negative emotion” [31]. Their solution was to develop the Facial Action Coding System (FACS). FACS ([30]) is a fine-grained facial measurement technique that allowed them to capture the subtle differences among the different types of smiles in question. In the years since, FACS has been instrumental in furthering scientific understanding of facial expression and has recently been used as the foundation for computational analysis and synthesis of facial expression as well[2].

1.2 Gesture and FORM

Gesture is another important channel for conveying intent and meaning. And, just as with facial expression, a freshman psychology student would give any number of clear cut examples of gestures that are either part of our culture’s gesture vocabulary—e.g., the two-fingered peace sign or the one-fingered curse—or have a clear emblematic

[2] Cf. [36], [51], [86], [87]

meaning given the context of their use—e.g., an invitation to a hug or asking where the phone is across a crowded room. However, unlike the current state of research on facial expression, the field of gesture studies has only very coarse-level categorizations covering the types of gestures and very little in the way of fine-grained description techniques.[3] The current state of the science[4] has gestures divided into only four broad categories—beat, iconic, metaphoric, and deictic—and has each gesture decomposable into its constituent phases. The phases are essentially of only four types: preparation, stroke, hold and retraction[5]. However, the need for a more fine-grained system is well understood. In [106], Wittenburg et al., when describing the choices they made while designing their annotation scheme, state that "it was soon perceived that an exhaustive gesture encoding including all relevant characteristics would be ideal but impossible (except for small segments)."

We accept this state of the field as our starting point. That is, we accept that there are at least some gestures that are classifiable as alluded to above and that these gestures may be able to be broken down into their constituent phases. However, a coding (or annotation) scheme that only labels the gesture as a whole runs the risk of missing important variations in meaning created by subtle changes in the components of a gesture in question. Similar to facial expression, slight differences in the make-up of a beat gesture, for example, may well express very different things concerning the mood or intention of the speaker.

Accordingly, we have developed a fine-grained, gesture coding scheme—FORM—that allows annotators to exhaustively[6] capture the constituent parts of the gestures

[3] As we will see below, Adam Kendon in [60] laid the ground work for a fine-grained, descriptive notation scheme, but it has not been widely used.

[4] The categorization scheme given in [73] is the most widely accepted.

[5] Cf. [25], [26], [52], [57], and [73]

[6] The pun is intended. Not only do we attempt to encode all of the salient aspects of gesture, but it is also the case that the act of doing so is often exhausting. Future work will need to address this very real issue.

of video-recorded speakers. In this document we present the FORM annotation scheme, its underlying data representation, inter-annotator-agreement studies, and the results of some machine-learning experiments using the FORM dataset. Additionally, we compare FORM to the currently most accepted method of data gathering, motion capture. Note that the ultimate goal of the project is to develop an easy-to-use, human-based annotation system that is analogous to FACS in that it will be beneficial to both science and engineering. However, just as FACS took time to develop to maturity, FORM is still in its early stages. Our goal for this thesis is simply to lay a foundation and to point in the direction of what we hope will be future successes.

Our contribution, then, consists of:

1. the FORM Annotation Scheme;

2. inter-annotator agreement studies for FORM;

3. a dataset build using FORM; and

4. a set of experiments demonstrating the usefulness of FORM.

1.2.1 FORM and Science

In order to make important, fundamental scientific discoveries—as well as less-profound, incremental discoveries—large amounts of data are needed. Until the advent of easily-usable and widely-accessible database technology, it had been difficult to gather the data necessary to do fine-grained analyses of communicative interaction. Additionally, these data could not have been gathered without a scheme to represent the salient aspects of the phenomena being studied. The aforementioned

FACS is an excellent example of a coding system that allowed for the gathering of scientific data. These facial-expression data made possible the discoveries concerning deception described above. Another very important example of how strategically-gathered data facilitates science is the CHILDES database ([16]). It has become the standard for any experiments concerning parent/child communicative interaction[7].

From a scientific perspective then, our aim is to develop a sufficiently portable tool that will allow a researcher to gather the fine-grained, three-dimensional details of the gestures of subjects from two-dimensional video. Developing such a tool is important to doing research in a natural, non-laboratory setting. Other methods of gathering three-dimensional information concerning a gesturer, for example video or motion capture, require the subject's being in an artificial environment and/or require wire or other sensors to be attached to the subject's body. Additionally, the ability to gather data simply from two-dimensional video means that there is a wealth of pre-existing data—from old news real footage and *The Honeymooners* through today—that can be used to produce extremely large data sets, under many varying conditions.

1.2.2 FORM and Engineering

Communicative interaction resides in a very high-dimensional space (e.g, words, prosody, components of gesture, components of facial-expression, etc). Even if we were able to give a complete formal description of all the dimensions, computing the intent of the speaker may well be intractable. However, humans clearly communicate and, for the most part, do so unambiguously. Given this, there must be some subspace of the communication domain that contains the salient dimensions necessary to convey meaning. Unfortunately, we are still in the dark as to just what

[7] A Google search for "CHILDES psychology" will turn up an abundance of examples here.

the salient dimensions are.

One method of dealing with this is to use preexisting intent-recognition machines, i.e., humans, to describe and classify the aspects of communicative interaction we wish to study. (That is, we use humans to create annotated corpora.) It is an empirical question as to whether or not the classification space used by human classifiers is adequate to a particular task, but successes using annotated corpora in natural-language processing[8] are encouraging for other domains as well[9].

From an engineering perspective then, our goal is for FORM to facilitate the annotation of a large-scale, extensible dataset that can be used to train corpus-based, statistical algorithms to analyze gesture. Extending the data used to train these algorithms to include other aspects of the communicative act, speech and prosody, for example, will allow for a better understanding of the relationships among all of these facets and, therefore, allow us to build tools with increasing amounts of domain knowledge.

1.3 Our Approach

FORM comprises a tag set with annotation practices, a set of corpora, and an underlying data structure which allows the data to be easily aligned with other aspects of speech that share the same time line. Each of these aspects will be discussed in later chapters, but we will briefly touch on them in the following section. We should note first, however, the contribution of the present work is in the tag set with annotation practices, the corpora, and some experimental verification of the

[8] Examples include [70], [37], [69], [12], [38], and [94]

[9] A cautionary remark is necessary here. We should not underestimate how the fact that the natural language phenomena modeled by these tools are discrete may have contributed to their success. Gesture may simply be too continuous. Understanding this is part of our plan for future research

technique. The underlying data structure, the annotation graph, was developed by Steven Bird and Mark Liberman and presented in [6]. We should also note here that this research lies squarely within the tradition of using annotated corpora to do linguistic research. A fair characterization of the work presented in this thesis is that it is an extension of these corpus-based techniques, developed originally for natural-language processing, to gesture. Along with this extension, however, come the following presuppositions.

- There is sufficient regularity in gesturing and that the corpus represents a sufficient sampling of the gesture-space.
- The information encoded in the annotation scheme used to create the corpus captures this regularity.

The results of this thesis, to a large degree, bear out these presuppositions. However much more research is needed to see if this, indeed, is the best approach. Gesture may simply not be sufficiently like speech to warrant these presuppositions.

1.3.1 The FORM Tag Set

Adam Kendon, in [58], nicely lays out the need for a program to conduct research on the low-level movement aspects of gesture and their relationship to other parts of the communicative act, in particular, speech.

> Such a programme of work could be linked to, and would contribute importantly, to research on what might be called the 'kinetics' of gesture (in parallel to 'phonetics'). We really have little explicit knowledge about how gestures are organized as physical actions . . . [and the work we have represents] only the merest hints[10].

[10] It is important to note here that Kendon's use of "kinetics," below, was to draw the parallel

Kendon presents a candidate annotation scheme aimed at facilitating this research in [60]. There are, however, two important limitations to his scheme. Firstly, it is very difficult for humans to read and write. Secondly, as it was designed to be written by hand or by word processor, there is no way to precisely time-align the speech transcription and the gesture annotation.

FORM began as an almost direct translation from Kendon's original scheme to an XML-based tag set that is more human readable and more computationally useful. We made it more human readable by choosing descriptive variable names and values. We made it more computationally useful by embedding it within the Java-based tool Anvil ([61]), which allows for frame-by-frame alignment of any set of tags with a given video sample. Figure 1.1 shows an annotation created using FORM in Anvil. It is the annotation of the movement represented by the stills in Figure 1.2.

As we progressed, we discovered a number of ambiguity problems resulting from the way we encoded upper-arm location. Essentially, given the number of degrees of freedom of the shoulder, there is no way to uniquely specify a normal position for the upper-arm that remains consistent as the shoulder joint moves. We solve this problem by describing the position of the wrist in a 5 × 5 × 5 grid centered around the speaker's solar plexus. This allowed us to better specify the location of the arm in space. This prompts the question of why we did not simply use this grid, plus whatever else we may need to uniquely specify the arm positions, instead of

with phonetics—not to imply the technical meaning of the word. Technically speaking the *kinetics* of movement are given by measurements of forces, moments of forces, etc, while the *kinematics* of movement are given by measurements of linear position, angular position, velocity, acceleration, and the like in Euclidean space. The important point is simply that we have a useful low-level description of the movements of gesture. There is also a trend in the gesture-research community to sometimes use *kinetic* to mean *kinematic*. When this comes up in our discussions we will make it clear by either using quotation marks or by using *[sic]* if it occurs within a quotation. For our part though, we will mostly use *kinematics*, as FORM measures movement in Euclidean terms.

including all of the complex parameters given by Kendon. These include things like forearm rotation, upper-arm lift, movement descriptions, hand and wrist positions, etc. The answer is that we want FORM to be able to both uniquely *and* richly describe the gestures of a subject. Again, one of the goals of this work is to develop a 'kinetics' of gesture. Doing so requires both understanding and capturing all (or some large subset) of these parameters. Without these, we may not be able to develop explanatory theories about how each parameter contributes to the overall meaning and force of the gesture. This, indeed, is the long-term goal of the research for which we are laying the foundation here[11]

1.3.2 The FORM Code Book

When using human annotators to develop a large-scale corpus[12], it is imperative that the annotators produce both consistent and accurate data. The way to best ensure this is to settle on a set of *annotation practices* which have both been checked against all pairs of annotators, and against the phenomena the data are meant to model. The first check contributes to the data's internal consistency and the second contributes to its external accuracy.

For consistency, there are a number of tests—depending on the type of data—that can be used to measure *inter-annotator agreement.* In Chapter 4, we present one of these, the bag-of-arcs measure, and the results of inter-annotator agreement studies conducted using it.

Unfortunately, external accuracy—that is, how well the data reflects reality—is not always directly measurable. We can, however, check our annotations against

[11] Here we need another cautionary note. It is not clear the necessary dynamic aspects of gesture can be captured by our very coarse-grained annotation system. Future work should include studying what level of granularity will capture this information, if any.

[12] Cf. [70].

the original phenomena and make sure that they are done according to the model. This model is encapsulated in a set of *annotation practices* for FORM. Appendix A presents these. The transcript of the video training manual we used to train annotators on these practices, along with representative stills, is given in Appendix B. Additionally, in Chapter 6, we further check FORM's representation against the widely-accepted "gold standard", viz. motion capture.

1.4 Research Using FORM

As mentioned above, the goal of the FORM project is to develop a foundation for future research using fine-grained, time-aligned gestural data. In order to demonstrate the potential usefulness of the FORM method, then, some experiments using the FORM dataset must be conducted. We briefly describe, here, the experiments conducted and provide full results in Chapters 5 and 6.

As mentioned above, the current state of gestural theory has gestures divided into roughly four classes–Beats, Iconics, Metaphorics and Deictics—and each gesture subdivided into Preparation, Stroke, Retraction and Hold. Accepting this as a starting point, then, our goal is to work from the bottom up, so to speak. That is, if gestures are made up of phases and phases are made up of sequences of movements, we propose—as our initial experiments with FORM—to utilize these sequences of movements, given as FORM descriptions, to predict the phases that comprise the movements. In order to do this, though, we first added an additional annotation layer to the FORM corpus: Phase. Here, the gesture excursion was segmented into *preparations*, *strokes*, *retractions*, and *holds*. In order to judge the consistency of the new annotations, we conducted new inter-annotator agreement studies. These are presented Chapter 5.

Interestingly, although this was done simply as a way to add higher-order labels to the data, it also served as an experiment concerning the validation of PSR theory (preparation-stroke-retraction theory) itself. For, if PSR did not have at least some cognitive validity, we would expect the inter-annotator agreement results to be poor. As we will see though, they were remarkably high. It is important to note that our success here is a qualified one. There were many segments of movement about which the individual annotators could not make a judgment. Instead of insisting that our annotators force all segments into one of the four categories above, we added a fifth, Unsure. As we will see, Unsure and Hold were often confused in the inter-annotator agreement studies

1.4.1 Classification of Gesture Phases using HMMs à la Speech Recognition

Our goal for these experiments is to predict a mid-level phenomenon (*phase*) from the low-level, underlying FORM representation. The results given in Chapter 5 show that PSR-theoretic annotations can be consistently applied. So, we next divided these phases into five bins, one each for *preparations*, *strokes*, *retractions*, *holds* and *unsures*. For all of the experiments in this series we then used the following general algorithm[13]:

1. For each of the bins described above, train an HMM using the segments in the bin;
2. For each item i in the test set, compute $P(i|M)$ for each M in the set of HMMs

[13] This algorithm is similar to a large component of the way speech-recognition systems recognize words. This is the reason for the title of this section. We note here that speech systems are far more complex than this, incorporating a great deal of domain knowledge and a language model. In the future, we hope to have as much domain knowledge and a useful gesture model as well. However, we must note that gesture is highly-dimensional, while speech here is 1-dimensional. It is an open question as to whether or not such a model will be useful, and if so, how.

created above;

3. Label the test item after the M which maximizes $P(i|M)$.

As we will see in Chapter 6, we conducted a large number of experiments, modifying both the way in which the sequence data (which makes up a phase) was created and whether or not context from surrounding phases was included. The results were very promising.

1.5 Outline of Thesis

The rest of this document is structured as follows:

- Chapter 2 discusses other work related to the FORM project. This includes discussions of the annotation of facial expression, the current state of gestural theory as it applies to annotation, and a brief discussion of alternatives that were not used in developing FORM. Here, we will also discuss FORM's relationship to other ways of gathering three-dimensional kinematic information, including motion capture and video capture.

- Chapter 3 presents the complete FORM annotation scheme.

- Chapter 4 presents the results of the general inter-annotator agreement studies for FORM.

- Chapter 5 presents the results of our phase-annotation inter-annotator agreement studies described in section 1.4.

- Chapter 6 presents the design and results of the HMM phase experiments described in section 1.4.1.

- Chapter 7, the conclusion, restates our results and gives suggestions for further research.

- Appendix A presents the best practices discovered by the FORM annotators.

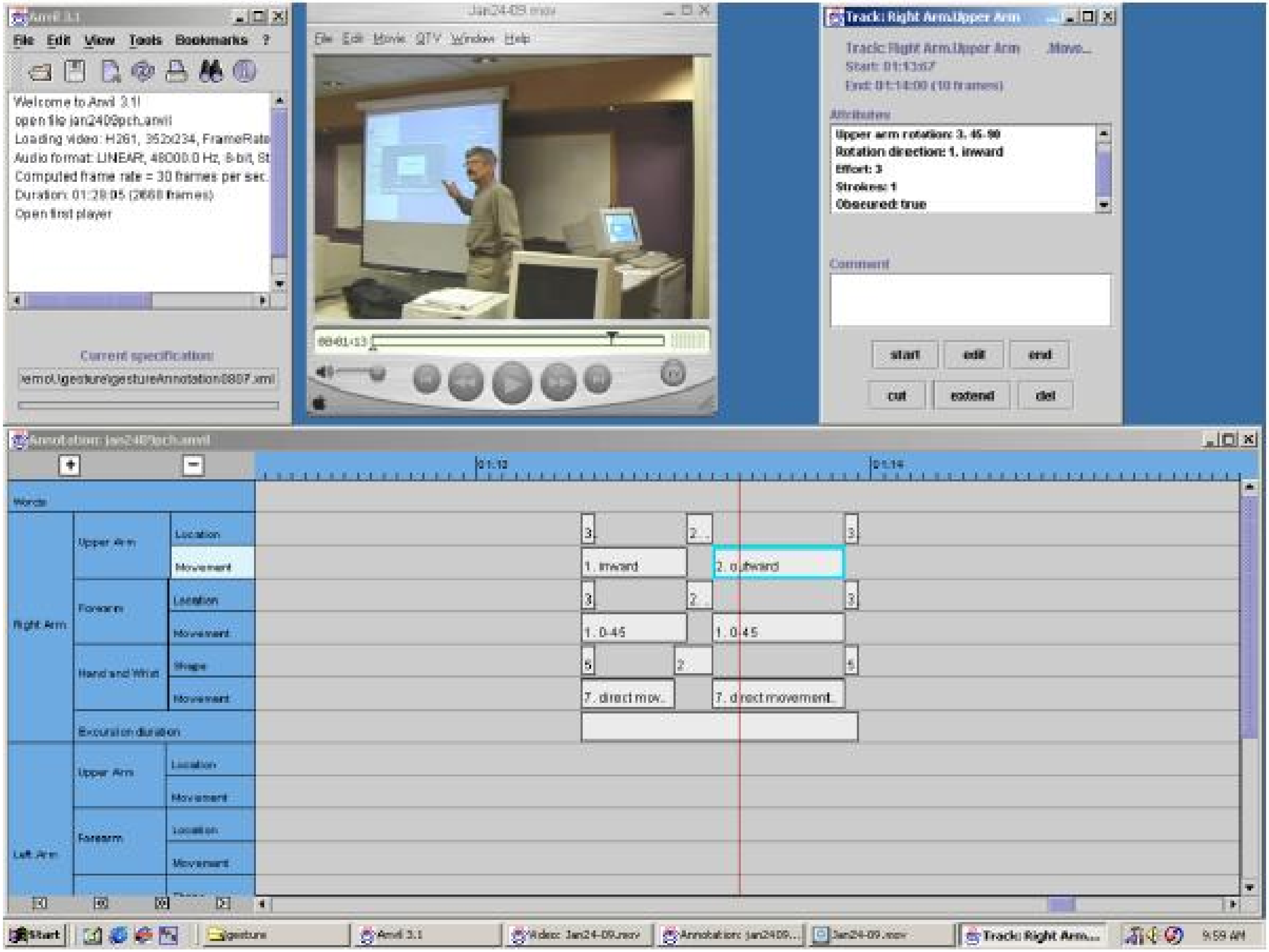
- Appendix B presents the transcripts of the FORM *Video Code Book*—our training manual—along with representative still.

- Appendix C presents the inter-annotator agreement code used for the bag-of-arcs experiments in Chapter 4.

- Appendix D presents the confusion matrices for two annotators for all FORM parameters.

- Appendix E presents the results of a series of McNemar's tests. These allow us to judge how significant the differences in results among the methods given in Chapter 6 are.

- Appendix F presents a review of the mathematics of hidden Markov models.



Welcome to Anvil 3.1!
open file jan2409pch.anvil
Loading video: H261, 352x234, FrameRate
Audio format: LINEAR, 48000.0 Hz, 8-bit, St
Computed frame rate = 30 frames per sec.
Duration: 01:29:05 (2668 frames)
Open first player

Current specification:
remo\.\gesture\gestureAnnotation0807.xml

Track: Right Arm.Upper Arm Move...
Start: 01:13:67
End: 01:14:00 (10 frames)

Attributes
Upper arm rotation: 3. 45-90
Rotation direction: 1. inward
Effort: 3
Strokes: 1
Obscured: true

Comment

Track						
Right Arm	Upper Arm	Location	3.	2. ..	3.
		Movement	1. inward	2. outward	
	Forearm	Location	3.	2. ..	3.
		Movement	1. 0-45	1. 0-45	
	Hand and Wrist	Shape	5.	2.	5.
		Movement	7. direct mov..	7. direct movement..	

Figure 1.1: FORM annotation of Jan24.mov, using Anvil as the annotation tool

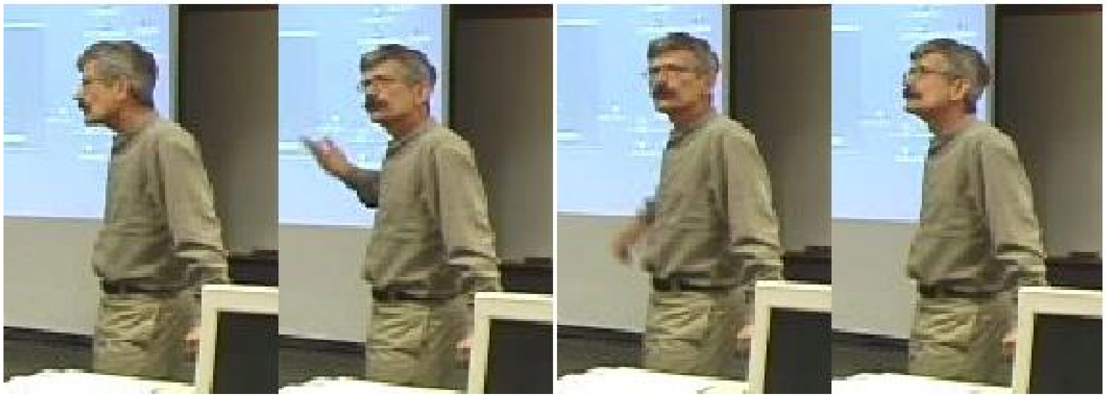


Figure 1.2: Snapshots of Brian MacWhinney on January 24, 2000

Chapter 2

Related Work

There are multiple threads of research that relate to our goal of building a database of gestural information in communicative interaction. In this chapter we will first explore the most important qualitative models of gesture developed in the psychological, linguistic and social-science literature. We will then look at some important work from the dance/movement community. Next we will survey various other computer-based annotation systems. Finally, we will explore the two major ways of gathering gestural information automatically and discuss their pros and cons vis-à-vis the goals of FORM.

2.1 Qualitative Theories of Gesture

2.1.1 David Efron: Gesture, Race and Culture

As highlighted in [108], David Efron's work [25][1] is the seminal work on describing how a gesture is performed. Efron's primary concerns were anthropological, and his contribution to the theory of gesture grew from these concerns. Primarily, he

[1] Reprinted as [26]

wanted to understand how race and environment influenced culture as given in the use of gestures. That is, he asked whether the types of gestures used by a certain *race* were genetic or environmental. He compared Jewish immigrants to New York against Italian immigrants.

To undertake this study, however, Efron had to develop a means of coding the gestures and comparing them against each other. He divided gestural movement into three distinct phases: the preparation phase, the stroke phase, and the retraction phase. For Efron, it is in the stroke phase that the gesture is performed. During preparation and retraction the hands are, respectively, raised to the starting point of the stroke and then return to a resting position. The interesting theoretical issue for us here is his assumption that a stroke—the gesturally salient part of the movement—is unambiguously perceived by the human observer.

This assumption that the stroke is clearly, and unambiguously, recognized by an observer is part of the motivation for the experiments later in this thesis. Ideally, we wish to know whether this perceived “stroke” is given solely by the physical signal or if there are other cognitive dimensions involved as well.

Following this work by Efron, two researchers stand out as particularly relevant to the work here: Adam Kendon and David McNeil.

2.1.2 Adam Kendon

Adam Kendon’s contribution to gesture research is manifold. For a good overview with many references to other good review articles see [56]. In this section we will highlight four of the important contributions he has made to the field. These are:

1. Evidence that humans easily recognize gestural movement;

2. Further development of Efron’s Preparation-Stroke-Retraction analysis;

3. An analysis of the various types of gesticulation; and
4. An annotation scheme to capture the "kinetics" of gesture.

Humans Easily Recognize Gestures

In [56], Kendon defines a "gesture" as:

> ... any visible bodily action by which meaning is given voluntary expression. "Gesture" is to considered separate from emotional expressions; it does not include the various minor ticks ... or "nervous movements" which, though informative to the eye of another, are not treated in the intentions or part of the individual's "official" or "given" expression. Practical action will also not be considered as gesture, even if, as is sometimes the case, such expressions have an expressive purpose. Only if a person is seen to pantomime a practical action will this be considered a part of gesture[2].

Although this definition is not exact, it nicely captures the intuition that to be a gesture the bodily movement concerned must be designed to convey some meaning. And, although, we can often glean important information from watching the bodily movements of a person, we, again, only call gestures those movements that are intended by the individual to carry meaning. The important point in this section is that people have no trouble making these distinctions between gestural and non-gestural movement. In [52] and [55], Kendon had subjects view movies of speakers interacting in a language unknown to the subjects. The viewers had no trouble picking out which body movements were gestures and which were not.

Further Development of Efron's Analysis

The second important contribution made by Kendon, from our point of view, is the extension of the work on strokes by Efron. Kendon generalizes the analysis of

[2] [56], p. 13

preparations, strokes and retractions to that of a "gesture unit" and its constituent parts. Again, from [56]:

> Gesticulation of the forelimbs can be observed to be organized into phrases of movements, and these phrases of movement can often observe some complexity, incorporating smaller phrases of movement within them .... A unit of gesticulation, which may be called a Gesture Unit, is composed of an excursion of the forelimbs from a position of rest into free space in front of the speaker and back again to a position of rest. During the course of such an excursion, one or more Gesture Phrases may be observed to occur. A Gesture Phrase is distinguished by a nucleus of movement, termed the stroke, in which the gesticulating limb performs some definite pattern of movement, set apart within the overall flow by having distinct dynamics and spatial qualities.

As mentioned in Chapter 1, the current work takes the theory just expressed as its tentative starting point. We do, however, make some terminological changes. What Kendon calls the Gesture Unit, we term the Gesture Excursion. And what Kendon calls the Gesture Phrase, we call the Gesture Phase. Both systems agree that

$$\text{Gesture Phrase/Gesture Phase} \in \text{Gesture Unit/Gesture Excursion}$$

However, we find, in practice, that Kendon's terminology was often taken to mean the exact opposite, that is,

$$\text{Gesture Unit} \in \text{Gesture Phrase},$$

since a "unit" seems as if it should be an atom and a "phrase" implies complexity.

Gesture and Speech

In [55], Kendon relates the structure of gestures, based on phrases, to the structure of speech, based on *tone units*. Just as the stroke phase is

> the nucleus of movement with definite form and enhanced dynamic qualities . . . preceded by a preparatory movement and succeeded by a movement which either moves the limb back to its rest position or repositions it for the beginning of a new [stroke] phrase,

so, too, is the *tone unit* a

> phonologically defined syllabic grouping with a single intonation tone[3].

Kendon found that the stroke phase occurs simultaneously—or just before—the tone unit.

However, Kendon, in [57] points out that gesture and speech are not necessarily equivalent modes of expression. For example, gesture and speech do not obey the same turn-taking rules. And, gesture and speech seem to have independent contexts, e.g., gestures might be more numerous or more pronounced in a noisy environment when speech reception is impaired. Further, gestures can often express better than speech spatial information like distance, size, or trajectory. Kendon devised a continuum of gesticulation which rates forelimb movement in accordance with how connected to speech. It is:

1. Gesticulation;
2. Language-like Gestures;
3. Pantomimes;
4. Emblems; and
5. Sign Languages.

As we move from Gesticulation to Sign Language the necessity for speech accompaniment declines and the presence of language-like properties increases.

[3] [55], p. 34

The "Kinetics" of Gesture

Kendon has also been a pioneer in developing what he calls the *kinetics* of gesture, analogous to the *phonetics* of speech. This goal, best articulated in "An Agenda for Gesture Studies" ([58]), is to be able to annotate and analyze at a much lower, that is, fine-grained, level than had been done up until then. As such, he develops in [60] a scheme which captures how joints are bent, how the different aspects of the arm move, and even how these different dimensions of gesture align with speech. Consider Figure 2.1 as an example. It is the annotation of a transcript of a video of a tour guide discussing the stone used in a specific building.[4] We will not explain fully the meaning of the annotations. It is sufficient to know that they describe physical position and changes in position of the speakers arms, hands, head and torso. For example, the annotation under "If its a good building material" says that a stroke was performed during the utterance of "building"; that the upper arm is at the side of the body, that is, there is no lift; that it is extended neither forward nor back; and that it is rotated inward. The elbow joint flexion is between 0^0 and 20^o. The thumb is at rest and is in contact with the index finger. The first and second fingers are extended, are wiggling and are rubbing against each other. The forearm is neutral and the metacarpal orientation is away from the speaker.

The important aspect of this annotation scheme, from our perspective, is the way in which the temporal aspects of the gesture are aligned with the temporal aspects of speech. Notice the series of tildes, stars and underscores. These represent different phases of the gesture (viz., preparation, stroke, retraction, etc) and the placement of the symbols represents the alignment of these phases with the speech. It is quickly

[4] Please note that this video, transcript and annotation were presented by Adam Kendon in a seminar held at the University of Pennsylvania. All copyrights belong to him. Please do not distribute without Professor Kendon's expressed permission.

clear that, from the perspective of doing computational analysis of gesture, that a clearer and more fine-grained method of annotation and temporal alignment is needed.

2.1.3 David McNeill

According to McNeill ([75] p. 8), the modern study of gesture—as opposed to the ancient study of rhetoric—has had two main shifts. The first, starting with Efron, was towards studying gestures in everyday life. The second, with Kendon, was towards considering gestures as a part of language itself. McNeill's work, to a very large degree, is a product of these two shifts. In this section we will first present the results of his key experiments with Elena Levy. Next we will present his second major contribution, as it relates to our goals, the development of the most widely used typology of gestures.

McNeill and Levy: Gestures and Speech are Parts of the Same Whole

In their seminal work ([76]), David McNeill and Elena Levy showed films and/or cartoons to subjects who then had to recount the story of the film/cartoon immediately to a third person from memory. The subjects thought they were engaged in a story–telling study and did not know that gestures were the interest of the researchers. The recounting of the stories was videotaped. These video recordings formed the raw data of their experiments. They were coded for prosody, pauses, and speech irregularities. Finally, they were coded for gestural information, including the type of gesture, gesture duration, the phases of the gesture, and the spatial aspects of the stroke of the gesture. The results of this work were similar to Kendon's: speech and gesture are part of the same common whole. In [73], McNeill summarizes this

GB 2:22:45

```
(1)      (2)                      (3)
Again/ good quality stone/It is like a lot of stone/ um

(4)                        (5)
Throughout the world /If its a good building material/
                           |~~~~~~~~~*******_______|
                           [          fgp 1         ]
                                o0051.G(1, 2wz)na.N
                           {          GUI              }

      (6)                                (7)                    (8)
 [........]When you knock something down /you don't throw it away/ you use it again/
RH|~~~~|*************|_________|~~~~|********|_______|~~~~~|*******_____
 [             fgp 1                ][        fgp 2               ][     fgp 3
          x0002.B2na.LVG                x116[2]2.B2nc[sqa].          x0002.B2pqa.LVG_
                                      LTP+WX

 {                  GU II

      (9)
 [........]because it will last forever/
RH ______________________*****_____________________|
                          ][              fgp 4              ] hand relaxes and is lowered
                           0002.B2pqa[na].N

HD|~~~~*******/****/***|******|_____________________|  head, then trunk turns left.
 [          hgp 1        ][   hgp 2                     ]

                          GU II                          }
```

Figure 2.1: “Again, good quality stone . . . ”

relationship as follows:

1. Gestures occur only during speech;
2. Gestures and speech are semantically and pragmatically co-expressive;
3. Gestures and speech are synchronous;
4. Gestures and speech develop together in children; and
5. Gestures and speech break down together in aphasia. . . .

> . . . [f]or all of these reasons, gestures and speech are most appropriately regarded as two sides of a single underlying verbal-gestural process of constructing and presenting meanings. . . . Despite the fundamental character of the differences between gestures and speech—one global and synthetic, the other linear and segmented—they are closely tied together in meaning, time, function, development, and dissolution[5].

Also in [73], McNeill reaffirms Kendon's timing results, viz. that gesture has timing constraints vis-à-vis speech. McNeill hypothesizes three types of rules which govern how gesture and speech synchronize:

- Phonological—the stroke of the gesture cannot follow the peak syllable of the accompanying speech;
- Semantic—gesture and speech must cover the same content if they co-occur; and
- Pragmatic—gesture and speech must serve the same pragmatic purpose.

McNeill claims that these principles, although not theoretically inviolable, have been consistently empirically verified.

Types of Gestures

McNeill is also important to the field for having developed the most widely used typology of gestures. [73] presents a scheme which divides the gesture space into four basic types:

[5] pp. 22-26

- Beat gestures;
- Iconic gestures;
- Metaphoric gestures; and
- Deictic gestures.

These categories are not meant to be mutually exclusive, although [73] has been blamed for making it appear so. According to the McNeill Lab web site ([77]):

> A misconception has arisen about the nature of the gesture categories described in *Hand and Mind*, to wit, that they are mutually exclusive bins into which gestures should be dumped. In fact, pretty much any gesture is going to involve more than one category. Take a classic upward path gesture of the sort that many subjects produce when they describe the event of the cat climbing up the pipe in our cartoon stimulus[6]. This gesture involves an iconic path-for-path mapping, but is also deictic. . . . Even "simple" beats are often made in a particular location which the speaker has given further structure (e.g. by setting up an entity there and repeatedly referring to it in that spatial location). Metaphoric gestures are de facto iconic gestures. . . . The notion of a type, therefore, should be considered as a continuum–with a given gesture having more or less iconicity, metaphoricity, etc.

This work has been very influential, and has been the basis for at least one major computational project (see the BEAT toolkit, below). However, this level of analysis only serves to categorize the gesture. It provides no useful computational information for automatic at a more fine-grained level. As asked in the Chapter 1, what are the more atomic parameters that make up a beat? Do small changes in these parameters cause subtle changes in the message the gesture conveys? Eventually, we hope FORM will be able to make a contribution toward answering questions like these.

[6] Cf. experiments described in the previous section.

2.1.4 Summary of Psychological and Linguistic Theories

This short review was not able to touch on all relevant gesture research, and we have purposefully concentrated on that which is most related to FORM[7]. However, Efron, Kendon, and McNeill constitute the background for much of the work being done in contemporary gesture research. The important key points can be summarized as follows:

1. Gestures are distinct from non-gestural movements, and the difference is easily observed;
2. The core of the gesture is the *stroke*, which has distinct qualities, is discernible by observers, and is connected with other aspects of communication; and
3. Gestures work in concert with speech and are a coordinated part of the whole communicative interaction.

2.2 Laban Movement Analysis and Labanotation

Those interested in gesture from a social-scientific perspective are not the only ones that seek a rigorous description or coding scheme for movement. The dance community has been striving for possibly millennia to develop a system that allows for correct and reproducible transcriptions of the movements of dancers. The "Holy Grail" here is a *score* of dance movements that is reproducible by a choreographer and dancers—just as the music encoded in a musical score is reproducible by a conductor and musicians. The preface to a book presenting one such scheme notes the following:

> The problem of recording movements of the human body is almost as old as the art of dancing. It has been said that the ancient Egyptians had a system of notation, but there is no evidence to prove this is so.

[7] Much other interesting work of a social-science nature has been done on gesture. For examples of collections offering excellent coverage of this field, see [78] and [75].

> To-day, when there is more general interest in the subject than ever before, there are several systems in use . . . and serious attempts are being made to put them into practice, both for recording ballets and for such things as research in industrial welfare[8].

In this section we will explore two systems for capturing the movement of dancers, both of which have their origins in the work of Rudolph Laban: Labanotation and Laban Movement Analysis (LMA)[9]. The primary aim of Labanotation is to record the physical and structural aspects of dance movements, e.g., direction, places, positions, and involved body parts. LMA is concerned more with a qualitative description of the movement, not simply the physical characteristics. As such, it, at least tacitly, assumes that these qualitative aspects are the more cognitively salient.

2.2.1 Labanotation

As stated above, the principle goal of Labanotation is the recording of the physical traits of a movement. However, the recording of these traits is at a very high level. Moreover, the recording of the information recorded is of a holistic nature. That is, the movement in question is described as a whole. If there are particulars of the movement that are of interest, these are captured in addition. The following extended quotation from [48] nicely demonstrates this issue.

> The basic principle of Labanotation is that simple, natural movement is written in the most simple, direct way. The second premise is that everything that occurs is recorded. These two statements may seem contradictory. Actually, they are not, but we must know where to draw the line. Let us take an example. Walking is a simple, natural movement. Each person varies slightly in the manner in which he walks, but the basic

[8] [99], p. i.

[9] Although the history of both systems is rich and varied—both because Laban's work evolved over his life and because each system has taken on a life of its own—a complete description of this history is beyond the scope of this short survey. For a more in depth survey of the concepts behind both Labanotation and LMA, see [68], [91], [43], [48], and [46].

> process is the same. The general process is uncomplicated, though the mechanical motions are really complex. The movements may be written in Labanotation in exact detail, as is sometimes necessary when a very stylized way of walking is desired.
>
> But for general purposes we write only the essentials: The weight of the body moves forward by means of a step taken by one leg. This is the essence of the movement which anyone can understand. In most cases, no more need be indicated; we describe the basic form simply and directly, without fear of misunderstanding. Once we have a grasp of this essential pattern, we can quickly perceive those particular cases in which modifications or embellishments have been added, and we then know what extra detail must be written. By describing the pure form directly, we learn to understand the basic categories of movement and to distinguish the manipulations of them which are the source of style[10]

As described, Labanotation, does allow for fine-grained descriptions. Consider, for example Figure 2.2. This is the notation for a cartwheel to the right “performed with a contracted arm, the degree of contraction being retained.” ([43], p. 106)

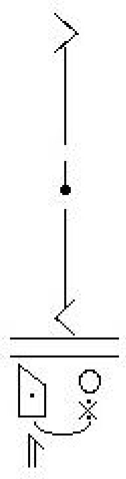


Figure 2.2: Labanotation of a Cartwheel with a Contracted Arm

Each of these symbols is given a precise meaning, the length of some imply duration, and there is a presupposed grid of three vertical lines indicating the center

[10] [48], pp. 9–10.

of the movement and movement away from that center.

Labanotation is widely used, and it seems to lead to positive results for choreography. However, from a computational perspective, it is a very difficult notation, as it is both complex and imprecise. FORM's goal, in circumstances such as these, would be to understand what sets of joint parameters, etc., go into classifying a movement as a cartwheel in the first place. However, we fully recognize that these higher-level categories are probably more cognitively salient—just as phonemes or words are more cognitively salient than an understanding of lip shape and tongue position. Further, these higher-level categories give us the equivalence classes we can use to analyze the lower-level parameters. That is, without first knowing what a cartwheel is, we can not explore the joint-parameter settings of all cartwheels.

2.2.2 Laban Movement Analysis

In the same way that Labanotation developed from Laban's ideas for notation, so too did Laban Movement Analysis—first just called Movement Analysis[11]—develop from Laban's idea that there are important qualitative descriptions of movements as well as physical. LMA divides a movement into five major components: Body, Space, Effort, Shape, and Relationship. The majority of the research applying LMA to gesture studies deals with only Effort and Shape, "because these two are the major direct specifications or indications of human movements."[12] As such, we will briefly touch on these two components here.

Effort can be broken into the following factors: Space, Weight, Time and FLOW ([108], [22]). Shape comprises: Horizontal, Vertical, Sagital and FLOW. For each of these eight factors there is an associated spectrum. For the Effort parameters, the

[11] Again, we are collapsing history here. Laban Movement Analysis was influenced and developed by a number of people. Cf. [22], [3], and [23].

[12] [108], p. 40.

spectra go from the "indulging" end to the "contending" end. For the Shape parameters, they move from "convex" to "concave." See Figure 2.3 for all the parameters and the values of their extrema.

Movements, and more precisely for our purposes, gestures, can be described as combinations of these parameters and their values along their respective spectra. It is important to note here that these Effort/Shape parameters (along with the other LMA parameters) are seen as irreducible and are necessary to giving a correct description of the movement. There has been a good deal of recent work at computerizing these parameters for computer animation[13] that has been somewhat successful in generating realistic gestures in avatars[14]. For the present project, however, Effort/Shape parameters are assumed to be too high-level. That is, we do not yet accept their irreducibility. For present purposes, we see Effort and Shape as describing equivalence classes of movements. Our interest is in understanding the low-level physical aspects that make up these classes of movements, although it would be fascinating, in the future, to experiment with the relationship between FORM and these LMA parameters. Future experiments might include augmenting the FORM corpus with LMA parameters to see, given a particular LMA-class of movements, whether or not they have common FORM-based descriptions. If so, this may be evidence that the Effort/Shape parameters are not irreducible. However, it may be the case that no underlying commonality is found, it which case, this furthers the irreducibility claim.

[13] [108], [14], and [15].

[14] Although, further psychological testing is required to back up the preliminary results.

Indulging	**EFFORT**	**Contending**
Indirecting	[Space]	Directing
Diminishing Pressure	[Weight]	Increasing Pressure
Decelerating	[Time]	Accelerating
Freeing	[FLOW]	Binding

Convex	**SHAPE**	**Concave**
Spreading	[Horizontal]	Enclosing
Rising	[Vertical]	Descending
Advancing	[Sagittal]	Retiring
Growing	[FLOW]	Shrinking

Figure 2.3: LMA Effort and Shape Parameters and their Spectra

2.3 Annotation Graphs

As briefly mentioned in Chapter 1, the underlying data structure for FORM is the Annotation Graph (AG). In this section, we describe the structure of AGs and why they are useful the FORM database.

As described in [6], Annotation Graphs are a formal framework for "representing linguistic annotations of time series data." AGs do this by extracting away from the physical-storage layer, as well as from application-specific formatting, to provide a "logical layer for annotation systems." Further, an annotation graph is a directed acyclic graph (DAG) such that the nodes represent time-stamps of some given signal and the arcs represent some linguistic event that spans the time between the time-stamps. For FORM, these arcs are as encoded as *attribute:value* pairs which capture the gestural information for each aspect of the arms and hands that has changed between the two time-stamps. In Figure 2.4, then, the arc labeled *HandandWrist.Movement* from 1:13.34 to 1:13.57 is short-hand for the arcs given in Table 2.1. They encode the kinematics of Brian's moving his right hand or wrist during this time period. Similarly, the arc from 1:13.24 to 1:13.67 is short-hand for Table 2.2, which encodes a change in his right hand's shape.[15]

The particular advantage to using AGs to encode the kinematics of gesture, or any linguistic signal, is the ease with which the annotation can be extended to include other data. The only constraint is that all the data share the same time line. As such, researchers can easily extend the FORM corpus to include, for example, grammatical information, discourse structure, facial expression, etc. Figure 2.5 is such an augmented AG. It is another representation of the video clip from Figure 1.2 (Jan24-09.mov) and is augmented with head/torso movement, speech transcription

[15] For the example given in Figure 2.4, Brian is only moving his right hand. Accordingly, the *Right.* which normally would have been prepended to the arc-labels has been left off.

and syntactic information, and intonation/pitch information. Note that this is a conservative extension of the original AG from Figure 2.4; the original AG remains unchanged and new information is simply added.

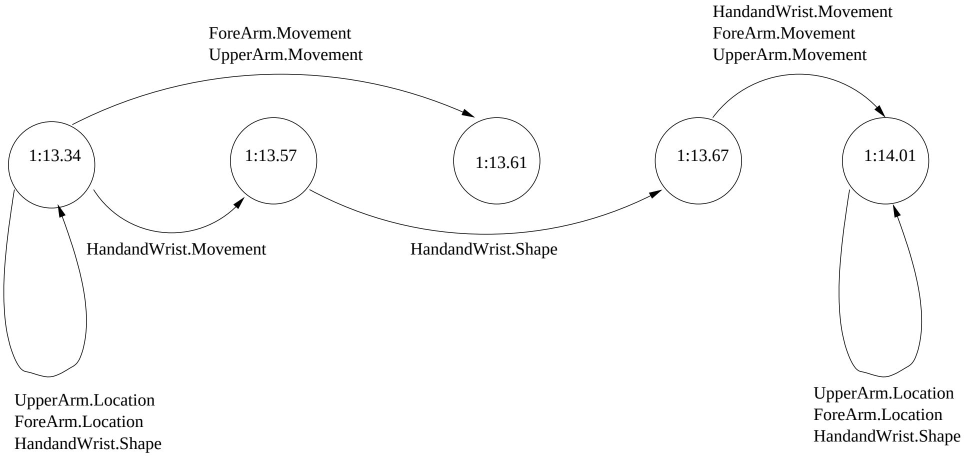


Figure 2.4: FORM/Annotation Graph representation of example gesture

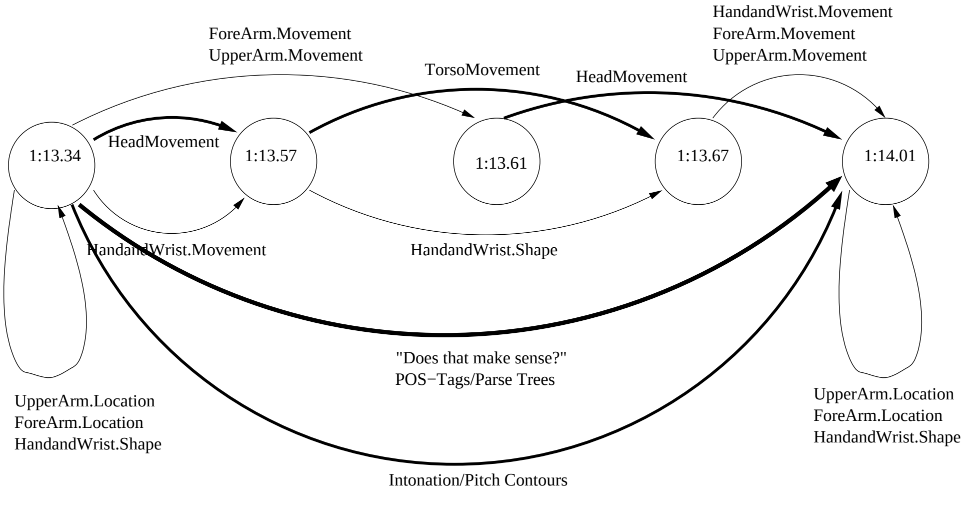


Figure 2.5: Augmented FORM annotation graph of Jan24-09.mov

Start	End	Attribute	Value
1:13:67	1:14:01	Obscured	True
1:13:67	1:14:01	Hand Movement	7
1:13:67	1:14:01	Wrist Up-Down Movement	1
1:13:67	1:14:01	Strokes	1

Table 2.1: Hand and Wrist Movement Attribute Arcs, 1:13:67 - 1:14:01

Start	End	Attribute	Value
1:13:57	1:13:67	Handshape Letter	C
1:13:57	1:13:67	Handshape Group	2
1:13:57	1:13:67	Wrist Bend: Side to Side	2
1:13:57	1:13:67	Wrist Bend: Up and Down	3
1:13:57	1:13:67	Tension	2

Table 2.2: Hand and Wrist Shape Attribute Arcs, 1:13:57 - 1:13:67

2.4 Computer-based Annotation Tools and Systems

In this section, we present some of the most important tools for building annotated databases of communicative interaction. Although there are many more than listed here, each of these has had a large impact on their respective communities.

2.4.1 CHILDES/CLAN

The CHILDES/CLAN system [16] is a suite of tools for studying conversational interactions in general. The suite allows for, among other things, the coding and analyzing of transcripts and for linking those transcripts to digitized audio and video. CLAN supports both CHAT and CA (Conversational Analysis) notation, with the alignment of text to the digitized media at the phrase level.

The CHILDES/CLAN system has the major advantage of being one of the first of its kind. The CHILDES database of transcripts of parent-child interactions has dramatically pushed forward both the theory and science of linguistics and language-acquisition. Additionally, it appears possible—in the future—to integrate FORM data with that developed by CHILDES/CLAN into a unified data set. This is due to the open-ended nature and extensibility of both systems. However, from the perspective of actually annotating videos with fine-grained, time-aligned gesture data, CLAN presents a problem. It is possible to describe the gesture that occurred during an utterance, but, given that time alignment is only at the phrasal level, we are unable to finely associate the parts of the gesture with other aspects of conversational interaction.

2.4.2 SignStream

SignStream [80] allows users to annotate video and audio language data in multiple parallel fields that display the temporal alignment and relations among events. It has been used most extensively for analysis of signed languages. It allows for annotation of manual and non-manual (head, face, body) information; type of message (e.g. Wh-question); parts of speech; and spoken-language translations of sentences.

Although SignStream would work with the FORM annotation scheme, and there has been some attempt at integrating the two projects, its interface is too comprehensive. Anvil, described below, more easily allows an annotator to quickly see the relationship among all the aspects of left arm, right arm, head and torso movement.

2.4.3 Anvil

Anvil [61] is a Java-based tool which permits multi-layered annotation of video with gesture, posture, and discourse information. The tags used can be freely specified, and can easily be hierarchically arranged. (Again, see Figure 1.1 as an example.) Additionally, the extensible nature of Anvil allows for the development of an Annotation Graph plug-in, so FORM data can be directly exported to AG format.

2.5 Computational Methods of Capturing Gesture Data

Considering FORM's goal of building a dataset of gestures during communicative interaction, it might seem to make more sense to choose an automatic method of data acquisition over the annotation method that FORM employs. In this section, then, we briefly present the two most important methods for automatic data gathering and explain, for each, why it is not as satisfactory as annotation for our purposes.

2.5.1 Motion Capture

Motion capture methods use electro-magnetic or optical sensors placed on a subject's body to capture position and orientation information at many points. These sensors are usually attached by Velcro straps and/or by wearing a full-body Lycra suit. Additionally, they must perform the gestures within a constrained area—either a skeletal metal cube that has cameras to sense the optical markers or an invisible cube containing the magnetic field of the system—and cannot move outside of it[16].

The problems with motion capture—from FORM's point of view—are three-fold. Firstly, it requires that the subject perform the motions in a highly constrained laboratory setting. As one of FORM's goals is a portable system that allows for gestural information to be gathered from a variety of natural settings, motion capture presents

[16] Cf. [102] and [21] as a representative sample.

a problem. Secondly, the movements of the subjects themselves are constrained by the equipment. It is true that motion capture has been successfully used to produce the movements of avatars and animated characters ([4], [39], and [40]). However, in these cases we are only concerned that the movements appear correct. The subject him/herself may well move differently under natural conditions. Even if all motion-captured movement seems "normal," this does not mean it does not constrain the movements to a smaller set—thereby precluding other normal movements useful for scientific study. Finally, motion-capture only allows for the acquisition of new data. Given the wealth of video data that already exists, a method that allows for it to be used is preferable. Annotation, although time consuming, allows for the use of pre-existing data, allows the subjects to remain in a natural setting, and does not necessitate the use of equipment that may restrict the subjects' movement.

2.5.2 Computer-Vision Methods

At first glance, a computer-vision-based approach may seem to be the ideal solution. Firstly, computer-vision methods can be applied to pre-existing video. Secondly, as video can be taken anywhere, there is no requirement that the subject be in a laboratory setting. And, finally, as the subject is simply being video recorded, there are no artificial constraints on his/her movement. The problem, however, is that current vision methods are not yet sufficient to achieve the first two points.

Current computer-vision methods have yet to fully solve the problem of inferring 3-dimensionality from a 2-dimensional video source. Additionally, researchers in these fields are just at the beginning of solving other problems like object recognition over time, segmenting the arm, and tracking arm/hand positions relative to the environment. Accordingly, many successful attempts utilize passive markers. That is, they put colored markers on the arms of the subject, who must wear appropriately colored clothing, and stand in front of an appropriately contrasting background. Additionally, for the information to be most useful, there must be a fairly precise arm-model of the subject in question.

So, although, computer-vision methods allow for full movement freedom, they still require a subject to be in a laboratory setting, positioned so that the marker work their best. This inability to extract appropriate information outside of laboratory conditions also precludes using current vision techniques on pre-existing data. Additionally, FORM can be used in scientific field-studies in a way that video capture can not. All that is needed is a digital video camera and a notebook computer.

2.6 Systems for Computational Analysis and Generation of Gesture

2.6.1 BEAT Toolkit: Justine Cassell, et al.

The Behavior Expression Animation Toolkit (BEAT) was developed at the MIT Media Lab in the Gesture and Narrative Language Research Group([10]). This work is advanced and is, by far, the most influential to date. It allows for the easy generation of synchronized speech and gesture in computer-animated characters. The animator simply types in the sentence that he/she wishes the character to say, and the BEAT Toolkit generates marked-up text which can serve as the input to an animation system. The system is extensible to many different communicative behaviors and domains. The output generated for a given input string is domain specific, and the training data for that domain must be provided.

The main purpose of BEAT is to appropriately schedule gestures (and other non-verbal behaviors) so they are synchronized with the speech.

> The BEAT toolkit is the first of a new generation (the *beat* generation) of animation tool that extracts actual linguistic and contextual information from text in order to suggest correlated gestures, eye gaze, and other nonverbal behaviors, and to synchronize those behaviors to one another. For those animators who wish to maintain the most control over output, BEAT can be seen as a kind of "snap-to-grid" for communicative actions: if animators input text, and a set of eye, face, head and hand behaviors for phrases, the system will correctly align the behaviors to one another, and send the timings to an animation system. For animators who wish to concentrate on higher level concerns such as personality, or lower level concerns such as motion characteristics, BEAT takes care of the middle level of animation: choosing how nonverbal behaviors can best convey the message of typed text, and scheduling them.[17]

BEAT is most concerned with the automatic generations of the timings, and the higher and lower levels, as aforementioned, are left to the animator. In particular, it is with the lower level of specifying motion characteristics that FORM is most concerned. We see FORM as potentially a more robust way to specify the gestures for which BEAT schedules the timings. The typology of gestures that BEAT uses is based on the work of McNeill[18]. As such, it sees gestures through the eyes of his ontology. It is, then, left up to the animator to specify exactly how a beat or a deictic, for example, is to be animated. We believe the data generated by the FORM annotation system could allow for a more robust output from BEAT-like systems, which could further alleviate the work of animators.

2.6.2 VISLab: Francis Quek

The VISLab project ([104]) is a large-scale, low-level-of-analysis research project developed and led by Francis Quek at Wright State University. It has achieved

[17] [10], p. 8

[18] Cf. section 2.1.1, above.

significant results in understanding the relationship of speech to gesture ([92]). The long-term intent of the project is to create a large-scale dataset of videos annotated with information about gesture, speech and gaze.

This project is in the same spirit as the FORM project. There are, however, major differences between the two projects. Firstly, FORM aims at developing a lower-level representation that humans can use to annotate gestures *and* that machines can use to analyze and gestures. However, the VISLab system's level of representation is much lower still. They use multiple cameras to extract 3D information about position, velocity, acceleration, etc. concerning a gesture. They are doing the *physics* of gesture, where FORM is looking at something closer to the *phonetics* of gesture. Secondly, the VISLab system requires a complex set up of multiple, precisely-positioned cameras and proper placement of the subjects in order to gather their data. Again, FORM allows any researcher with a notebook PC and a video camera to generate useful data. Thus, FORM can be used in the "field," where the VISLab system requires a laboratory setting.

Chapter 3

The FORM Annotation Scheme

In this chapter we describe the FORM annotation system in detail. Additional best-practices information is given in Appendix A. These best practices were written by Kari Myers, Lisa Britt, Paul Howard and Chris Osborn.

3.1 The FORM Annotation Scheme

FORM is designed as a series of tracks representing different aspects of the gestural space. Generally, each independently moved part of the body has two tracks, one track for Location/Shape/Orientation, and one for Movement. When a part of the body is held without movement with respect to the gesturer, a Location object describes its position and spans the amount of time the position is held. When a part of the body is in motion, Location objects of zero duration are placed at the beginning and end of the movement to show where the gesture began and ended. Location objects with zero duration are also used to indicate the Location information at critical points in certain complex gestures, i.e., Location objects = keyframes.

An object in a movement track spans the time period in which the body part in question is in motion. It is often the case that one part of the body will remain static while others move. For example, a single hand shape may be held throughout a gesture in which the upper arm moves. FORM's multi-track system allows such disparate parts of single gestures to be easily annotated separately. Once all tracks are filled with the appropriate information, it is easy to see the structure of a gesture broken down into its anatomical components. (This aspect is also seen in Figure 1.1.)

At the highest level of FORM are groups. Groups can contain subgroups. Within each group or subgroup are tracks. Each track contains a list of attributes concerning a particular part of the arm or body. At the lowest level (under each attribute), all possible values are listed. The structure, then, is as follow:

Group

Subgroup

Track

ATTRIBUTE

Value

The following descriptions will follow this structure. The groups described are Right/Left Arm, Gesture Obscured, Excursion Duration, and Two-Handed Gesture. Not described are Head and Torso Movement/Location. These will be implemented in a later version of FORM.

Right/Left Arm

Upper Arm (from the shoulder to the elbow).

Location

UPPER ARM LIFT (from side of the body)

no lift

0-45

approx. 45

45-90

approx. 90

90-135

approx. 135

135-180

approx. 180

Relative elbow position: The upper arm lift attribute defines a circle on which the elbow can lie. The relative elbow position attribute indicates where on that circle the elbow lies. Combined, these two attributes provide information about the location of the elbow and location information (in relation to the shoulder) of the upper arm.

extremely inward

inward

front

front-outward

outward (in frontal plane)

behind

far behind

The next three attributes individually indicate the direction in which the biceps muscle is pointed in one spatial dimension. Taken together, these three attributes provide the orientation of the upper arm.

Biceps: Inward/Outward

none

inward

outward

Biceps: Upward/Downward

none

upward

downward

Biceps: Forward/Backward

none

forward

backward

Obscured: This is an binary attribute which allows the annotator to indicate if the attributes and values chosen were "guesses" necessitated by visual occlusion. This attribute is present in each of FORM's tracks.

Movement

The next three attributes individually indicate the direction of elbow movement in one spatial direction. When diagonal movement occurs, a non-none (i.e.not *none*) value for more than one of the attributes is chosen. Each attribute has combination values so repeated or back-and-forth motions can be annotated as such.

Linear movement (horizontal plane): Indicates the direction(s) of inward or outward elbow movement.

none

inward

outward

inward-outward

outward-inward

Linear movement (median plane): Indicates the direction(s) of upward or downward elbow movement.

none

up

down

up-down

down-up

Linear movement (frontal plane): Indicates the direction(s) of elbow movement towards or away from the body.

none

towards

away

towards-away

away-towards

Upper arm rotation: The degree of change of bicep direction. Ranges are exclusive. Direction of change is not included, as it can be inferred from the information in the Location track.

0-45

approx. 45

45-90

approx. 90

90-135

approx. 135

135-180

approx. 180

greater than 180

Arc-like movement: This boolean attribute indicates whether or not the elbow movement was arc-like. When checked, Location objects will co-occur to note the location of the elbow at the beginning, apex, and end of the movement.

Circular movement: A non-none value indicates that elbow movement is circular in shape and notes the plane in which the movement is performed as well as its direction (clockwise or counter-clockwise). As was the case for arc-like movements, the Location track will be simultaneously utilized, in this case noting the location of the elbow at the start and halfway mark of the circle. This convention allows the size of the circle to be inferred.

parallel to horizontal plane (c=clockwise)

parallel to horizontal plane (cc=counter-clockwise)

parallel to median plane (c)

parallel to median plane (cc)

parallel to frontal plane (c)

parallel to frontal plane (cc)

Effort[1]: Indicates the effort of the movement on a 1 to 5 scale.

Strokes: Indicates the number of strokes of a movement.

[1] Although we use the same term—*Effort*—as does LMA, we do not mean the same. For us, we are simply measuring the annotators assessment of the strength, force or tension with which the gesture in question is made.

1 . . . 20

More than 20

Indeterminate

OBSCURED

Forearm: the part of the arm extending from the from elbow to wrist)

Location

ELBOW FLEXION: The angle made by the bend in the elbow.

0-45

approx. 45

45-90

approx. 90

90-135

approx. 135

135-180

straight

FOREARM ORIENTATION: Describes the orientation of the forearm if the upper arm were to be by the side and the elbow flexed at 90 degrees.

supine

supine/neutral

neutral

neutral/prone

prone

prone/inverse

inverse

Obscured

Movement

Elbow flexion change: The amount of change in elbow flexion measured in degrees. Direction of flexion change is not indicated, as it can be inferred from information in the Location track.

0-45

approx. 45

45-90

approx. 90

90-135

approx. 135

135-180

approx. 180

Forearm rotation: Direction of change of forearm orientation. Amount of change is not indicated, as it can be inferred from information in the Location track.

none

inward

outward

inward-outward

outward-inward

Effort

Strokes

Obscured

Hand and Wrist

Shape: Information about the static shape of the hand and orientation of the wrist.

The next two attributes give values to describe the shape of the hand. The values are represented in a catalog of hand-shapes (Figure 3.1), which is organized as a two-dimensional matrix. This method is employed because the complexity of the hand would make purely physical descriptions too unwieldy.

Hand-shape Group: Indicates the group (organized by number of extended fingers with 0 representing fist and 6 referring to miscellaneous shapes) in the hand shape catalog.

Hand-shape Letter: Indicates the appropriate hand-shape within the selected group, ranging from *A* to *M*.

Tension: Describes the amount of tension apparent in the performer's hand. An average amount of tension corresponds to the "slightly tense" variable.

relaxed

slightly tense

very tense

Wrist bend: Up and down: How far the wrist is bent towards the upper side or under side of the forearm.

up

up-neutral

neutral

down-neutral

down

Wrist bend: Side to side: How the wrist is bent towards the thumb or little finger.

towards thumb

neutral

towards little finger

extremely towards little finger

Part of body touched:

none

top of head

eye (same)

eye (opposite)

ear (same)

ear (opposite)

temple (same)

temple (opposite)

nose

cheek (same)

cheek (opposite)

chin

neck (same side)

neck (center)

neck (opposite side)

chest

groin

Obscured

Movement

Hand movement: Describes type of hand movement (if any). The A joint refers to the knuckle furthest from the fingertip and the B joint refers to the first joint above the A joint. Information about the C joint (the joint closest to the fingertip) is not recorded because C joint movement is usually dependent upon movement of the B joint. The numbering scheme of the first three variables is explained in the Finger Coordination attribute.

none

1) A joint movement

2) B joint movement

3) A and B joint movement

wrist circular

thumb rubbing index finger

thumb rubbing multiple fingers

direct movement between two shapes

Wrist up-down movement: Describes the up-down movement (to the underside or upper side of the arm) of the wrist.

up

down

up-down

down-up

Wrist side-to-side movement

towards little finger

towards thumb

towards little finger-towards thumb

towards thumb-towards little finger

Finger coordination: Describes the motion of the fingers in relationship to each other. A non-none value is only applicable if one of the choices labeled 1, 2, or 3 was selected from the Hand movement attribute.

parallel movement without thumb

random movement, without thumb

parallel movement, with thumb

random movement, with thumb

movement in sequence

Effort

Strokes

Obscured

Excursion Duration: Marks the duration of the gesture excursion of the arm from a resting position to another resting position. Since there is ambiguity about what constitutes a single gesture, this convention for grouping movements was adopted.

Gesture Obscured: Similar to above except this attribute refers to the entire gesture excursion, rather than just one track.

Two-handed Gestures

Right-hand contact

none

thumb

index finger

middle finger

ring finger

little finger

palm

back of hand

more than one digit

holding

Left-hand contact: The list of values is identical to that of the Right-hand Contact attribute.

The following seven attributes are all boolean-valued.

moving in parallel

moving apart

moving towards

moving around one another

moving in alternation

crossed

Obscured

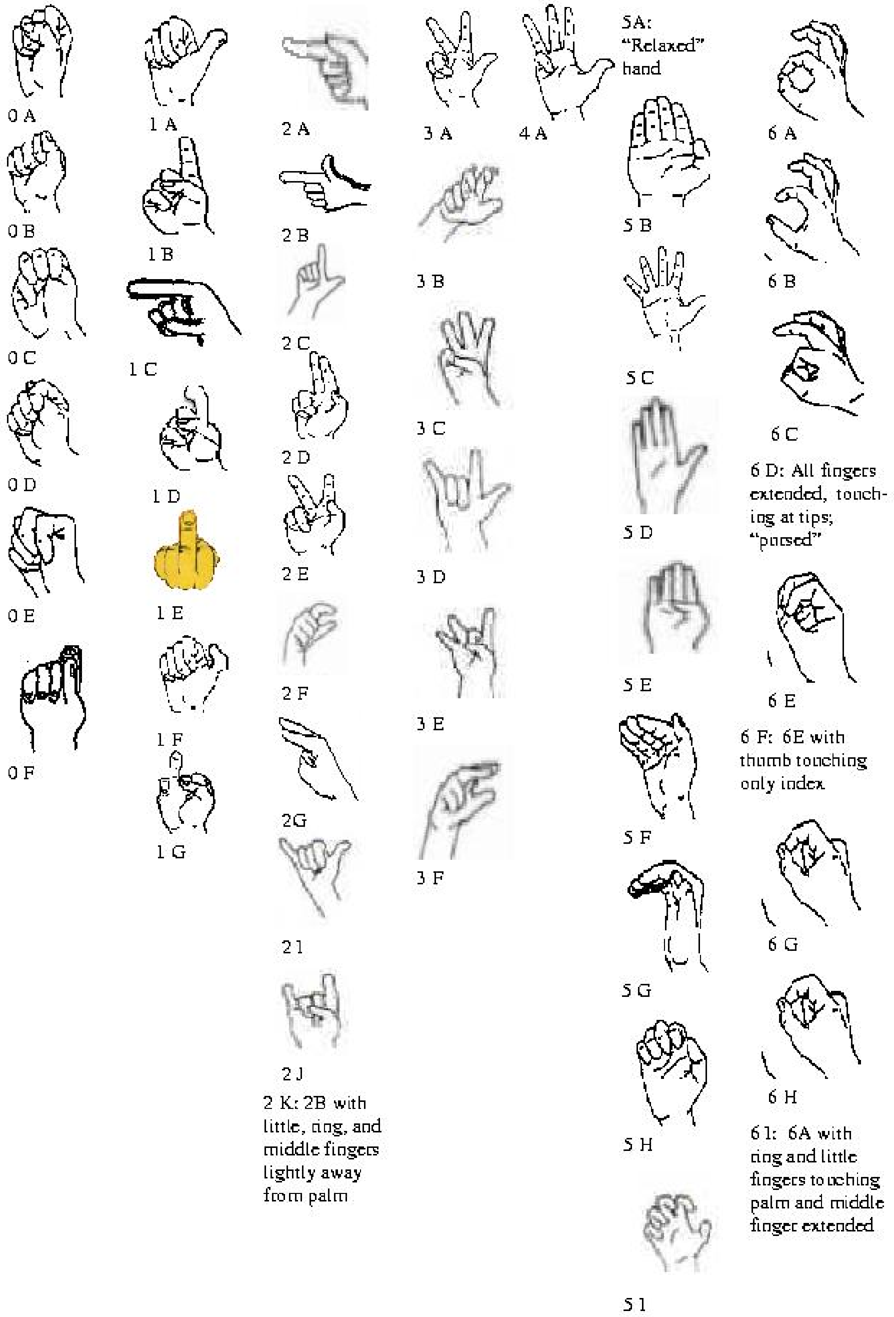

0 A
0 B
0 C
0 D
0 E
0 F
1 A
1 B
1 C
1 D
1 E
1 F
1 G
2 A
2 B
2 C
2 D
2 E
2 F
2G
2 1
2 J
2 K: 2B with little, ring, and middle fingers lightly away from palm
3 A
3 B
3 C
3 D
3 E
3 F
4 A
5A: "Relaxed" hand
5 B
5 C
5 D
5 E
5 F
5 G
5 H
5 1
6 A
6 B
6 C
6 D: All fingers extended, touching at tips; "pursed"
6 E
6 F: 6E with thumb touching only index
6 G
6 H
6 1: 6A with ring and little fingers touching palm and middle finger extended

Figure 3.1: Catalog of Hand Shapes. Based on the HamNoSys catalog (http://www.sign-lang.uni-hamburg.de/Projects/HamNoSys.html)

Chapter 4

The FORM Corpus: Ambiguity and Evaluation

4.1 Ambiguities in FORM

There are two known ambiguities in the FORM system as described in Chapter 3.

The first concerns the ***Upper Arm***:*Location* attributes that specify biceps direction. While anatomically it seems more accurate to describe the upper arm rotation by degrees of rotation rather than by using the direction of the biceps in free space, a problem arises when defining the neutral position of the arm rotation. For example, we could define *normal* as the position when the arm is held at the side with the biceps facing forward, as in Figure 4.1. However, if the elbow is then flexed to make a 90-degree angle with the upper arm, we get the position shown in Figure 4.2. If one then lifts the upper arm to the side so it is at 90 degrees with the body and still in the frontal plane, the upper arm has not rotated at all (see Figure 4.3). Let's call this position 1. If, however, one returns to the position in Figure 4.4, raises the upper arm forward so it is at 90 degrees with the body but parallel to the median plane, as in Figure 4.5, and then moves the upper arm 90 degrees to the side so that it is in the frontal plane again, as shown in Figure 4.6, it can be seen that this position is also reached without rotating the upper arm. Let's call this position 2. It is clear that position 1 is not the same as position 2, but both were reached by keeping the upper-arm in the *normal* position.

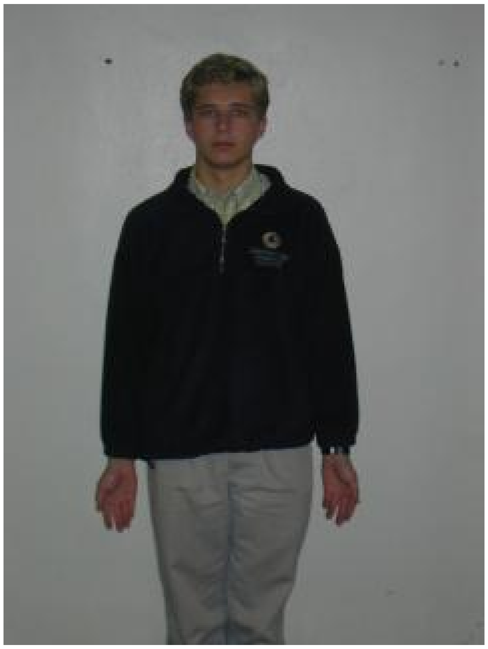


Figure 4.1: Position 1A

To solve this issue we could define a normal that is rotated 45 degrees when the Upper arm lift is at what we've deemed "approx. 90" and Relative elbow position is "front-outward." This convention, however, is hard to conceptualize by annotators and thus we opted to use the direction of biceps in free space, since it is more intuitive. The downside to this approach is that it allows for a large range of positions for each combination of values. Many positions could be called "forward-inward-upward," for example.

The second area of concern is in the ***Upper arm***:*Movement* track. This track

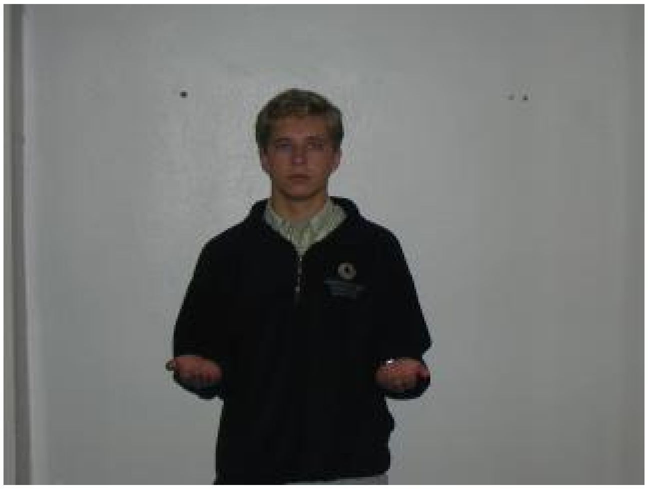


Figure 4.2: Position 1B

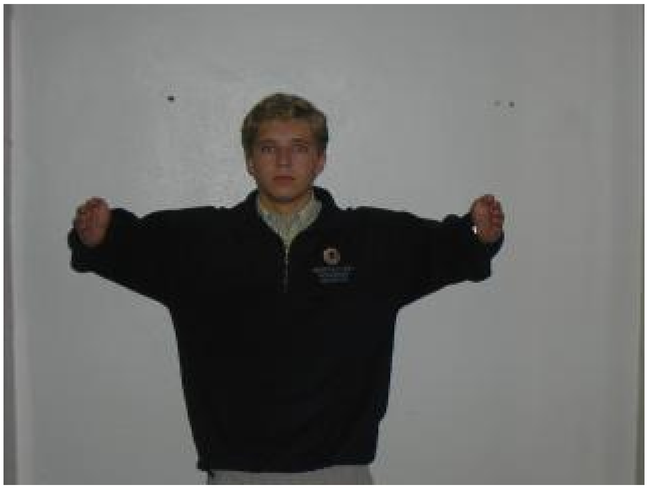


Figure 4.3: Position 1C

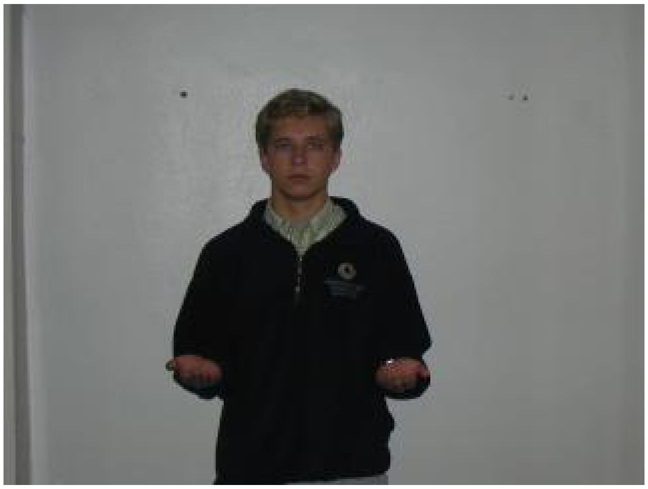


Figure 4.4: Position 2B

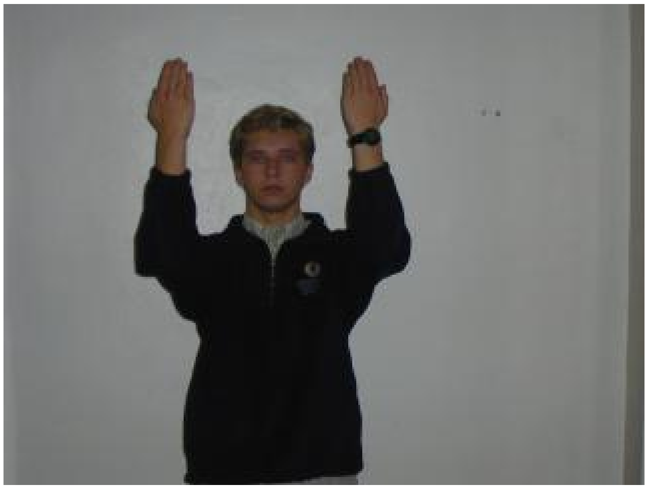


Figure 4.5: Position 2C

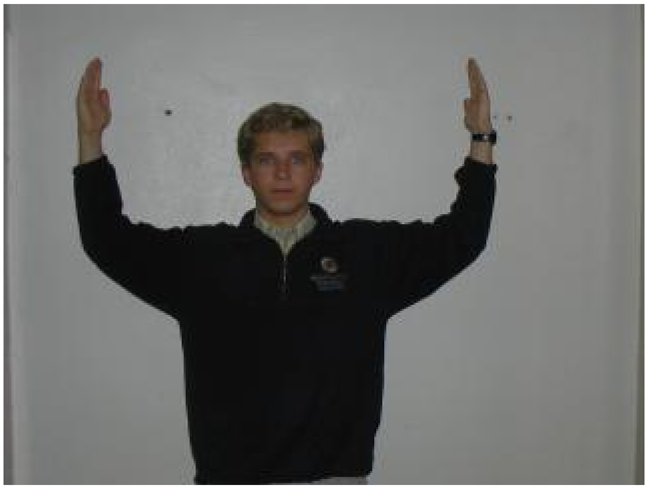


Figure 4.6: Position 2D

describes the movement of the upper arm independent of the forearm, elbow flexion, and hand-and-wrist. This movement can be described either as a combination of linear movement in different planes or as arc-like movement (using Location points to denote points along the curve). Since the upper arm is only able to move on a partial sphere with the shoulder as the center, it does not make sense anatomically to describe its movement as linear. However, since most movements are small enough not to appear as distinct arcs, linear values sufficiently approximate the movement.

4.1.1 Adding a $5 \times 5 \times 5$ Grid to FORM

In addition to the solution described above, we have also extended FORM to include additional attributes and values for wrist location. These allow us to specify in a $5 \times 5 \times 5$ grid the x, y, and z coordinates of the wrist. For some purposes the full description of location and movement will be desired, e.g., an experiment concerning how change in elbow flexion correlates with some aspect of pragmatics. However, for other purposes, we need simply specify the location of the wrist—along with

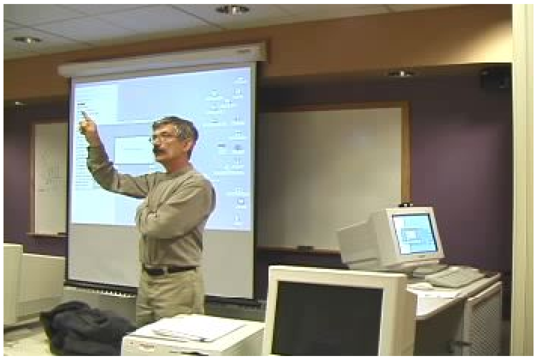


Figure 4.7: Right Arm at (1,4,1)

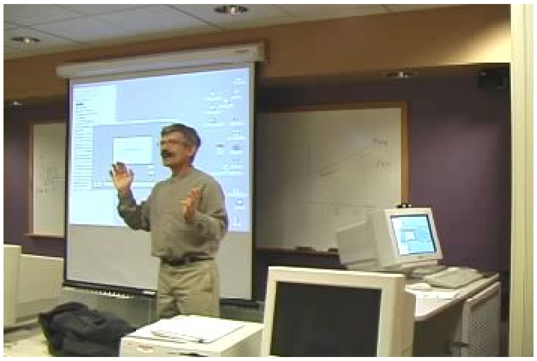


Figure 4.8: Right Arm at (2,4,3)

the upper-arm lift— at key points along the movement. This should suffice for us to recreate the motion. Experiments using this grid system are explained in Chapter 6. To better visualize the grid system, see Figures 4.7 - 4.19. They show a representative sampling of points using the $5 \times 5 \times 5$ coordinates.

4.2 The Corpus

FORM is a corpus of about 22 minutes of Brian MacWhinney teaching a Research Methods course at Carnegie Mellon University. These data were chosen since they were freely available via the TalkBank project (http://www.talkbank.org). They

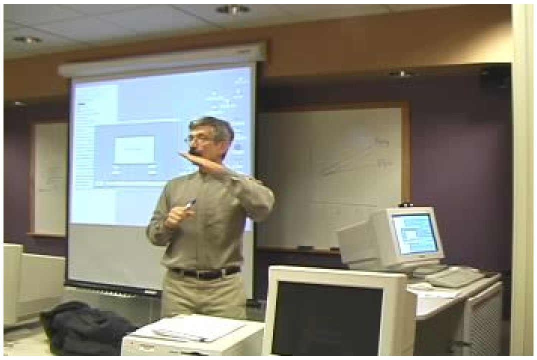


Figure 4.9: Right Arm at (3,2,2)

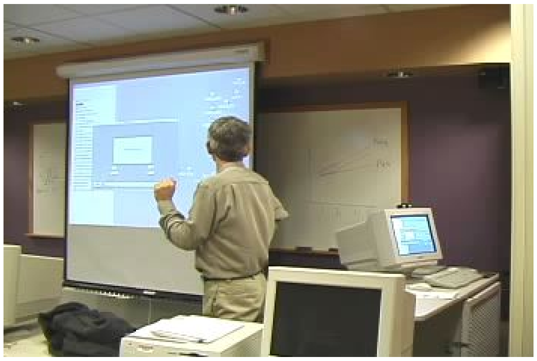


Figure 4.10: Left Arm at (4,3,2)

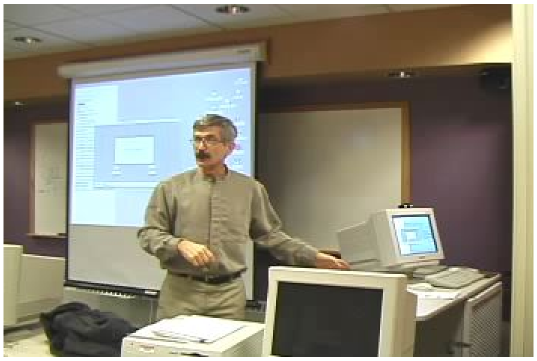


Figure 4.11: Left Arm at (5,1,3)

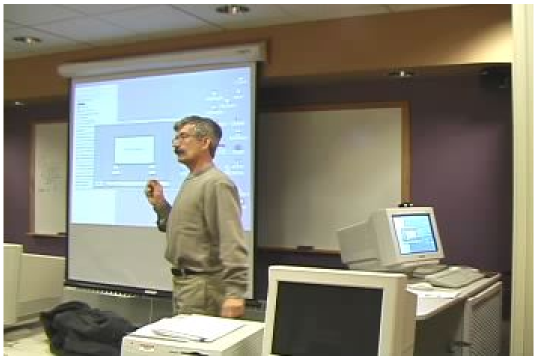


Figure 4.12: Left Arm at (2,1,3)

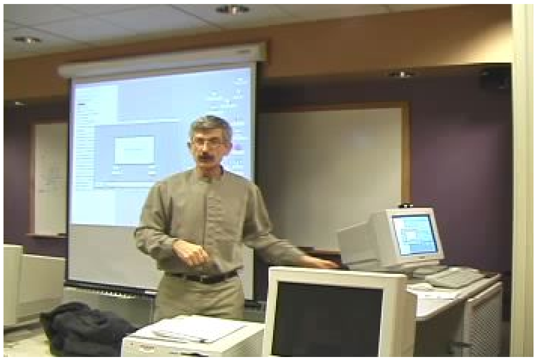


Figure 4.13: Right Arm at (3,2,3)

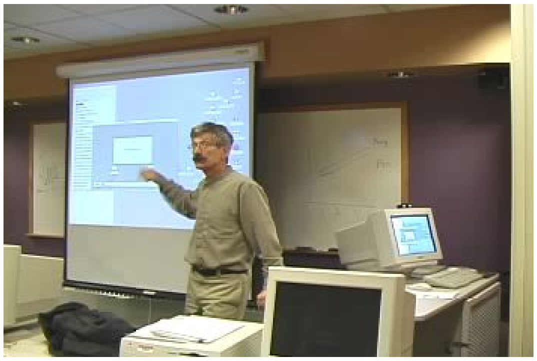


Figure 4.14: Right Arm at (1,3,3)

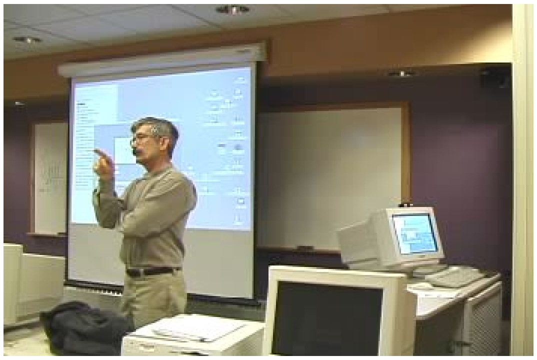


Figure 4.15: Right Arm at (3,4,2)

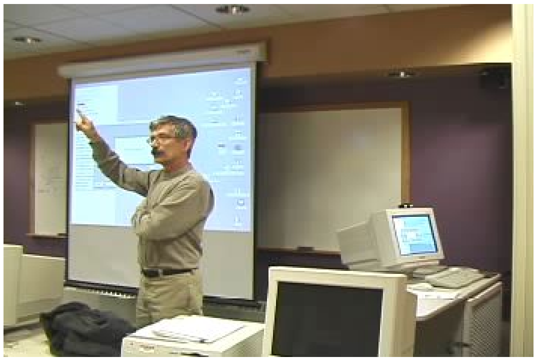


Figure 4.16: Right Arm at (1,5,2)

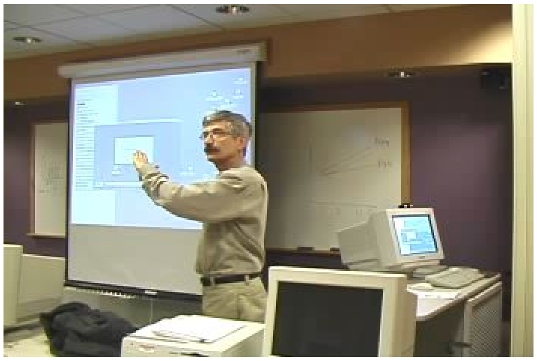


Figure 4.17: Left Arm at (3,4,1)

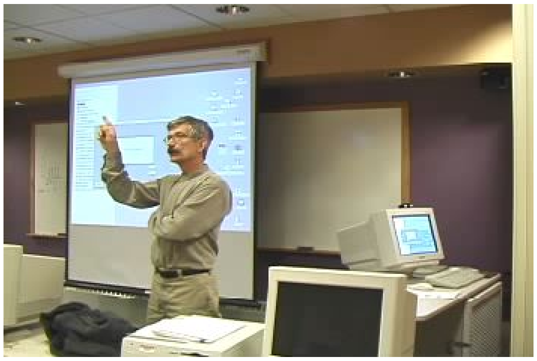


Figure 4.18: Right Arm at (2,4,2)

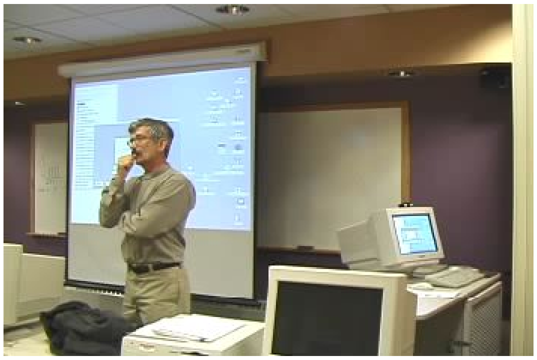


Figure 4.19: Right Arm at (3,4,3)

have been very useful for the project as people often gesture in a clear and exaggerated fashion while teaching.

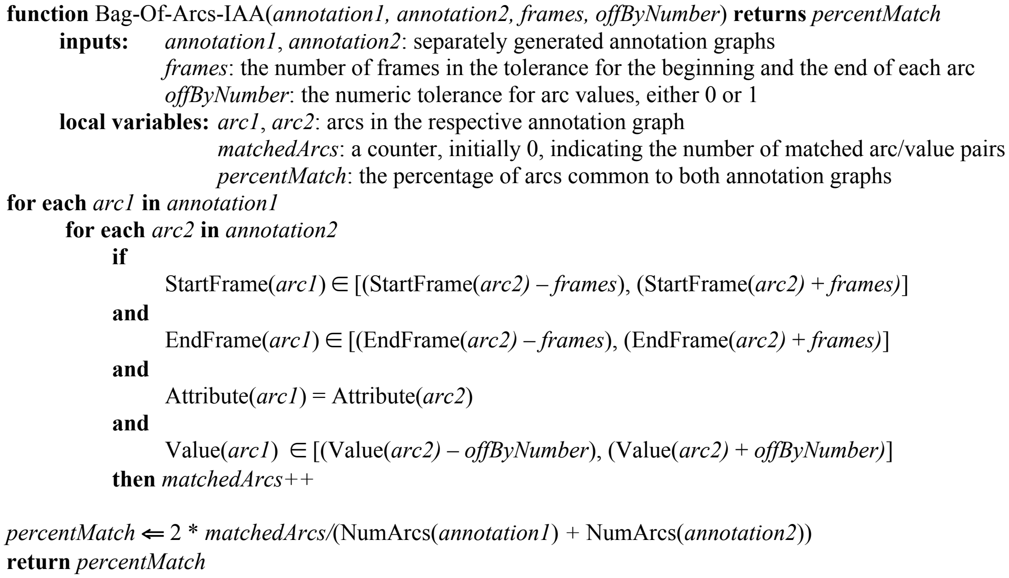
4.3 Annotation Difficulty

An experienced annotator can create approximately 3 seconds of annotation per hour. He/she can annotate for at most 6 hours per day, generating 18 seconds/day. Accordingly, it will take an experienced annotator 5 work days to annotate a 90-second video of conversational interaction.

Generating only 90 seconds of annotation per work week makes such an annotation project seem a daunting task. However, the amount of information contained in conversational gesturing is substantial—on the order of 3500 distinct ATTRIBUTE:Value pairs per minute. This underscores the potential value of such a corpus, viz. there is seemingly much more information in 90 seconds of communicative interaction than we are currently capturing by only transcribing speech.

4.4 Inter-Annotator Agreement Results: Bag-Of-Arcs

Our experiments with FORM-annotation show that with sufficient training, agreement among the annotators can be very high. Table 4.1 shows inter-annotator agreement results for two annotators annotating a file of four gesture-excursions. Table 4.2 shows the results of an intra-annotator agreement study with one of the annotators from the above study annotating the same video at different times. The results were generated by the bag-of-arcs algorithm, as given in Figure 4.20. Essentially, all the arcs for each annotator are combined into a bag. Then all the bags are combined and



```
function Bag-Of-Arcs-IAA(annotation1, annotation2, frames, offByNumber) returns percentMatch
     inputs:     annotation1, annotation2: separately generated annotation graphs
                 frames: the number of frames in the tolerance for the beginning and the end of each arc
                 offByNumber: the numeric tolerance for arc values, either 0 or 1
     local variables: arc1, arc2: arcs in the respective annotation graph
                      matchedArcs: a counter, initially 0, indicating the number of matched arc/value pairs
                      percentMatch: the percentage of arcs common to both annotation graphs
for each arc1 in annotation1
     for each arc2 in annotation2
          if
               StartFrame(arc1) $\in$ [(StartFrame(arc2) – frames), (StartFrame(arc2) + frames)]
          and
               EndFrame(arc1) $\in$ [(EndFrame(arc2) – frames), (EndFrame(arc2) + frames)]
          and
               Attribute(arc1) = Attribute(arc2)
          and
               Value(arc1)  $\in$ [(Value(arc2) – offByNumber), (Value(arc2) + offByNumber)]
          then matchedArcs++

percentMatch $\Leftarrow$ 2 * matchedArcs/(NumArcs(annotation1) + NumArcs(annotation2))
return percentMatch
```

Figure 4.20: Bag-of-Arcs Algorithm

the intersection is extracted. This intersection constitutes the overlap in annotation, i.e., where the annotators agreed. The percentage of the intersection to the whole is then calculated to get the scores presented. Appendix 4 contains the Perl code used to generate these results.

Each of the annotators agreed that there were four gesture excursions. The *Precision* column in the tables gives the number of frames (at 29.97045 fps) that the annotators can be off from one another and still be counted as having agreed. A precision of 0 frames means that the two annotators had to agree on the exact start and end times of an arc in order to be counted as agreeing. Given that it is vague as to exactly where a gesture phase starts and ends[1], we first loosened this restriction to within 7 frames (or approximately ± .25 seconds), and then to

[1] Cf. Durell Bouchard's *WPEII* on the Motion Segmentation Problem.

within 15 frames (or approximately ± .5 seconds). Anything over 15 frames was deemed too tolerant. We also relaxed the algorithm by looking not only at exact matches on the value of an attribute, but also counted as matching any values that were off by no more than one. This is given in the *Off-by-one-or-less* column of the tables. As examples, let *arc1* be $\langle 428, 446, ForearmRotation, 1\rangle$ and *arc2* be $\langle 427, 451, ForearmRotation, 2\rangle$. Then *arc1* will match *arc2* if the tolerances are set to *frames* = 15 and *Off-by-one-or-less* = *Yes.* However, they will not match if *frames* = 0 or if, instead of *Off-by-one-or-less*, *Exact Match* = *Yes.* The bag-of-arcs method is similar to the one used by the IBM BLEU project to judge quality of a machine translation ([88]). The FACS project, mentioned in Chapter 1, also used a similar metric in 1974. There they called two facial encodings a match along a dimension if the first choice of one annotator was the first or second choice of another annotator ([32]).

The most tolerant measure, then, is given by the ⟨15 *frames, Off-by-one-or-less*⟩ cells. Using these cells (in bold face type) as our measure strikes us as fair since, again, both relaxations concern problems that are known to be difficult even for machines, and we are not relaxing too much. Given the above, the results are as following. For inter-annotator agreement the first three excursions have agreement of approximately 80%. The fourth excursion had an agreement of 71.62%. Note, however, that this is not so far off from intra-annotator agreement results. The average for inter-annotator agreement was 77.96%, while the average for inner-annotator agreement was 81.29%. If we were to relax the constraints to 30 frames, our results would be better.

A more interesting analysis, however, entails understanding where the annotators disagree. Appendix D contains all the confusion matrices for all the FORM parameters for the two annotators used in these studies. Here, let us note Tables 4.3—4.7.

Table 4.3 shows the start and end frames for the "three" annotators for all four gesture excursions. Annotator A is the same as annotator D. Annotations B and D were used for the inter-annotator agreement study, while A and D were used for the intra-annotator agreement study. Note that the frames chosen by each annotator are very close. The mean difference is only 2.375 frames. Table 4.4 and Table 4.5 give the confusion matrices for *upperArmLift* for inter- and intra-annotator agreement respectively. Note in both of these case, there is some confusion among the annotators. Although most confusion is close to the diagonal, some is fairly far away. This, however, is not surprising, as humans are not very good at assigning mathematical values. On the other hand, Table 4.6 and Table 4.7, which are the inter- and intra-annotator confusion matrices for RelativeElbowPosition—a "human-level" parameter described above. Here—as in Table 4.3—the agreement is very strong. This suggests that gesture-annotation agreement will increase as we make the parameters more *psychologically* salient and less mathematically precise. Further work needs to be done to explore which parameters can be set with computational assistance, and which work best using human-level annotation.

Finally, let us note that these results are relatively strong. For many difficult, multi-dimensional phenomena, the confusion matrices have a heavy off-diagonal population. For example, [89] is concerned with annotating the Buckeye corpus of conversational speech with phonemeic information. This is a fairly complex task. Table 4.8 gives their labeling consistency. The *% Agree* column is the percentage for which there was agreement between some pair of annotators. The *% Unanimous* is the percentage for which there was unanimous agreement[2]. Although unaimous agreement was fairly low, pairwise agreement was much higher. The two rows labeled *Vowels* are interesting. The former is the agreement across all vowel categories,

[2] The table in the original paper contains other columns, including Kappa score information.

Gesture Excursion	**Precision**	**Exact Match**	**Off-by-one-or-less**
1	0 Frames	44.78	46.77
	7 Frames	64.68	68.66
	15 Frames	74.63	**80.60**
2	0 Frames	29.05	33.94
	7 Frames	61.47	70.64
	15 Frames	70.03	**80.43**
3	0 Frames	41.42	47.34
	7 Frames	47.34	56.81
	15 Frames	63.91	**79.19**
4	0 Frames	40.65	43.23
	7 Frames	59.35	64.51
	15 Frames	64.52	**71.62**

Table 4.1: Inter-Annotator Agreement on Jan24-09.mov

Gesture Excursion	**Precision**	**Exact Match**	**Off-by-one-or-less**
1	0 Frames	0.98	1.96
	7 Frames	82.35	93.13
	15 Frames	82.35	**93.13**
2	0 Frames	13.99	18.03
	7 Frames	53.29	65.40
	15 Frames	63.52	**79.40**
3	0 Frames	25.47	33.02
	7 Frames	39.62	50.94
	15 Frames	57.55	**74.53**
4	0 Frames	13.02	15.98
	7 Frames	57.40	70.42
	15 Frames	62.72	**78.10**

Table 4.2: Intra-Annotator Agreement on Jan24-09.mov

Start	**Annotator A**	**Annotator B**	**Annotator D**
Excursion 1	378	380	380
Excursion 2	1241	1241	1241
Excursion 3	1839	1839	1839
Excursion 4	1946	1963	1947
End	**Annotator A**	**Annotator B**	**Annotator D**
Excursion 1	447	452	452
Excursion 2	1431	1431	1431
Excursion 3	1892	1878	1878
Excursion 4	2092	2092	2092

Table 4.3: Inter- (B and D) and Intra- (A and D) Annotator Agreement on Excursion Start/End Frames.

D **B**	**1**	**2**	**3**	**4**	**5**	**6**
1 (no lift)	1	1	1	0	1	0
2 (0-45)	0	7	4	4	1	0
3 (approx. 45)	0	3	6	4	0	0
4 (45-90)	0	2	4	6	3	0
5 (approx. 90)	0	1	0	0	6	1
6 (90-135)	0	1	0	0	2	3

Table 4.4: Inter-Annotator Agreement for Right Arm UpperArm Location: Upper arm lift

D **A**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1 (no lift	0	2	1	0	0	0	0
2 (0-45)	1	5	1	0	2	0	0
3 (approx. 45)	0	5	7	5	0	0	0
4 (45-90)	0	2	4	7	0	0	0
5 (approx. 90)	0	0	0	1	2	0	0
6 (90-135)	0	0	0	1	5	2	0
7 (approx. 135)	0	0	0	0	0	2	0

Table 4.5: Intra-Annotator Agreement for Right Arm UpperArm Location: Upper arm lift

B \ **D**	**2**	**3**	**4**	**5**	**6**
2 (inward)	0	1	0	0	0
3 (front)	0	6	1	0	0
4 (front-outward)	0	4	10	0	0
5 (outward)	0	0	1	4	0
6 (behind)	0	0	0	0	5

Table 4.6: Inter-Annotator Agreement for Right Arm UpperArm Location: Relative elbow position

A \ **D**	**2**	**3**	**4**	**5**	**6**
2 (inward)	0	1	0	0	0
3 (front)	0	7	3	1	0
4 (front-outward)	0	5	11	0	1
5 (outward)	0	1	1	2	2
6 (behind)	0	0	0	0	5

Table 4.7: Intra-Annotator Agreement for Right Arm UpperArm Location: Relative elbow position

	N	**% Agree**	**% Unanimous**
Overall	2159	80.3	62
Stops	368	92.9	74
Fricatives	507	91.2	76
Nasals	331	87.5	68.5
Liquids	251	86.5	56
Vowels	907	69	49
Vowels (-iu, -ux)	907	73.6	53
Overall (-iu, -ux)	2159	82.2	63.5

Table 4.8: Measures of Labeling Consistency in the Buckeye Corpus

while the latter is the agreement with two particularly difficult categories removed, viz. *ix* and *ux*. For these vowels the *% Agree* values were 17% and 7% respectively. Even less agreement can be found in the results given in [67]. Researchers had judges categorize the emotional intent present in the utterance of a word. Table 4.9 gives the confusion-matrix results. The rows represent the *ground truth*—here generated by subjects uttering the word "Sarah" while simulating the various emotional intentions—while the columns represent how they were labeled by an individual annotator[3]. Even though for most cases the diagonal represents the largest number, the off diagonal numbers are significant.

It is not that surprising that these results are worse than those for both the FORM and Buckeye corpora. In the case of FORM and the Buckeye corpus, the phenomena being annotated were psychologically fairly clear to the annotators. In the case of [67], the annotators were making a judgement of an entire utterance, the constituent parts of which are not necessarily psychologically discernable.

[3] Three randomly-chosen annotators were used to create three versions of Table 4.9. Using chi-square tests, the authors determined that there were no significant differences.

Labeled Simulated	Nam	Com	Ang	Scorn	Cont	Adm	Plea	Sad	Fright	Ast
Naming	28	20	10	9	7	2	11	5	2	6
Commanding	14	28	33	10	2	0	5	2	2	4
Angry	3	20	43	11	1	1	6	1	5	9
Scornful	18	10	14	17	7	9	9	12	1	3
Content	12	4	2	7	10	23	18	17	1	6
Admiring	2	2	2	5	14	27	7	4	1	36
Pleading	10	2	2	6	9	10	30	18	8	5
Sad	9	1	1	4	3	4	21	48	8	1
Frightened	4	1	1	2	6	2	16	6	50	12
Astonished	8	4	3	5	17	7	10	2	3	41

Table 4.9: Confusion Matrix for Emotional-Intent Classification

Chapter 5

Segmentation of Gestures into Phases

In the last chapter, we discussed how we checked the consistency of the FORM dataset. In this chapter, we present an augmentation of the data to include *phase* information. We will then present a new inter-annotator agreement study just for *phase*. The results of this are that Preparation, Stroke, and Retraction appear to be cognitively valid gesture-phase categories.

5.1 Preparation, Stroke, and Retraction

In order to judge the value of the FORM dataset, it is not sufficient to simply judge the inter-annotator agreement. We must also demonstrate that it is useful in some manner. Consistent, but wrong data, doesn't do anyone any good. To demonstrate its usefulness, then, we decided to use the FORM representation of gesture—which is fairly *low-level*—to predict the *medium-level* phenomenon of gesture-phase. This is medium-level in the sense that it is above FORM and below the higher-level phenomenon of *gesture*[1]. We have purposively avoided defining what constitutes an individual gesture in this project, as it is very difficult to clearly pin down the beginning and end of the constituent movements that make up a gesture excursion. Further, there is not yet a theory to describe in what ways these "kinetic" simples should combine to create a gesture. So far, in this work, we have simply picked out the beginning and end of the gesture excursion—viz., rest position to rest position. As we have seen, this is done with suprising consistency. The medium-level phenomenon of phase, however, does not seem to have this trouble. To pick out the phases of an excursion, we do not need to explain which "gesture" they make up. Instead, we only need to segment the excursion and label these segments. It is methodologically much cleaner; and, as we shall see, people do it fairly consistently.

To do this experiment, we added a *Phase* track to both the LeftArm and RightArm Groups of FORM. The annotators segmented the gesture excursion into gesture phases and labeled the phases[2]. Phases were initially of four types: Preparation, Stroke, Retraction, and Hold. Interestingly, though, the annotators were often comfortable claiming there was a phase change, while they were, at the same time, uncomfortable with classifying the new phase. For these cases, we added a fifth type: Unsure. We call the sequence of phases that describe a gesture excursion the *PSR-theory* description, and *PSR theory* the theory that says excursions can be so divided.

As in Chapter 4, a necessary condition for a useful annotation scheme—in this case PSR theory—is consistency; we need to see how well the annotators agreed in their classification of phases. Additionally, as every annonation scheme is a theory

[1] Cf. Chapter 2 for a brief discussion of some of the various schemes used to classify gestures.

[2] The annotators were trained by Adam Kendon as to what constituted each phase type.

about the phenonemon in question, inter-annotator agreement results give us one validation of the theory, particularly if the phenomenon has at least some cognitive components.

5.2 Inter-annotator Agreement for PSR Theory

Our inter-annotator agreement study for PSR theory was done differently than the general FORM agreement study. The reason for this concerns the Unsures. Most of the time, annotators placed an Unsure in the space transitioning between two clear-cut phases. By this, we mean that Unsure served as a way to mark the penumbra between the two phases. In these cases, agreement judged using bag-of-arcs would return very low results. This is because the penumbra between two phases is often larger than 15 frames. This would prevent a match even under the most relaxed conditions. To counter-act this, we did the following. We divided the gesture excursion into frames—each one equivalent in length to the frames of the original video—and labeled each of the frames according to the phase of which it was a part. We then simply judged the degree of agreement on the labels of the frames. So, even if one annotator had a large Unsure between a Preparation and a Stroke while the second annotator had the Preparation directly adjacent to the Stroke, as long as there was enough agreement on which frames made up the Preparation and which frames made up the Stroke, the agreement score would be high. Tables 6.1 and 6.5 present the results of these expriments.

Table 6.1 is particularly interesting. This presents the result of judging agreement over *all* phase categories, including Unsure. Note that the total agreement over all frames was only 68.28%. This low number is largely explained by how Unsures are

	P	**S**	**R**	**H**	**U**
P	701	90	36	30	4
S	57	739	0	0	16
R	0	0	288	3	0
H	5	0	21	313	30
U	169	136	138	290	165

Table 5.1: Agreement: 68.28%

	P	**S**	**R**
P	701	90	36
S	57	739	0
R	0	0	288

Table 5.2: Agreement: 90.42%

used, as described above. The row annotator[3] simply used Unsure much more often. However, we can see that—although there was strong consistency for Preparation, Stroke, and Retraction—there was also some confusion concerning Hold. In particular, the row annotator almost equally divided the column annotator's holds between Hold and Unsure. In other words, the column annotator was more comfortable saying that there was a Hold in between two other phases than the row annotator was. Inspection of the video at this point reveals that in many of these cases the speaker's hand are preforming what we call "incidental movement." Incidental movement is movement during a phase that is, for example, cognitively considered a Hold, even though there is some bouncing or jittery movement of the hand. Some annotators pay attention to the arm as a whole, while others concentrate on the particular part of the body. The latter method could lead to calling this incidental movement an Unsure rather than a Hold. It needs to be explored if further training can help with this phenomenon.

[3] By this we mean the annotator represented by the row labels.

After the above results, we ran the study again, but, for this second run, we only judged agreement on Preparation, Stroke, and Retraction. The results were significantly better. Overall agreement across these three phases was 90.42%.

5.2.1 Using Only Preparation, Stroke, and Retraction

Given the above results, most of the hidden-Markov-model experiments performed in the next chapter only use the Preparation, Stroke, and Retraction categories. We do, however, perform some with Hold added back and some with both Hold and Unsure added back. As can easily be predicted from Table 6.1, these experiments were not as sucessful as those which just used Preparation, Stroke, and Retraction. As holds are presumably important for understanding human gesturing, more work is warranted so that we can consistenly annotated hold phases.

5.3 Comparing PSR Theory to Gesture Theory

As promising as the above PSR-theory-based results are, it is possible that the more cognitively-salient level is the gesture level and not the phase level. [27] presents a study where individual gestures were annotated as belonging to one of the following classes: Deictic, Action, Other, and Unknown.

> Text and video examples were used to instruct participants about the gesture classification scheme. The instructions described both the kinetic [sic] and verbal components of each gesture class. The label "Action" was used in place of "Iconic", since pilot participants found the latter term to be confusing. Similarly, the label "Other" was used to capture "Beat" gestures, as well as any additional gestures that the listener felt did not belong to either of the other two categories. As reported in [20], metaphoric gestures are extremely infrequent in [the corpus used here]. A subset of participants were also allowed to classify gestures as "Unknown."

	Deictic	**Action**	**Other**	**Unknown**
Deictic	.270	.069	.060	.017
Action	.069	.249	.032	.009
Other	.060	.032	.079	.015
Unknown	.017	.014	.015	.004

Table 5.3: Confusion Matrix for Gesture-Level Annotation

Table 5.3 gives the results of the study. It is important to note that the annotators were not required to judge whether or not a gesture occured. A beep was used as an indicator for the start of the stroke of a gesture. The annotator's were only required to judge the class of the gesture that contained the beep-indicated stroke. One of the more interesting aspects of these results is how well annotators did on deictics and actions. These were the only two categories that had positive descriptions. As with the FORM results in the current and prior chapters, labels like "Other" and "Unknown" were more widely used. The first two elements of the diagonal notwithstanding, the off-diagonal cells are significant.

Although much more work needs to be done here, the fact that annotators had a hard time with the *orthodox* gesture typology, combined with the annotation results in Table 5.3, strengthens our view that the gesture-theory level may be too high.

Chapter 6

Hidden-Markov-Model Experiments

In the last chapter, we described how we augmented the FORM corpus to include phase information. In this chapter, we will describe how we use the underlying FORM representation to generate a matrix of vectors for each of these phases. We will then describe how we used this labeled data to run a series of hidden Markov model (HMM) experiments with the goal of *predicting* phase labels from the FORM representation.

6.1 Experimental Overview

As described in Chapter 4, one of the ways we overcame ambiguity in FORM was by adding end-effector position. This position was given as $\langle x, y, z \rangle$ coordinates in a $5 \times 5 \times 5$ grid. If we combine these coordinates with the value of the *upperArmLift* parameter, we get a vector in $\mathbb{R}^4$ which describes the position of an arm at a particular frame. So, a sequence of these vectors encode the movement of an arm through-out a gesture excursion. If we divide the excursion into sub-sequences of these vectors such that they are co-extensive with the phase segmentation described in the last chapter, we have created a set of labeled data.

However, as described in Chapter 3, FORM annotators only put *Location* markers at critical points in the gesture. The goal was to approximate zero-crossings in the first and second derivatives. In order to create the requisite interpolated vectors, then, we take the $\mathbb{R}^4$ vectors for each *Location* point in the gesture excursion and utilize various interpolation methods to fill in the values for the intervening frames. This generates a large matrix in $\mathbb{R}^4$, the number of columns of which is determined by the number of frames—at 29.97045 fps—in the excursion. We then divide this large matrix in accordance with the phase segmentation to generate bins of matrices representing the different phases. When we are done, we have a bin of preparations, a bin of strokes, a bin of retractions, a bin of holds, and a bin of unsures.

For each of the various interpolation methods, we then run the hidden Markov model experiment described in Figure 6.1. It is a version of a cross-validation method known as *Leaving-one-out* ([83]). For each iteration of the experiment the training set is of size $N-1$, while one data point, i, is used as held-out testing data. This process is repeated N times so each data point gets left out once[1]. Our particular algorithm works as follows. Of the combined set of *all* phase matrices—which we will call *observations*—choose one, $observation_i$, at each iteration and remove it from the set of observations. Then, for each of the sets of phases Preparation, Stroke, Retraction, etc, generate a hidden Markov model representing that phase and train with all the

[1] It is important to note that this method has both advantages and disadvantages. An advantage is that it allows for exploration of how the model changes for any particular piece of data. In addition, it is useful for doing cross-validation when the total number of data points is low, as is the case with the current FORM dataset. On the other hand, Chen and Goodman ([13]) argue that using larger deleted chunks gives better results. Here, we should consider this method as giving something like an upper bound, as it is dangerously close to testing on the training data. As the size of the FORM dataset grows, we should be able migrate to safer methods.

samples for that phase only. Label $observation_i$ after the hidden Markov model, M, which maximizes $P(observation_i|M)$. If the label generated for $observation_i$ matches the actual label of $observation_i$, call it a match. Finally, return $observation_i$ to the set of observations. We do this for all i. Our total percentage of matches is computed as 100 × (total matched/total number of *observations*).

The following sections present the results of these experiments. Appendix F provides an overview of hidden Markov models.

6.2 Baseline

6.2.1 Call-all-*X*

In all of the experiments present in this chapter, the baseline used is Call-all-x. Actually, Call-all-x is a combination of multiple baselines—one per phase—that produces particularly conservative results. For each of the phases, x, in the experiment, we assumed an algorithm that labels all observations as x. For example, the Call-all-Prep baseline labels every observation as a preparation. Precision is calculated simply as the proportion of actual preparations in the dataset. Recall will always be 1. *Mutatis mutandis* for all other phases.

These baseline results are conservative since the recall score for each phase is 1, which will drive up the baseline f-score for each phase. The important point here is that the high recall is *not* at the expense of precision. If it it were, then the f-score (Equation 6.1), being simply a special case of the harmonic mean (Equation 6.2), would be lower. The harmonic mean has the very nice property that you are punished if you privilege one dimension at the expense of another. To see this, first compute the f-score with $p = .5$ and $r = .5$. You will see that the answer is .5, i.e., the same

```
function Leave-One-Out($phaseSet_1, . . ., phaseSet_n$) returns percentage of correct classifications
    inputs: $phaseSet_1, . . ., phaseSet_n$: a list of sets of phase matrices,
                                          e.g. prepSet, strokeSet, retractionSet
    local variables: observation: the current held-out phase matrix
                     $M_i$: the HMM for $phaseSet_i$, e.g. $M_{stroke}$ is the HMM trained on strokeSet
                     match: a counter, initially 0, indicating number of correct classifications
                     percentMatch: a number $\in [0,1]$, indicating percentage of correct
                                 classifications
for each observation in $\{phaseSet_1 \cup \ . . . \cup \ phaseSet_n\}$
    for each $phaseSet_i$ in ($phaseSet_1, . . . , phaseSet_n$)
        $M_i \Leftarrow$ createHMM($phaseSet_i$)
        if $observation \in \ phaseSet_i$ then
            Train($M_i$) using $phaseSet_i - \ observation$
        else
            Train($M_i$) using $phaseSet_i$

    PredictedLabel(observation) $\Leftarrow$ argmax$_i$ P(observation | $M_i$)
    if PredictedLabel(observation) = actualLabel(observation) then
        match++

percentMatch $\Leftarrow$ 100 * (match/total number of observations)
return percentMatch
```

Figure 6.1: Leave-one-out Training Algorithm using One HMM per Phase

as the arithmetic mean. However, if you assume some change to your algorithm that pushes up recall to 1 while pushing down precision to .25, the harmonic mean is lower than the arithmetic mean. That is, the arithmetic mean would be .625, while the f-score would be .4.

Finally, please note that sometimes baseline numbers are different across experiments using the same data set. This is an artifact of the HMM system we used[2]. Depending on how a particular HMM was trained, testing does not always complete. That is, the test observation may contain too few frames for one or more of the testing HMMs to return a probability. In these cases, GT2K/HTK simply returns an answer of "too few frames" and the observation is not labeled. However, to compare the results of experiments with different baseline numbers, we can simply look at the percent difference from the baseline f-score a particular algorithm is at predicting the phases. This is what is reported in the ±***Baseline*** column of Tables 6.1–6.16.

$$\text{F-Score} = \frac{2}{\frac{1}{P} + \frac{1}{R}} \tag{6.1}$$

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{6.2}$$

6.2.2 Maximum Likelihood: Call-all-Stroke

As described above, we choose Call-all-x for our baseline, because we wanted to compare the results of the three phases independently across experiments. However, this has the effect of skewing the results against Stroke. This is because Stroke is the most common category in all of our datasets. Another common baseline we could have used would be to label all phases after the most common category: maximum

[2] We used HTK ([107]) and the Georgia Tech Gesture Toolkit ([105]) to build and train the HMMs.

likelihood. In our case, this is equivalent to simply using Call-all-Stroke as our baseline and comparing its precision and recall numbers against the total precision and total recall across all three categories. Again, though, this collapses the results of all three categories into one set of numbers. In order to be thorough, and to test against a more common baseline, we did recalculate our results against the maximum likelihood baseline. These results are presented in Appendix G[3]. They produce numbers that are slightly scaled, but the trends described in this chapter are exactly observed against the maximum-likelihood baseline as well.

6.3 PSR vs PSRH vs PSRHU

In Chapter 5, we saw that the annotation of the Unsure category was done inconsistently across the annotators. Additionally, there was a particular confusion between the Hold and Unsure categories. One would expect, then, that running *Leaving-one-out* using all five categories would produce lower results than running it with the Unsure left out. Further, one would expect that experiments using just Preparation, Stroke, and Retraction would produce the best results. As the experiments in Tables 6.5–6.2 show, these predictions turn out to be the case. For all of these experiments, we interpolated using cubic splines and vector quantized using the k-means algorithm with $k = 1000$.

Table 6.1 gives the results of running *Leaving-one-out* using all phases—Preparation, Stroke, Retraction, Hold, *and* Unsure—as inputs. As expected here, the Unsure category did very poorly with respect to the baseline f-score, as did Stroke. The other

[3] NB: For almost every experiment in this chapter, a maximum likelihood baseline is calculated and the results presented in Appendix G. Excepted from this are the "unbalanced" experiments described in Section 6.4.2. These are excepted as they are explicitly designed to explore ways to do better on one phase, Stroke, for example, at the expense of the others. Using the maximum-likelihood baseline would prevent us from seeing any desired results.

categories did better than baseline or were essentially the same. The poor results are not surprising given Table 6.2, the confusion matrix for this experiment. Both Unsure and Hold caused a lot of off-diagonal confusion.

Table 6.3 gives the results of running the experiment without the Unsure category. Here we see that removing Unsure increases our results as expected. From Table 6.4, though, we can see that Hold is still creating a large amount of confusion. Again, this is due to the inconsistency of the annotation. As mentioned in Chapter 5, training annotators better on how to deal with incidental movement may help with this.

Finally, Table 6.5 gives the results of just using Preparation, Stroke and Retraction. Table 6.6 shows that this experiment has the least amount of confusion. Although, preparations are still often labeled as Stroke.

6.4 First Experiments

In this section, we present the results of our first experiments using FORM. These are all location-based experiments and are essentially the experiments described in Section 6.1. We are calling them location-based experiments, because each vector of the phase matrices represents an arm location for that frame. This section is divided into balanced and unbalanced experiments. The balanced experiments attempt to maximize precision and recall for all three phases, while unbalanced experiments try to maximize precision at the expense of recall. The unbalanced experiments were done since, for some scientific purposes, we are far more concerned that the things we say are x actually are x.

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.15	1.00	0.26	
PSHRU	0.38	0.66	0.48	+85%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.30	1.00	0.46	
PSRHU	0.46	0.49	0.47	+2.2%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.11	1.00	0.20	
PSRHU	0.49	0.56	0.52	+160%
	Hold			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Hold	0.20	1.00	0.33	
PSRHU	0.43	0.26	0.32	-3.3%
	Unsure			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Unsure	0.25	1.00	0.40	
PSRHU	0.34	0.24	0.28	-30%

Table 6.1: Precision, Recall, and F-Score Results for *S1000* on the *Brian* Data Set using Preparation, Stroke, Retraction, Hold and Unsure

Prediction **Truth**	**Preparation**	**Stroke**	**Retraction**	**Hold**	**Unsure**
Preparation	69	22	6	2	5
Stroke	47	102	6	17	35
Retraction	5	3	42	9	16
Hold	24	38	14	36	24
Unsure	36	59	18	19	41

Table 6.2: Confusion Matrix for S1000-PSRHU

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.19	1.00	0.32	
PSRH	0.45	0.67	0.54	+69%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.40	1.00	0.57	
PSRH	0.58	0.60	0.59	+3.5%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.14	1.00	0.25	
PSRH	0.65	0.69	0.67	+168%
	Hold			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Hold	0.26	1.00	0.41	
PSRH	0.53	0.27	0.36	-12%

Table 6.3: Precision, Recall, and F-Score Results for *S1000* on the *Brian* Data Set using Preparation, Stroke, Retraction, and Hold

Prediction **Truth**	**Preparation**	**Stroke**	**Retraction**	**Hold**
Preparation	70	26	6	2
Stroke	53	125	8	21
Retraction	8	5	52	10
Hold	25	60	14	37

Table 6.4: Confusion Matrix for S1000-PSRH

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.27	1.00	0.43	
PSR	0.50	0.68	0.58	+35%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.54	1.00	0.70	
PSR	0.80	0.61	0.69	-1.4%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.19	1.00	0.32	
PSR	0.72	0.82	0.77	+141%

Table 6.5: Precision, Recall, and F-Score Results for *S1000* on the *Brian* Data Set using Preparation, Stroke, and Retraction

Prediction **Truth**	**Preparation**	**Stroke**	**Retraction**
Preparation	71	26	7
Stroke	63	127	17
Retraction	8	6	62

Table 6.6: Confusion Matrix for S1000-PSR

6.4.1 Balanced Experiments

The results of the balanced experiments are given in Table 6.7. All of the experiments are versions of *Leaving-one-out* with the differences being in the interpolation and vector-quantization methods used. In all of these experiments, the *UpperArmLift* parameter is simply linearly interpolated.

- Fixed-Grid: In this experiment we first linearly interpolated between the points given in the data. Then, we vector quantized by labeling each vector within a cube of the $5 \times 5 \times 5$ grid, e.g, $\langle 1, 1, 1 \rangle$, by the name of that cube. To do this, we simply rounded the results of the linear interpolation. So, for example, $\langle 1.235, 1.45, 1.75 \rangle$ becomes $\langle 1, 1, 2 \rangle$.

- L500: For this experiment, we first linearly interpolated between the points given in FORM, and then vector quantized using the fast k-means algorithm given in [34][4]. In this case $k = 500$.

- L1000: This experiment was the same as above except $k = 1000$.

- S500: In this case, we first used *spline3()* function from Matlab 7.0 to generate cubic splines between the FORM points. We then vector quantized with $k = 500$.

- S1000: As above, but $k = 1000$.

- SnoVQ: In this case, we used *spline3()* to generate cubic splines, but utilized all the vectors as given. That is, we did not vector quantize.

[4] All vector quantizing for these experiments was done using the Matlab code available at http://www.cse.ucsd.edu/users/elkan/fastkmeans.html. We found it to be orders of magnitude faster than standard k-means.

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.27	1.00	0.43	
Fixed-Grid	0.45	0.63	0.53	+23%
L500	0.50	0.64	0.56	+30%
L1000	0.46	0.67	0.55	+28%
S500	0.45	0.62	0.52	+21%
S1000	0.50	0.68	0.58	+35%
SnoVQ	0.47	0.69	0.56	+30%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.54	1.00	0.70	
Fixed-Grid	0.78	0.60	0.68	-2.9%
L500	0.79	0.61	0.69	-1.4%
L1000	0.78	0.59	0.67	-4.3%
S500	0.78	0.58	0.67	-4.3%
S1000	0.80	0.61	0.69	-1.4%
SnoVQ	0.79	0.59	0.68	-2.9%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.19	1.00	0.32	
Fixed-Grid	0.76	0.83	0.79	+147%
L500	0.67	0.81	0.73	+128%
L1000	0.77	0.79	0.78	+144%
S500	0.70	0.80	0.75	+134%
S1000	0.72	0.82	0.77	+140%
SnoVQ	0.77	0.81	0.79	+147%

Table 6.7: Precision, Recall, and F-Score Results for balanced HMM Methods Using the *Brian* Data Set

6.4.2 Unbalanced Experiments

The unbalanced experiments we did are as follows. The results are in Table 6.8. For all of them we used cubic-spline interpolation.

- S1000.50: This is the same as S1000, above, but we added a measure of uncertainty. If the difference between the log-probability of the most likely model and the second most likely model was greater than 50% of the difference between the most likely and the least likely, we deemed the labeling to be *uncertain*. The 50% mark was chosen empirically to give the best results. We did this in the hope of increasing our precision, even at the expense of recall. As is evident, we did raise precision, but only from 0.5 to 0.51, while recall dropped from .68 to .38.

- P.25.500: For this experiment, we explored adding context from the previous phase, in the hopes that more context would increase the results. The *P* in the title indicates that we *prepending* context from the prior phase to the current phase. The *25* indicates that we are prepending 25% of the prior phase. The 500 indicates that we vector quantized with $k = 500$.

- P.25.1000: As above, but $k = 1000$.

- P.25.1000.373: As above, but if the difference between the highest log-prob and the second-highest log-prob was .373 or greater of the difference between the highest log-prob and the lowest log-prob, we called the labeling *uncertain*. Again, this number was empirically chosen to maximize precision.

- A.25.500: In this experiment, we explored *appending* 25% of the following phase to the current phase. The idea is to use context from later in the gesture. Again, vector quantization was done with $k = 500$.

- A.25.500.148: As before, we appended 25% of the following phase to the current phase; and, if the difference between the highest and second highest log-probability was .148 or greater of the difference between the highest and lowest log-probability, the labeling was deemed *unsure*. As before, .148 was determined empirically to maximize precision.

- A.25.1000: Appended 25% of the following phase and vector quantized with $k = 1000$.

- A.25.1000.35: Same as above, but the difference cut off was empirically set to .35.

6.4.3 First Results

It is interesting to note that retractions seemed the easiest to classify, with its highest F-score being .79 for the Fixed-Grid, balanced method. If the stroke is the most definite aspect of the gesture, we would have expected it to be the easiest to classify. Retractions do have the easily identifiable characteristic of ending in a rest position. As well, there is usually only one per excursion; there are many preparations and strokes per excursion. Additionally, although the first preparation of an excursion starts from a rest position, the subsequent ones do not. We were able to use the difference-cut-off technique in the unbalanced section to increase the precision of stroke recognition, but it was at great expense to recall. A.25.500.148 allowed for a precision of .92, but a recall of .18. Although this may not be useful for an automatic phase detector, it could be very useful for scientific exploration of strokes. It would give high assurance that those things we automatically identified as strokes were actually strokes. One more thing of note here is that, although we were able to

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.27	1.00	0.43	
S1000.50	0.51	0.38	0.44	+2.3%
P.25.500	0.44	0.41	0.42	-2.3%
P.25.1000	0.35	0.25	0.29	-33%
P.25.1000.373	0.48	0.29	0.36	-16%
A.25.500	0.49	0.60	0.54	+26%
A.25.500.148	0.50	0.41	0.45	+4.7%
A.25.1000	0.45	0.49	0.47	+9.3%
A.25.1000.35	0.49	0.29	0.36	-16.3%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.54	1.00	0.70	
S1000.50	0.92	0.06	0.11	-84%
P.25.500	0.74	0.65	0.69	-1.4%
P.25.1000	0.69	0.73	0.71	+1.4%
P.25.1000.373	0.90	0.18	0.30	-57%
A.25.500	0.74	0.59	0.66	-5.7%
A.25.500.148	0.92	0.18	0.30	-57%
A.25.1000	0.72	0.57	0.64	-8.6%
A.25.1000.35	0.83	0.20	0.32	-54.3%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.19	1.00	0.32	
S1000.50	0.87	0.61	0.72	+125%
P.25.500	0.15	0.56	0.24	-25%
P.25.1000	0.50	0.66	0.57	+78%
P.25.1000.373	0.54	0.71	0.61	+90%
A.25.500	0.30	0.69	0.42	+31%
A.25.500.148	0.30	0.69	0.42	+31%
A.25.1000	0.49	0.80	0.61	+91%
A.25.1000.35	0.53	0.72	0.61	+91%

Table 6.8: Precision, Recall, and F-Score Results for unbalanced HMM Methods Using *Brian* Data Set

use the segmentation "tricks" described above to increase certain statistics, the best overall was simply using cubic-splines and $k = 1000$. Further, this was not that much better than the simplest, Fixed-Grid method that simply linearly interpolates and rounds the x, y, and z coordinates to the nearest integer value. This is clearest in the graphs given in Figures 6.2–6.4, where the first five entries are all fairly similar in shape and value.

6.5 McNemar's Test

Of all the experiments above, S1000 performed the best overall. However, the important question of whether or not the results are significant needs to be addressed. Given that all of the above experiments labeled the same pieces of data, and given that there is not a strong difference among the results of the various experiments, McNemar's test is very useful. It is essentially a modified sign test. Sign tests are used when you can divide your results into positive cases and negative cases. They are fairly simplistic and are very useful as sanity checks. By this, we mean that if a more sophisticated statistical test says there is a significant difference, while a sign test says there is not, then it is time to rethink your sophisticated test.

For McNemar's test we create a 2×2 table that represents *correctness* and *error* agreement between two experiments. In the $\langle$Correct, Correct$\rangle$ cell of the table is the number of items that both experiments labeled correctly. In the $\langle$Error, Error$\rangle$ cell is the number of items that they both labeled incorrectly. These are the uninteresting cells. The interesting cells are $\langle$Correct, Error$\rangle$ and $\langle$Error, Correct$\rangle$. These are the numbers of items that were labeled correctly by one test and incorrectly by another. McNemar's test becomes a sign test, then, simply by stating that one of these latter two cells is called *positive* and one is called *negative*. It doesn't matter

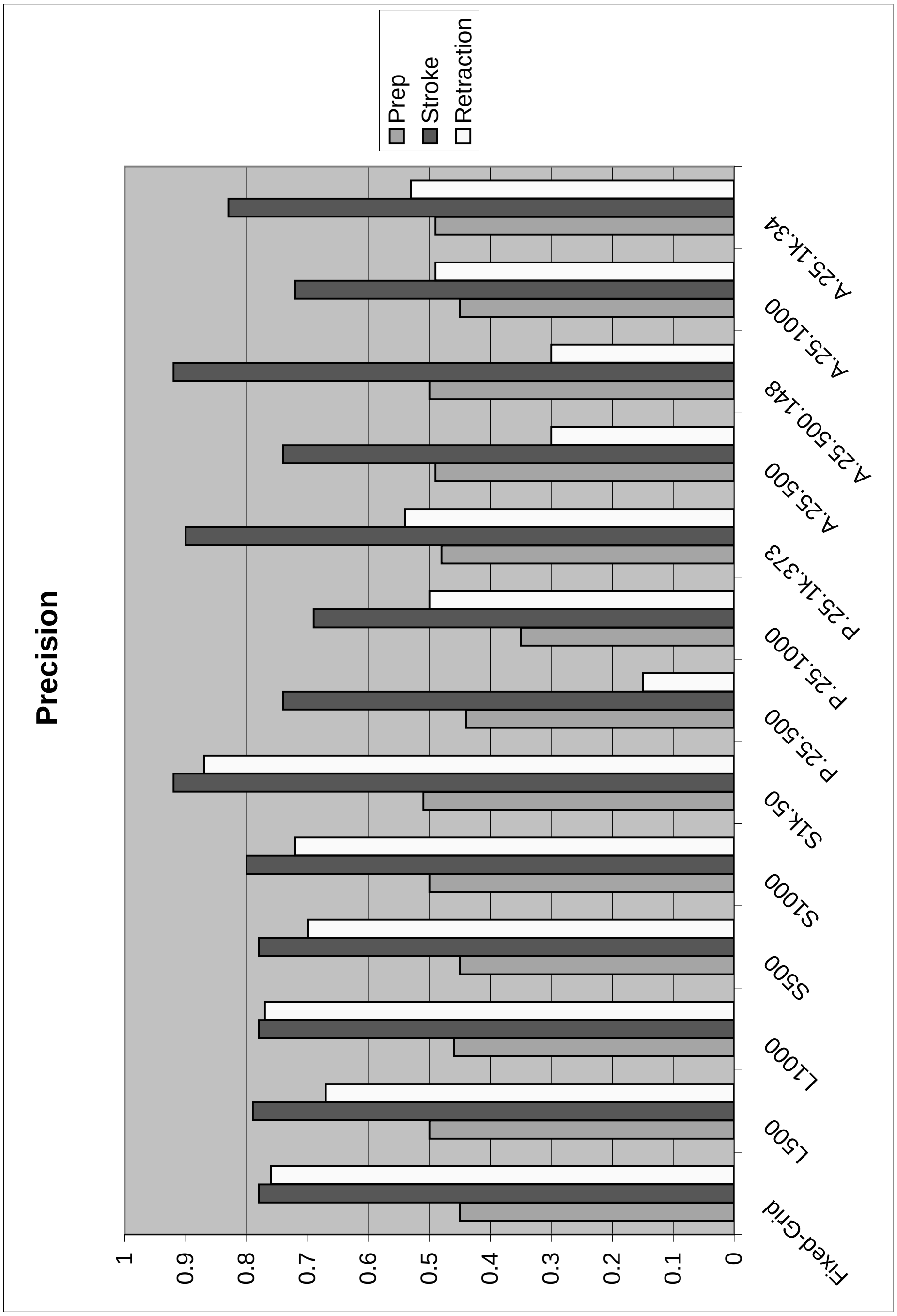


Figure 6.2: Hold-one-out Precision-score Results

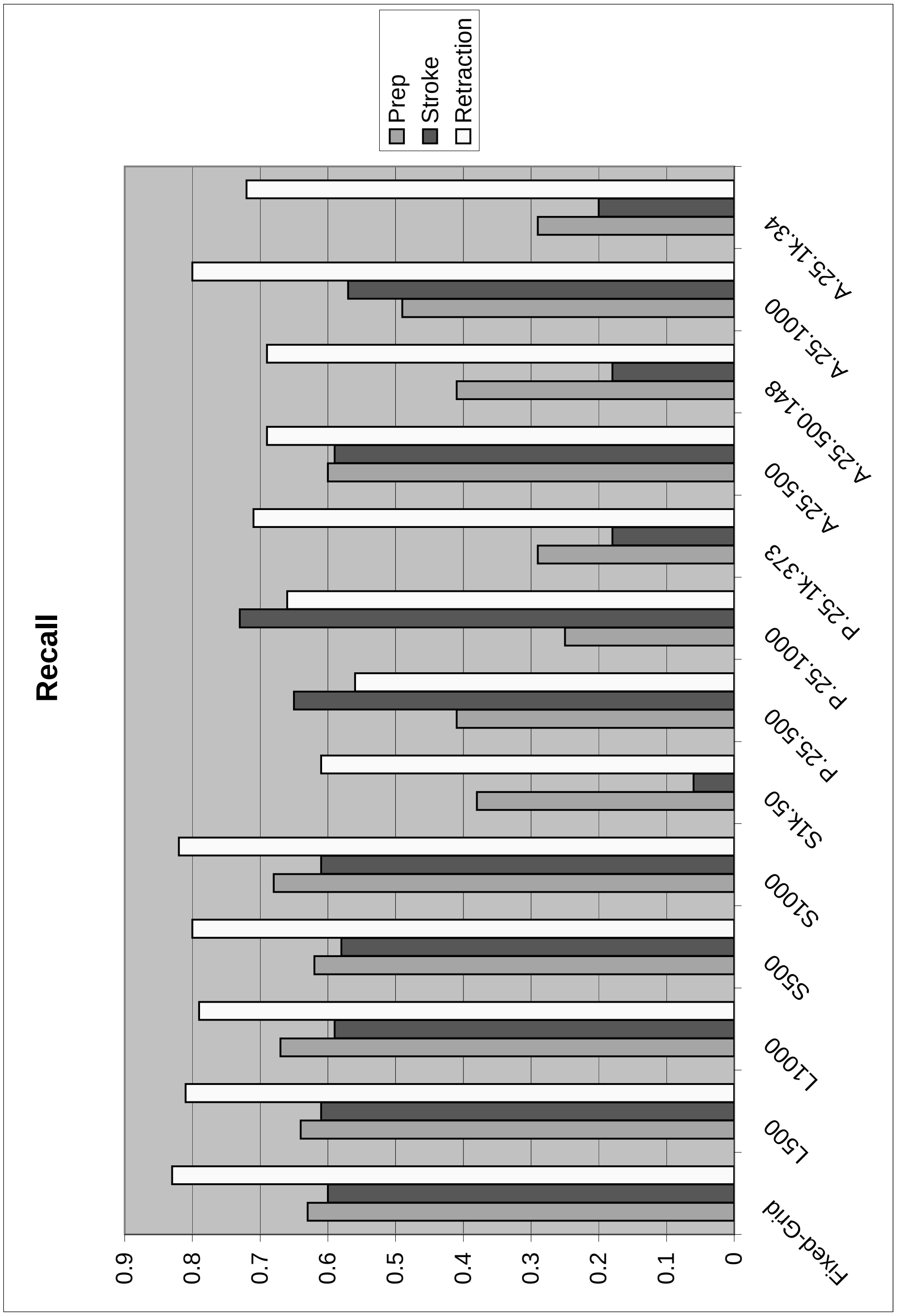


Figure 6.3: Hold-one-out Recall-score Results

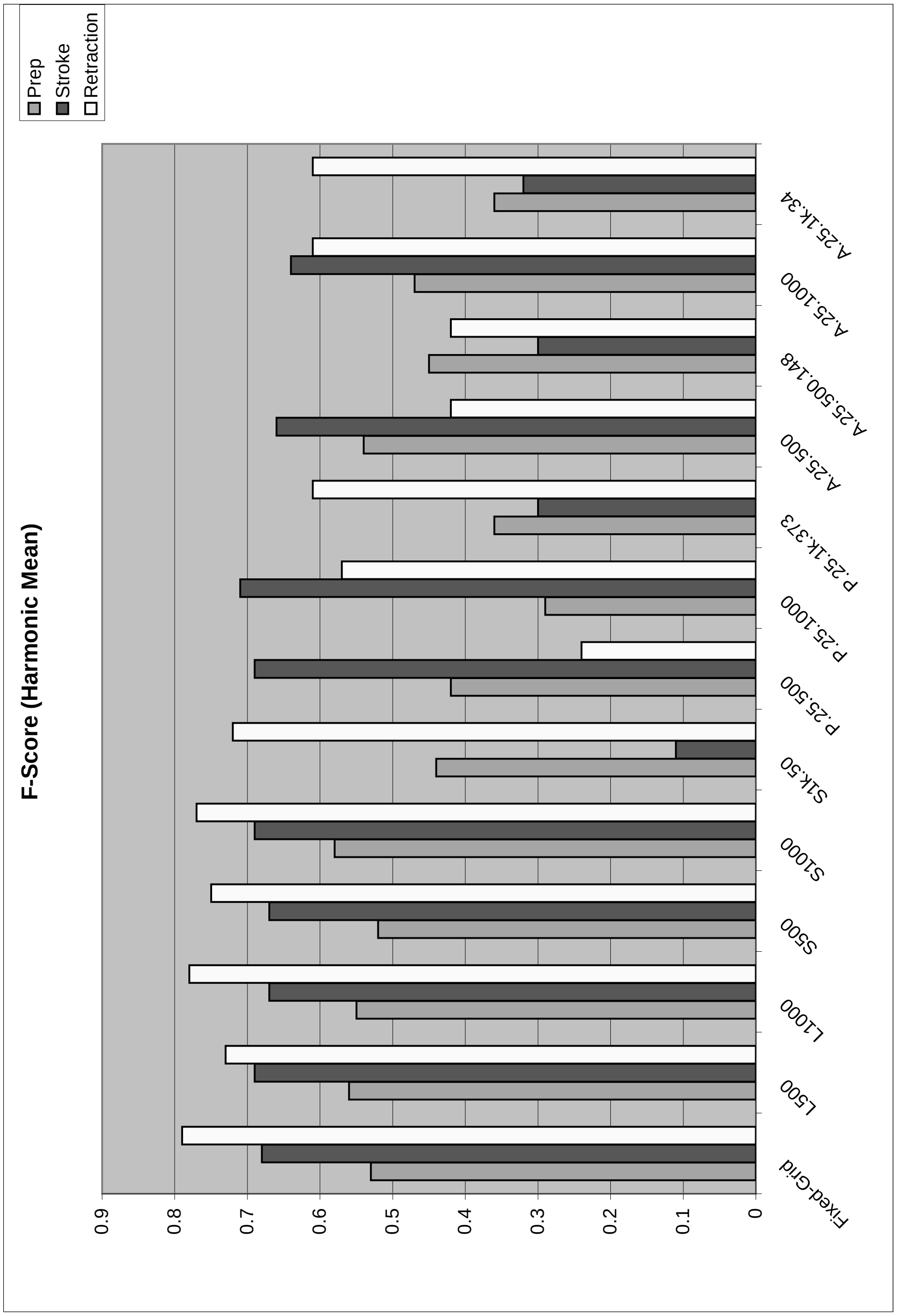


Figure 6.4: Hold-one-out F-Score Results

which. The null hypothesis, H_0, for the McNemar's test is that $\langle$Correct, Error$\rangle$ = $\langle$Error, Correct$\rangle$, or that both experiments would make the same number of mistakes.

Appendix E contains the results of running McNemar's Test on all pairs from the balanced experiments above. There was no significant between the different methods except for between S1000 and S500. Table 6.9 shows the results of this test. The *p-value* is two-tailed and is computed using the following formula:

$$P \leq 2 * \sum_{i=0}^{k} \frac{\frac{N!}{i!(N-i)!}}{4}, \tag{6.3}$$

where

$k = min\{\langle Error, Correct\rangle, \langle Correct, Error\rangle\}$,

and

$N = \langle Error, Correct\rangle + \langle Correct, Error\rangle$.

We can see that the differences among the two experiments for each of the phases is not significant, although for Preparation it is close. However, the overall the difference is significant.

6.6 Dual Data: FORM vs Motion Capture

The results described above may or may not be of interest depending on how well FORM compares to more precise ways of gathering gestural data. Additionally, the experiments above only concern one subject, Brian MacWhinney. In order to address these issues we built another data set for comparison: the *Craig* data set. It comprises approximately three minutes of Craig Martell in lecturing mode discussing his teaching methods. The data for this set were gathered in two ways: motion-captured and video-recorded. The video recordings were than annotated using FORM. This

Preparation	**Correct**	**Error**	**p-Value**
Correct	64	7	0.07
Error	1	32	
Stroke	**Correct**	**Error**	**p-Value**
Correct	112	15	0.3
Error	9	71	
Retraction	**Correct**	**Error**	**p-Value**
Correct	58	3	1.0
Error	2	12	
All	**Correct**	**Error**	**p-Value**
Correct	234	25	0.049
Error	12	115	

Table 6.9: S500 (along the top) vs. S1000 (down the side)

data set, then, allows us to compare FORM to motion-capture *vis-á-vis* prediction of preparations, strokes, and retractions. It also allows us to compare the prediction results of two FORM datasets of different speakers in similar situations.

Tables 6.10–6.14 give the results of these experiments. Again, they are on different tables because their baselines were sometimes different[5]. They are described below in order of their listing in the tables.

6.6.1 Location-Based Experiments

Our first set of experiments on this dataset were, as in all experiments so far, location-based. These were as follows.

- S1000-Craig: This experiment is the same as the original S1000 experiment for *Brian*. The frames between the location points were interpolated using cubic splines and then vector quantized to 1000 vectors.

[5] Cf. Section 6.2.

- SnoVQ-Craig: This experiment is the same as the original SnoVQ for *Brian*. It is just as above but without the vector quantization.

- mocap1000: For this experiment, we utilized a subset of the 32 motion-capture marker points to generate end-effector position for each arm, as well as *upperArmLift*. This created vectors in $\mathbb{R}^4$ analogous to those used in FORM. We then vector quantized to 1000.

- mocapNoVq: Same as above but without vector quantizing.

- simulatedFORM: In this experiment, we generated a set of key frames in the motion-capture data, performed cubic-spline interpolation, and vector quantized to 1000. The key-frames were chosen so as to match the location points given by the FORM annotation. This experiment was done to see if the location information of FORM combined with the fidelity of motion-capture could increase results.

Results

The first result of note is that for *Craig*, both FORM-annotated and motion-captured, the SnoVQ experiment did better than the S1000 experiment. This is likely due to the fact that the *Craig* data set has only three minutes worth of gesturing, as opposed to 20 minutes worth for the *Brian* data set. In the later case, the total number of *raw* vectors, so to speak, is much greater than with the *Craig* data. This would account for the benefits gained from using vector quantization, since, as the number of training vectors increases, the chance of only seeing any particular vector once—space data—also increases. Additionally, Brian MacWhinney is teaching using a white board. This results in more movements away from the solar plexus than for Craig

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.35	1.00	0.52	
S1000-Craig	0.65	0.50	0.57	+5.8%
SnoVQ-Craig	0.67	0.50	0.57	+5.8%
mocap1000	0.56	0.45	0.50	-3.8%
mocapNoVQ	0.61	0.49	0.54	+3.8%
simulatedFORM	0.37	0.91	0.53	+1.9%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.45	1.00	0.62	
S1000-Craig	0.72	0.69	0.70	+13%
SnoVQ-Craig	0.72	0.73	0.72	+16%
mocap1000	0.64	0.53	0.58	-6.5%
mocapNoVQ	0.69	0.60	0.64	+3.22%
simulatedFORM	0.25	0.01	0.02	-9.7%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.20	1.00	0.33	
S1000-Craig	0.58	0.86	0.69	+109%
SnoVQ-Craig	0.61	0.86	0.71	+115%
mocap1000	0.46	0.78	0.58	+76%
mocapNoVQ	0.46	0.76	0.57	+73%
simulatedFORM	0.41	0.22	0.29	-12%

Table 6.10: Precision, Recall, and F-Score Results for Various HMM Methods Using the *Craig* Data Set

Martell, who is only addressing an audience. Vector quantization is useful under these circumstances as it creates equivalence classes of vectors with a representative vector for each. Each of these representatives (the k means) will be seen on average $\frac{N}{k}$ times, where N is the total number of distinct vectors in the data set. On the other hand, if the total number of raw vectors does not contain too many singletons, then the fact that vector quantizing collapse information is relevant.

The second interesting result is that FORM did better than motion-capture in predicting phase labels. Look at SnoVQ-Craig vs mocapNoVQ. The former does 5.8% better than Call-all-Preparation, while mocapNoVQ only does 3.8% better. This is not a large difference, but now consider Call-all-Stroke. Here SnoVQ-Craig did 16% better than baseline, while mocapNoVQ only did 3.2% better. The largest difference is for Call-all-Retraction. There the gain over baseline was 115% and 73% respectively. Table 6.11 gives the result of the McNemar's test for SnoVQ-Craig vs mocapNoVQ. Again, it is the case that although there is not a significant difference for any one phase, the difference overall is significant.

These results are contrary to expectation. One would think that the "physical truth" given by motion-capture would do better at predicting a mid-level *physical* phenomenon like phase than would a cognitively-laden annotation scheme. These results, then, may indicate that phase prediction is more cognitively laden than was originally thought. One possible explanation for this was mentioned in Chapter 5. There we explored the phenomenon of incidental movement as a possible explanation of why there was confusion between Hold and Unsure. If it is the case that holds are considered holds even when there is incidental movement, then something similar may be the case for other phases as well. For example, there may be some subset of the parameters of a stroke that humans use to classify the movement, and the smoothing of the curve done by the FORM method may better approximate these

Prep	**Correct**	**Error**	**p-Value**
Correct	35	16	1.0
Error	15	35	
Stroke	**Correct**	**Error**	**p-Value**
Correct	57	31	0.09
Error	18	22	
Retraction	**Correct**	**Error**	**p-Value**
Correct	41	10	0.18
Error	4	4	
All	**Correct**	**Error**	**p-Value**
Correct	133	57	0.05
Error	37	61	

Table 6.11: mocapNoVQ (along the top) vs. SnoVQ-Craig (down the side)

parameters. Much more work needs to be done here, but incidental movement is a known difficulty for gesture and phase prediction[6].

Simulating FORM by using the location points given by annotators to pick out key-frames in the motion captured data, however, did not work very well. If the above theory is correct, though, it should have. That is, if the reason for FORM's doing better than motion-capture is simply the smoothing of the curve and the removal of incidental movement, then simulatedFORM should have achieved this. A possible answer is that the FORM key-frames smooth further by picking out very *chunky* locations in space. The motion-capture time-stamp at some form location frame, i, will pick out a much more precise region of space. The FORM method reduces a large number of paths from $location_i$ to $location_{i+1}$ to just one, further reducing potential sparse-data problems.

[6] Cf. [92] for pointers.

6.6.2 Motion-Based Experiments

The location-based experiments given in Section 6.6.1 allow us to compare FORM to motion-capture. Additionally, however, the existence of the motion-captured, "gold-standard" data allows us to explore the relationship of motion to phase in a way that we cannot do with FORM data alone; the motion-capture data simply has a much higher fidelity. Further, if features of movement other than location are more predictive, knowing what they are would be instructive for building an enhanced version of FORM. Accordingly, we ran the following experiments using estimates of end-effector velocity, $\hat{v}$, and end-effector acceleration, $\hat{a}$, as our primary motion features[7]. They are defined as follows.

$$\hat{v}_i = \frac{l_i - l_{i-1}}{t_i - t_{i-1}} \tag{6.4}$$

and

$$\hat{a}_i = \frac{\hat{v}_i - \hat{v}_{i-1}}{t_i - t_{i-1}}, \tag{6.5}$$

where l_i is the end-effector location (in $\mathbb{R}^3$) for frame i and t_i is the time-stamp for frame i.

- 1zcKeyFramesXYZ: For this experiment, we determined the frames which represent zero-crossings in the first derivative and used these as key-frames. They were determined by looking for sign changes between $\hat{v}_i$ and $\hat{v}_{i-1}$. There was no interpolation between the frames; a phase was defined simply by its key-frames. The XYZ indicates that only location of the end-effector was used.

[7] Much of this work follows that given in [108], although the equations we use are slightly different. [108] also presents experiments which use *curvature* and *torsion* as important motion features. We ran a set of preliminary experiments using these features on *Craig*, but the results were disappointing.

- 1zcKeyFramesXYZTh: This experiment is the same as the one above with the exception that both end-effector location and *upperArmLift* were used.

- 2zcKeyFramesXYZ: For this experiment, we used as key-frames those which represent zero-crossings in the second derivative. They were determined by looking for sign changes between $\hat{a}_i$ and $\hat{a}_{i-1}$. As above, there was no interpolation between these frames and only the end-effector position was used.

- 2zcKeyFramesXYZTh: As above, but using end-effector location and *upperArmLift*.

- 6deltas: In this experiment, we generated a vector for each frame by computing the Δx, Δy, and Δz between frames i and $i-1$. Additionally, we computed $\Delta\Delta x$, $\Delta\Delta y$, and $\Delta\Delta z$ for each frame. For the first frame the Δs were zero, and for the first and second frames the $\Delta\Delta$s were zero. This produced a 6-element vector per frame containing only motion information.

- 9deltas: This experiment was the same as above with the addition of the x, y, and z location coordinates to each vector. The end result is a 9-element vector representing both location and motion information.

Results

The motion-based experiments in this section are somewhat out of place in this thesis, as they do not relate directly to the current version of FORM. Instead, as aforementioned, we ran these to see if motion parameters may be a useful addition to FORM—at least with regard to phase detection.

The general result here, interestingly, is that the location experiments did better; this is for both the FORM and motion-capture ones. The results, in order of success

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.40	1.00	0.57	
1zcKeyFramesXYZ	0.46	0.38	0.42	-26%
1zcKeyFramesXYZTh	0.57	0.40	0.47	-18%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.36	1.00	0.53	
1zcKeyFramesXYZ	0.55	0.64	0.59	+11%
1zcKeyFramesXYZTh	0.58	0.64	0.61	+15%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.24	1.00	0.39	
1zcKeyFramesXYZ	0.52	0.53	0.52	+33%
1zcKeyFramesXYZTh	0.47	0.63	0.54	+38%

Table 6.12: Precision, Recall, and F-Score for 1zcKeyFrames for *Craig* Data Set

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.38	1.00	0.55	
2zcKeyFramesXYZ	0.65	0.34	0.45	-18%
2zcKeyFramesXYZTh	0.57	0.30	0.39	-29%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.39	1.00	0.56	
2zcKeyFramesXYZ	0.58	0.63	0.61	+9%
2zcKeyFramesXYZTh	0.58	0.61	0.59	+5%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.23	1.00	0.37	
2zcKeyFramesXYZ	0.46	0.76	0.57	+54%
2zcKeyFramesXYZTh	0.47	0.79	0.59	+59%

Table 6.13: Precision, Recall, and F-Score for 2zcKeyFrames for *Craig* Data Set

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.35	1.00	0.52	
6deltas	0.45	0.63	0.53	+2%
9deltas	0.48	0.62	0.54	+4%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.45	1.00	0.62	
6deltas	0.66	0.45	0.54	-13%
9deltas	0.63	0.47	0.54	-13%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.20	1.00	0.33	
6deltas	0.37	0.36	0.36	+9.1%
9deltas	0.45	0.47	0.46	+39%

Table 6.14: Precision, Recall, and F-Score Results for 6deltas and 9deltas for *Craig* Data Set

per phase, are given in Table 6.15. No one method was best overall, but taking a combination of the best per phase, we get motion-based methods doing 4% better than baseline on Preparation, 15% better on Stroke, and 59% better on Retraction. Compare this to the best of all the location methods, SnoVQ: +5.8% for Preparation, +16% for Stroke, and +115% for Retraction.

It is clear that that much more experimentation is needed to find the correct motion parameters for gesture phases. Further experiments should include *jerk*, i.e. the third derivative. Jerk measures the rate of change in concavity/convexity. The rate of change in jerk, the 4th derivative[8] may, in fact, contribute to our ability to predict gesture phase as well. Further exploration along these lines, however, will require a new way of generating the data; cubic splines are not guaranteed to be continuous in the 4th derivative.

[8] The 4th derivative is sometimes referred to as snap, while the 5th and 6th derivatives have jokingly been called *crackle* and *pop*!

Prep	**±Baseline**
9deltas	+4%
6deltas	+2%
1zcKeyFramesXYZTh	-18%
2zcKeyFramesXYZ	-18%
1zcKeyFramesXYZ	-26%
2zcKeyFramesXYZTh	-29%
Stroke	**±Baseline**
1zcKeyFramesXYZTh	+15%
1zcKeyFramesXYZ	+11%
2zcKeyFramesXYZ	+9%
2zcKeyFramesXYZTh	+5%
6deltas	-13%
9deltas	-13%
Retraction	**±Baseline**
2zcKeyFramesXYZTh	+59%
2zcKeyFramesXYZ	+54%
9deltas	+39%
1zcKeyFramesXYZTh	+38%
1zcKeyFramesXYZ	+33%
9deltas	+9.1%

Table 6.15: Motion-based results, Ordered by Best per Phase

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.42	1.00	0.59	
S1000-Paul-Lec	0.67	0.61	0.64	+8.5%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.44	1.00	0.61	
S1000-Paul-Lec	0.70	0.62	0.66	+8.2%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.13	1.00	0.23	
S1000-Paul-Lec	0.37	0.64	0.47	+104%

Table 6.16: Precision, Recall, and F-Score Results S1000 for *Paul-Lecture* Data Set

6.7 More Data: *Paul*

The experiments in the prior sections have primarily cut across two dimensions. We first looked at FORM location-based experiments and then compared them to motion-captured location-based experiments. Secondly, we compared location-based experiments to motion-based experiments. In this section we will look at two more comparisons: FORM location-based experiments for multiple-subjects in an analogous contexts and FORM location-based experiments for the same subject in different contexts. We do this to check that changes in subjects or changes in contexts do not radically affect our results.

In order to do these experiments, we first created a third FORM data set, *Paul*. It contains roughly six minutes of Paul Howard; for the first three Paul is lecturing, for the second three Paul is having a conversation with someone off camera. S1000 was run using each of these subsections of the data. The results are given in Tables 6.16 and Tables 6.17.

	Prep			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.45	1.00	0.62	
S1000-Paul-Con	0.63	0.65	0.64	+3.2%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.43	1.00	0.60	
S1000-Paul-Con	0.64	0.58	0.61	+1.7%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.12	1.00	0.21	
S1000-Paul-Con	0.56	0.68	0.61	+190%

Table 6.17: Precision, Recall, and F-Score Results S1000 for *Paul-Conversation* Data Set

Results

The results of this comparison are given in Table 6.18. We see that they follow the same pattern for all three subjects. Retractions are the easiest to classify, followed by preparations and, finally, strokes. Although strokes for Brian were below the baseline, the difference is very small. This can, again, be accounted for by the fact that Brain was lecturing to a class and, therefore, using the white board quite a bit. This increased both the location and shape ranges of his strokes. The most interesting thing here is that the 20 minutes of data in *Brian* didn't do as well as the 3 minutes of *Craig* or the 3 minutes of *Paul*. Brian's preparations may be cleaner than either Paul's or Craig's. Craig's strokes were the most easily picked out and, therefore, may also be cleaner. Further research should be conducted to discover the characteristic of the cleanest members of each phase. We should then explore whether these prototypes can be used in learning algorithms. For example, we may label a phase based on which prototype it is closest to.

Table 6.19 compares the results of the two sets of *Paul* experiments. *Paul/*

Prep	**±Baseline**
S1000 (Brian)	+35%
S1000-Craig	+5.8%
S1000-Paul-Lec	+8.5%
Stroke	**±Baseline**
S1000 (Brian)	-1.4%
S1000-Craig	+13%
S1000-Paul-Lec	+8.2%
Retraction	**±Baseline**
S1000 (Brian)	+140%
S1000-Craig	+109%
S1000-Paul-Lec	+104%

Table 6.18: *Brian* vs *Craig* vs *Paul*

Conversational didn't do as well at predicting preparations or strokes as did the *Paul/Lecturing*. This could simply be caused by the fact that Paul's gestures were, subjectively judged, more exaggerated when he was lecturing than in conversation. Retractions were predicted better from the conversational data.

6.8 Summary

In this chapter we have presented a series of experiments designed to serve as a verification of the FORM method for gathering gestural data. Although, at this point, these results are preliminary, we have seen that FORM can be used to predict the mid-level annotation of PSR theory at least as well as the baseline when using cubic splines to interpolate. And, depending on the person and the amount of data collected, it may sometimes be better to vector quantize and sometimes not. Furthermore, these results hold for the cross-context experiment as well. Finally, it turns out that motion-capture does not do as well at predicting phases as does

Prep	**±Baseline**
S1000-Paul-Lec	+8.5%
S1000-Paul-Con	+3.2%
Stroke	**±Baseline**
S1000-Paul-Lec	+8.2%
S1000-Paul-Con	+1.7%
Retraction	**±Baseline**
S1000-Paul-Lec	+104%
S1000-Paul-Con	+190%

Table 6.19: *Paul in Conversation* vs *Paul Lecturing*

FORM, nor do motion-based experiments do as well as location-based ones. There is still too little data at this point to generate a full theory. However, these first results look promising.

Chapter 7

Conclusion: Contributions and Further Work

In this thesis we have presented an annotation scheme that allows for gestural information to be gathered simply from videos of subjects in communicative interaction. FORM is designed to be extensible, portable, and useful for both engineering and science. Additionally, we built a dataset useful for pilot studies using corpus-based statistical methods applied to gesture. We also conducted an experiment in the annotation of the PSR-level of gesture, the results of which serve as a preliminary indication that the decomposition of gestures into phases is cognitively valid. Finally, we showed that FORM—as a representation of the kinematics of gesture—is useful for predicting PSR-level phenomena. This suggests that information at the kinematic level, although difficult to obtain, has value both for science and engineering. However, although these results are promising, much more work needs to be done in order to more fully validate our approach. We need more data and better, faster ways of gathering it. In the short run, we may revert to motion-capture. However, as we discovered, data gathered in this way may not be as useful as FORM-annotated data for some purposes.

7.0.1 A Word about Levels

Throughout this theses we have referred to low-, mid-, and high-level analyses of gesture; FORM and Motion Capture are low-level, LMA[1] and PSR-Theory are mid-level, and BEAT[2] is high-level. It is important to note that at this point we mean this only as a description of the levels of abstraction. We do not yet have enough information to fully articulate a hierarchical theory such that it is clear how a higher-level description is deterministically reduced to a lower-level description.

To make this distinction clearer, let's assume our goal is to find a reduction from PSR to FORM, for example, that is analogous to the reduction of biology to chemistry. In the latter, a particular biological description has associated with it one description—or clearly defined set of descriptions—at the chemical level. As importantly, this (set of) chemical description(s) maps to only one biological description. In other words, the descriptions at the two levels are not orthogonal. For gesture theory, however, our levels are not yet so cleanly related to each other. Even given that case that we can "reduce" a particular BEAT-level description to a PSR-level description, it is not a reduction in the same sense as described above. A given PSR-level description can be mapped to many BEAT-level descriptions, and a particular BEAT-level description can be mapped to many different PSR-level descriptions. For a complete theory of gesture, we would hope that this orthogonality would not be

[1] Depending on how you slice it, so to speak, LMA can also be seen as a high-level theory. On such a model, though, there would probably be only LMA at the high-level and a description of the kinematics of movement at the low-level, with no intervening mid-level description.

[2] By BEAT here, we mean an analysis in terms of the type of gesture being produced, i.e., a beat, a deictic, a metaphoric, or an iconic. I hope the makers of the BEAT toolkit[10] are not offended by our co-opting the use of "BEAT" in this way.

the case. This thesis, then, serves as a potential starting point for such a reduction—both in that it provides for useful low-level descriptions and in that it presents some beginning experiments towards such a reduction. To move forward, we need much more data and many more experiments.

7.0.2 Different Levels for Different Uses

Although much more research needs to be done here, it seems likely that a completed theory of gesture would have different levels of analysis being useful for understanding different dimensions of gesture. For example, knowing the different types of gestures at a BEAT level makes it much easier for animators. Similarly, someone interested in analyzing the qualitative aspects of gestural movement might well choose Laban Movement Analysis. To test whether or not there is a correlation between the frequency of strokes and emotional state, we would need to use the PSR-level description. Our long-term hope for FORM is that, regardless of the level of analysis needed for a particular task, data gathered at the FORM level will be useful. Again, however, much more work is needed to show that we can infer these respective higher levels from FORM data.

7.1 Contributions

Although a large number of experiments have been conducted for this thesis, we believe that the contributions lie mostly in the annotation practices presented, the inter-annotator agreement metric, bag-of-arcs, we developed, and the data sets themselves. The experiments do serve to validate the approach, though. Our contributions, then, are:

- **The FORM Annotation System**: Chapter 3 presents the full FORM annotation system. Appendix A gives the best practices discovered by the FORM annotators, while Appendix B presents stills from the training videos in the *FORM Video Code Book*. Further, Chapter 4 presents an augmentation of FORM to include a $5 \times 5 \times 5$ grid for end-effector location. The work presented here is not just an extension of prior work. It is a first in that it is both low-level—and therefore highly descriptive—and computationlly useful.

- **Bag-of-Arcs Metric**: Chapter 4 presents a new metric for measuring inter-annotator agreement. In this thesis it is applied to data stored as Annotation Graphs, but it is applicable for any time-dependent data that can be represented as n-tuples.

- **The FORM Data Sets**: For this thesis we built a gesture-annotated data set of over 22 minutes of Brian MacWhinney teaching. This data set was validated using the aformentioned bag-of-arcs technique. It was found that the agreement was sufficiently higher than for other high-dimensional linguistic phenonmena. Further, we built two smaller data sets—one of Craig Martell and one of Paul Howard—that are useful for comparisons both across subjects and contexts. Additionally, the *Craig* data set contains both motion-captured and FORM-annotated data. This will allow for further research into the pros and cons of each method of data gathering.

- **Experimental Verification of PSR Theory**: As part of the experiments in this thesis, we needed to further augment all of the aforementioned corpora with PSR-theory annotations. As presented in Chapter 5, we then validated the quality of these annotation by measuring inter-annotator agreement. However, since phases are "fuzzy" at the edges, we did not use the bag-of-arcs method.

Instead, we simply judged the proportion of frames that were annotated the same by both annotators. The results were just over 90%. As every annotation scheme is a theory, these results serve as a preliminary study on the cognitive validity of PSR theory.

- **Prelimiary Experiments using FORM**: Finally, in Chapter 6 we presented a large set of experiments which demonstrated the usefulness of FORM for both science and engineering. Although, no one of these results did significantly better than our conservative baseline, Call-all-x, the set serves an excellent starting point for those interested in gesture annotation at a kinematic level.

7.2 Further Work

This thesis has laid the ground work for a more mature gesture-annotation field. However, in many places through out this thesis we have expressed the need for more research. Those areas that are the most important are listed here.

- Although the FORM annotation method produces strong inter-annotator-agreement results, it is still remarkably time consuming. To build an empirical science of gesture, as well as to build corpus-based algorithms for gesture analysis, we need data, data and more data. Fundamental to doing this, though, is building better tools and methodologes for gathering the data. Considering the results in Chapter 6, including that FORM is more predictive of the PSR-level than is motion-capture, the default assumption of most researchers—that motion-capture is the best way to gather gesture data—may need to be rethought.

- Further work is also needed for better undedrstanding which aspects of gesture are the most predictive. The aforementioned results notwithstanding, motion features certainly may play an important role here. To this end, further research into which aspects of motion are the most useful is needed.

- The inability of FORM to predict the PSR-level better than it did may actually be due to the machine learning techniques used. Further research on other techniques is warranted. Our experiments only used sequential hidden Markov models. An exploration of other HMM typographies, e.g., ergodic models, is needed. As well, we need to explore other types of statistical models. We may find that multiple models are needed simultaneously to capture all the different aspects of gesture simultaneously.

- Zhao ([108]) went a long way towards characterizing different aspects of gestural motion. These characterizations included parameterizations of the gesture space that were completely different from the location-based and motion-based parameterizations presented in this thesis. In particular, [108] is concerned with using Laban Movement Analysis parameters as a way to characterize gesture. Further FORM experiments should include adding this parameterization to the corpora, discovering whether it is more predictive of *phase* than FORM, and, if so, exploring the correlations between LMA and FORM. In particular, we see LMA as generating equivalence classes of gestural movements. It would be useful to see if the characterizations of these classes are predictive and whether there is a consistent, underlying kinematic description for each class.

- Finally, to further research the relationship between the kinematics of gesture and various other aspects of the communicative interaction, we need to augment the FORM corpora with annotations of these other dimensions. The first step here is to add time-aligned transcriptions of the speech that accompanies gesture, followed by part-of-speech tagging of these transcriptions. This will allow us to look for correlations between speech and gesture—a necessary stop on the road to understanding what gestures mean.

Appendix A

FORM: Best Annotation Practices

By Kari Myers, Lisa Britt, Paul Howard, and Chris Osborn

A.1 Techniques for Viewing and Controlling the Video

- Always look at gestures both frame-by-frame and also in real-time. It is possible to confuse one movement for a similar movement if it is only viewed at one speed. For example, a "wrist circular" hand movement looks very similar to a "down-up" movement combined with a "toward thumb-toward little finger" movement when only viewed frame-by-frame. However, the circular motion is apparent when the clip is played at normal speed.

- If helpful, look at the video in full screen; motion can be lost when looking at a small version of the video.

A.2 Techniques for Looking at the Speaker

- Use reference points, shadows and blurs to help determine where movement starts, stops and changes direction.

- Often it can be difficult to determine whether a movement results from actual arm movement, or simply a change in body position. If a frame of video is blurred around the hand or arm, it signifies that there is indeed movement. Likewise, inferences about movement can be made from the relative position of a shadow to the body or the background. This can help in determining not only if there is movement, but also a change in the direction of movement.

- The point where the hand or arm begins to blur is often the point where it starts to move.

- Changes in the wrinkles of clothing can signify movement.

- Much like blurs and reference points the clothing of the subject can give indications to aid in determining the movement that is occurring.

- If the subject's clothing is moving around the elbow, then it signifies elbow movement. Likewise, changes in the size and/or number of creases or folds around the shoulder would signify upper arm movement. This is most helpful in distinguishing which part of the arm is exerting the effort in a gesture.

- Be careful to look at body parts separately.

- While all parts of the arm work together, it is easy to lose upper arm motion and instead attach the movement to the forearm, or lose forearm movement to wrist movement, for example. One way to account for this is to cover the part of the arm that you are not looking at.

- Similarly, be careful to distinguish between wrist movement and small forearm rotations. They often occur simultaneously, but it can be difficult to distinguish when the rotation is small.

A.3 How to Use the FORM Scheme

- When choosing a hand shape, it is often difficult to find an exact match. In this case, tension may compensate for the difference. For example, a loose fist could be annotated as a relaxed version of a tight fist if it is the most similar shape.

- Compare how far your wrist bends towards your thumb with how far it bends towards your little finger. Use extremely towards little finger in cases where the wrist is bent further towards the little finger than it could bend towards the thumb.

- In the two-handed gestures track, only one of the Boolean values (interlaced, moving in parallel, moving apart etc.) can be chosen per object.

- Hand position is annotated as a location track.

- When hand position is not changing, a multi-frame object should be created to indicate position.

- Remember that it is possible for other parts of the arm to be moving without hand position changing.

- Single frame objects are used to indicate changes in direction.

- This includes the peaks of curved movements.

- It is sometimes necessary to use more than one object at different points within one continuous motion in order to properly capture the path of the motion.

- Often the hand will move within one box, so consecutive objects will have identical attributes. Annotate this as motion rather than unchanging location. If you wish to divide the body using different increments at a later time, the necessary objects will already exist at the proper timestamps.

A.4 Annotation Techniques

- Use your own body for reference.

- If the movement is ambiguous due to video quality or body position, it may be helpful to mimic the subject's movement. Performing the movement can help determine the precise action that is taking place and give an additional perspective to aid in annotation. When you cannot see a movement clearly, always mark it obscured to indicate your uncertainty.

- The best point of reference for a particular part of the arm is the joint connected to it. When annotating upper arm movement, looking at the elbow's position will show the upper arm lift. Likewise, to determine forearm orientation, it is most helpful to look at the direction the palm and wrist are facing.

- Use multiple strokes whenever possible.

- If part of a gesture is repeated in exactly the same manner, annotate it as one movement track and use the "strokes" value to indicate the number of times the action is repeated.

- If you have trouble deciding on values involving angles (upper arm lift, elbow flexion), make a "cheat sheet" with examples of several different angles (e.g. 45, 90, 135, and maybe a few in between), and keep it handy while you annotate. You can do this easily with a pen and a plastic protractor.

- When selecting a value for effort, it is often helpful to consider the amount of movement compared to the amount of time.

- A small movement that spans an entire second requires much less effort than a large one that spans a few frames.

- Viewing the video in real-time may help to make this decision.

- It can be difficult to visualize the dividing line between two segments. In this case, a small degree of relativism may be appropriate when making the judgment. For example, if the hand is in the "Right" segment of the X dimension, and it moves more to the right, but you are not sure if it is still "Right" or "Far Right," using the latter option is more likely to capture the motion properly.

Appendix B

Transcript and Stills of the *Video Code Book*

To allow for a better understanding of the process of annotation, we present here the complete transcript of the FORM *Video Code Book*, a best-practices video tutorial, along with representative stills. The actual *Video Code Book* is available at `http://www.cis.upenn.edu/~cmartell/`.

Excursion Duration

The *Excursion Duration* track is used to mark the beginning and end of a gesture excursion, which is defined as all gesturing occurring between two rest positions.

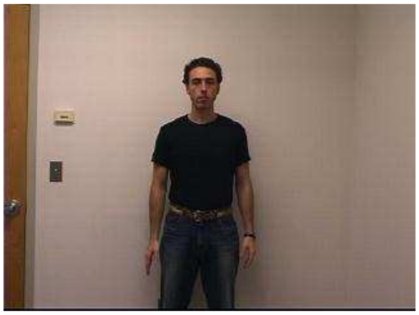
A rest position is a position requiring minimal effort to hold the arms stationary. Examples of rest positions would be:

folding of hands;

arms crossed;

hands on hips;

or arms directly at the side.



It is important to note that a gesture excursion may contain multiple, individual gestures.

Upper Arm

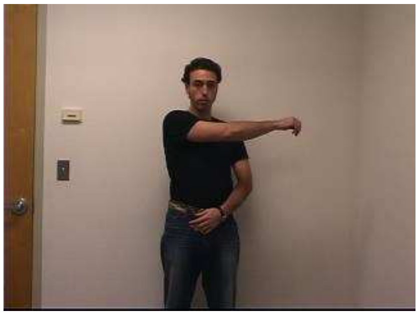
The upper arm location and movement tracks are used to record gesture information pertaining to the part of the arm between the elbow and the shoulder {indicate by pointing}.

These tracks are identical for the Right and Left arms.

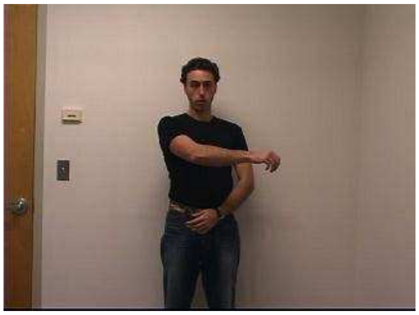
The *Upper Arm.Location* track contains the following attributes:

The “relative elbow position” describes the elbow’s position in relation to the body. The possible values are:

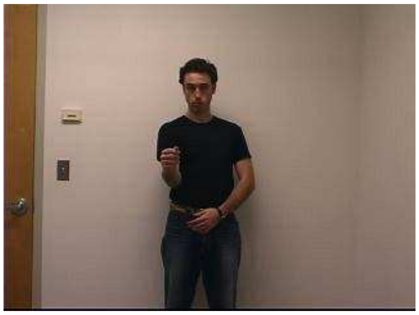
extremely inward;



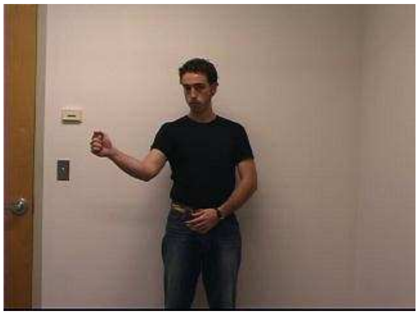
inward;



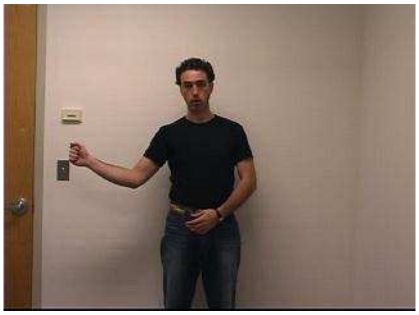
front;



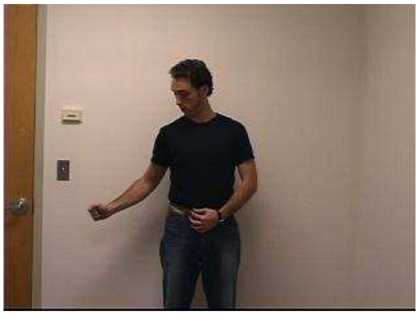
front-outward;



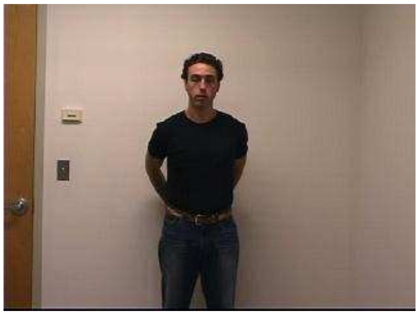
outward (in frontal plane);



behind;

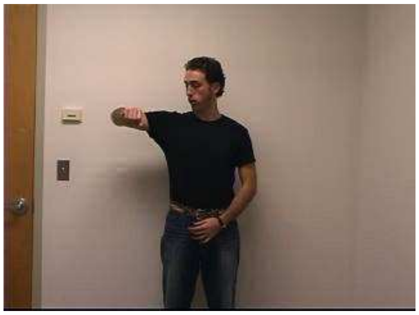


and far behind;

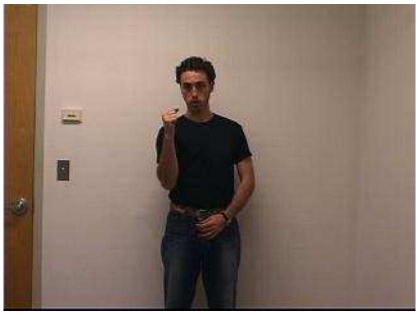


The “upper arm lift” is the measure of the angle created by the upper arm and the body. It is measured in increments of 45 degrees. For example:

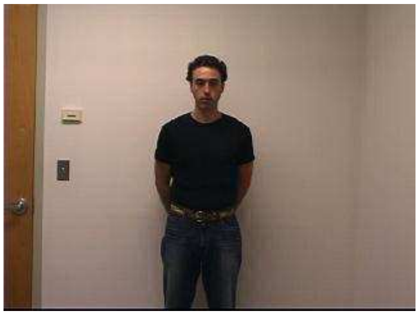
outward – 90 degrees;



front – 45 degrees;

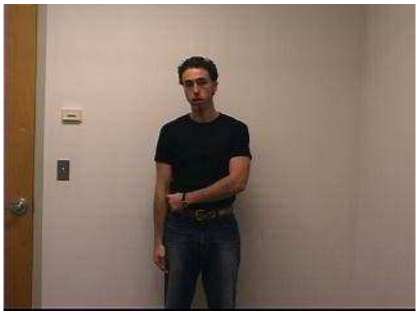


behind – 0-45 degrees.

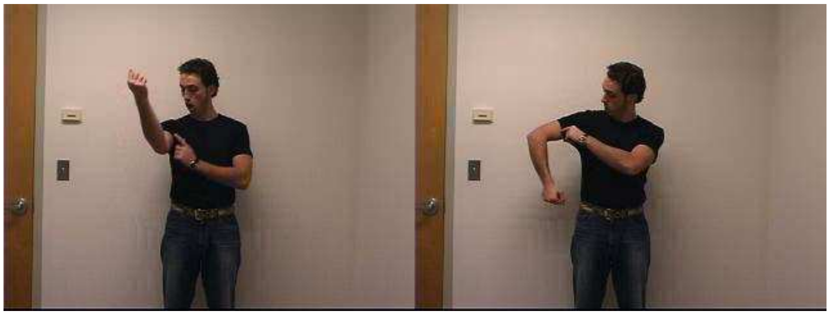


When the arm is directly at the side it is annotated as:

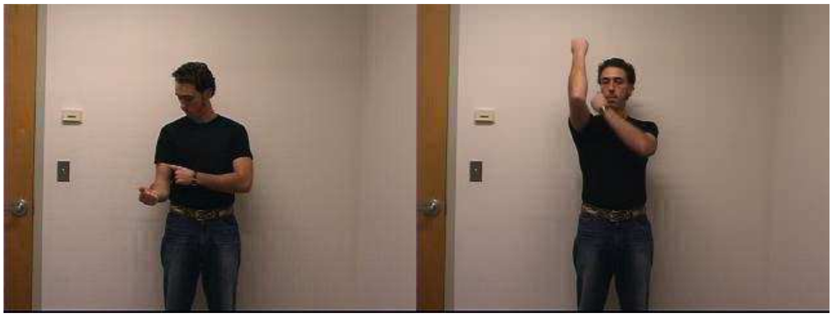
outward – no lift.



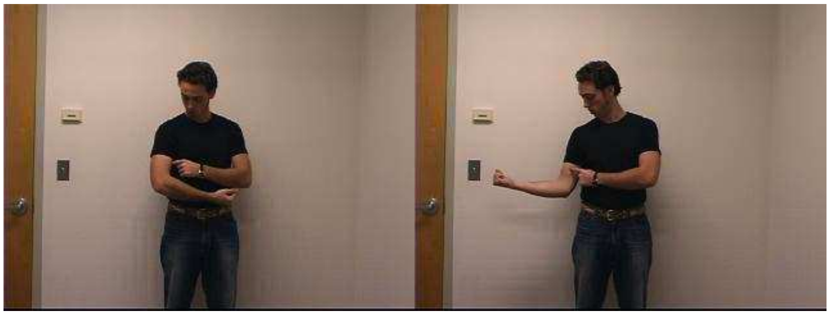
The direction of the bicep is annotated by describing its orientation in three dimensions. They are:



upward/downward;

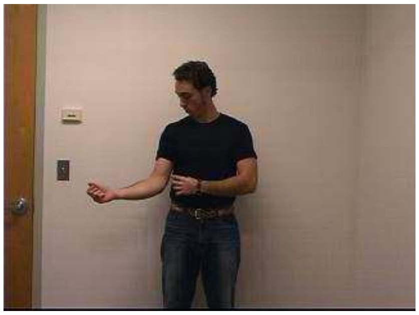


forward/backward;

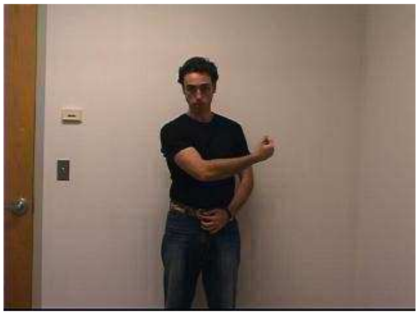


and inward/outward.

In many instances two or even three of these attributes may be used simultaneously, such as:

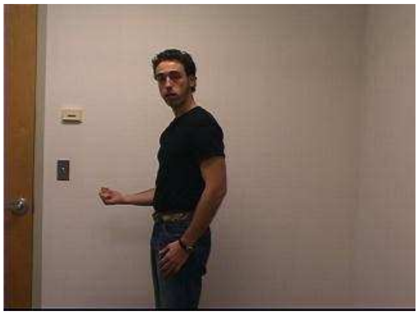


outward – forward;



or inward – upward – forward.

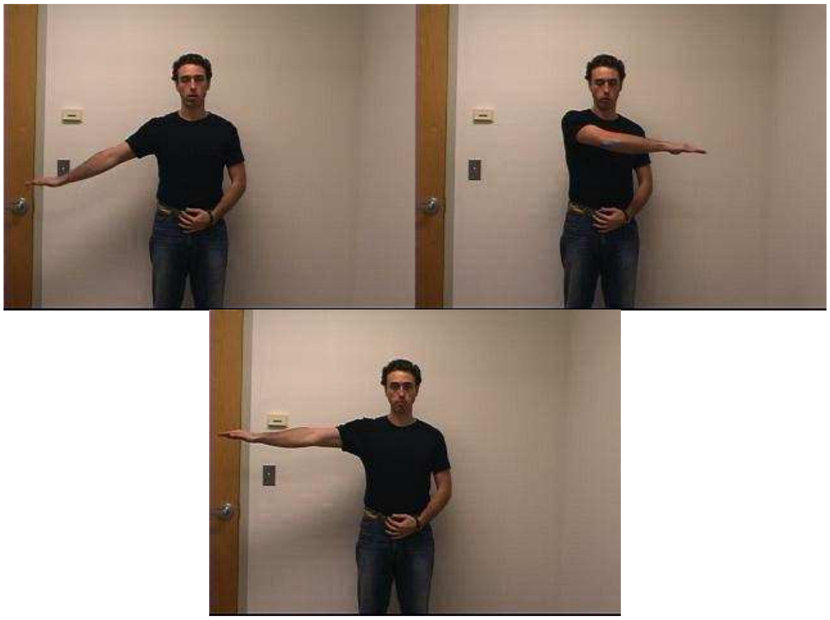
Within the upper arm location there is also an “obscured” attribute. This is used when the exact position of the arm cannot be seen, but can be inferred based on other information. For example, the following location would be annotated as:



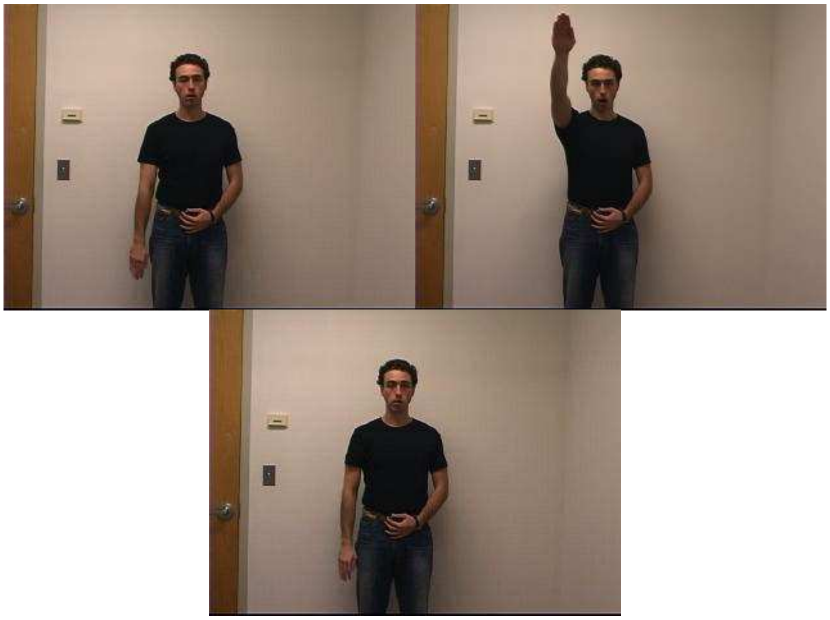
no lift–outward –bicep forward.

Although the upper arm cannot be seen, it is still possible to infer this position. The annotation should as above with the “obscured” attribute box checked as well. This “obscured” attribute is used similarly in all location and movement tracks.

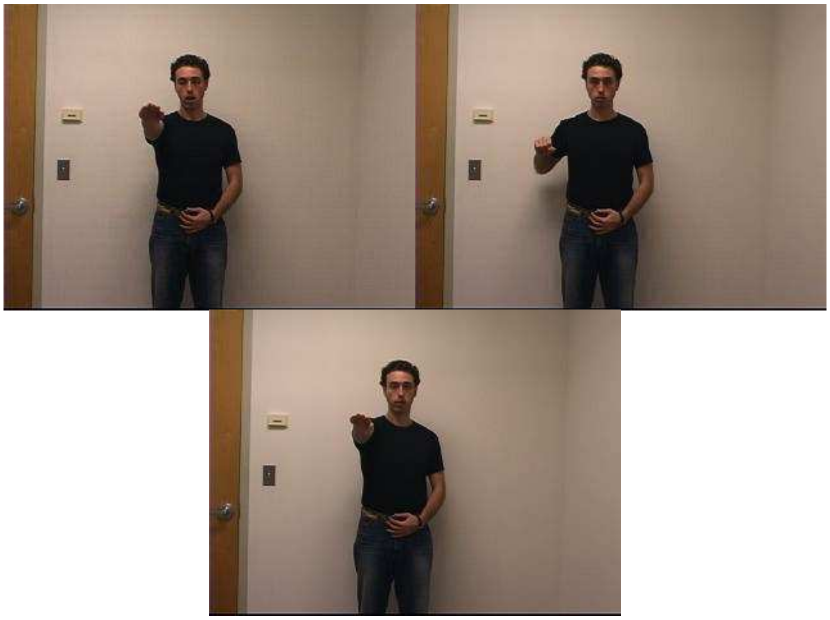
The “Upper Arm Movement” track contains the following attributes. There are three planes of movement used to describe linear movement. They are:



inward/outward (in the x plane);

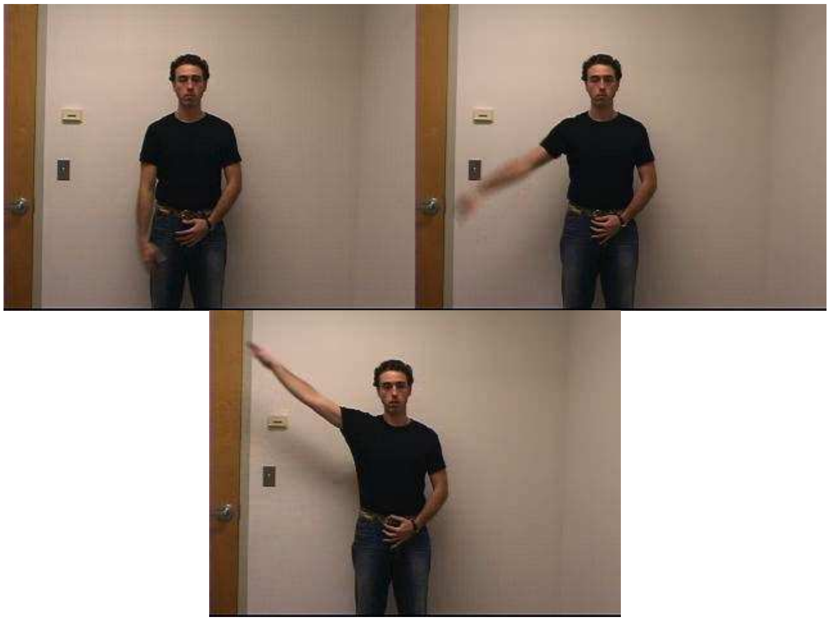


up/down (in the y plane);

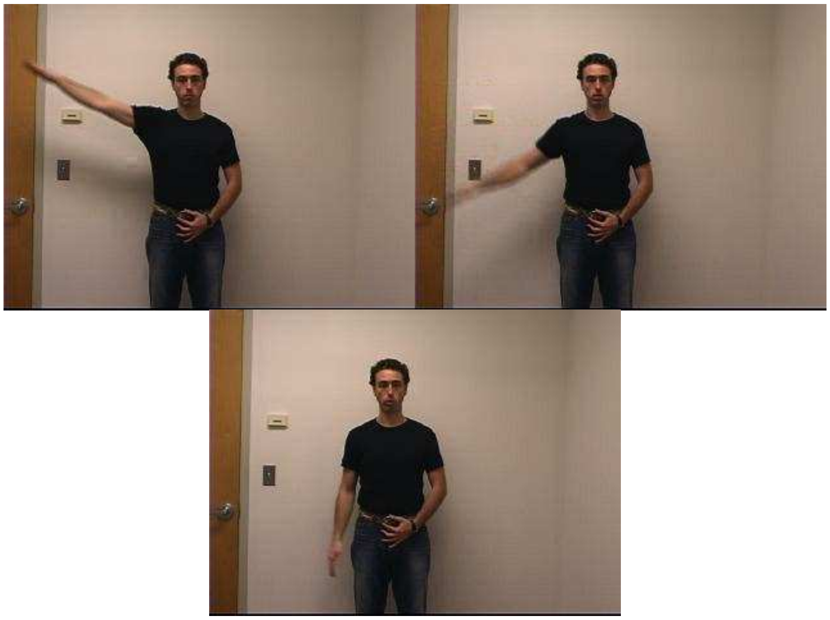


and towards/away (in the z plane).

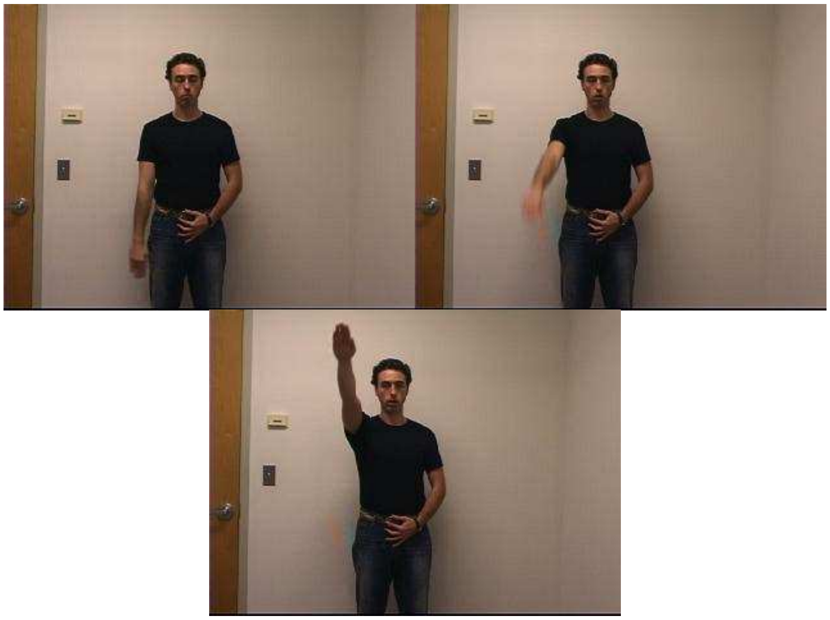
These can be used in conjunction with one another as well. For example:



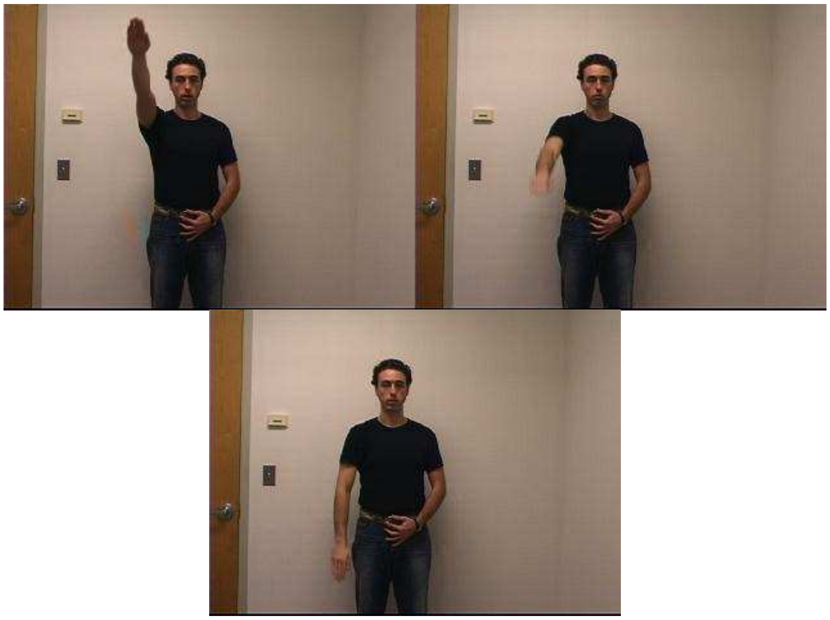
outward – up;



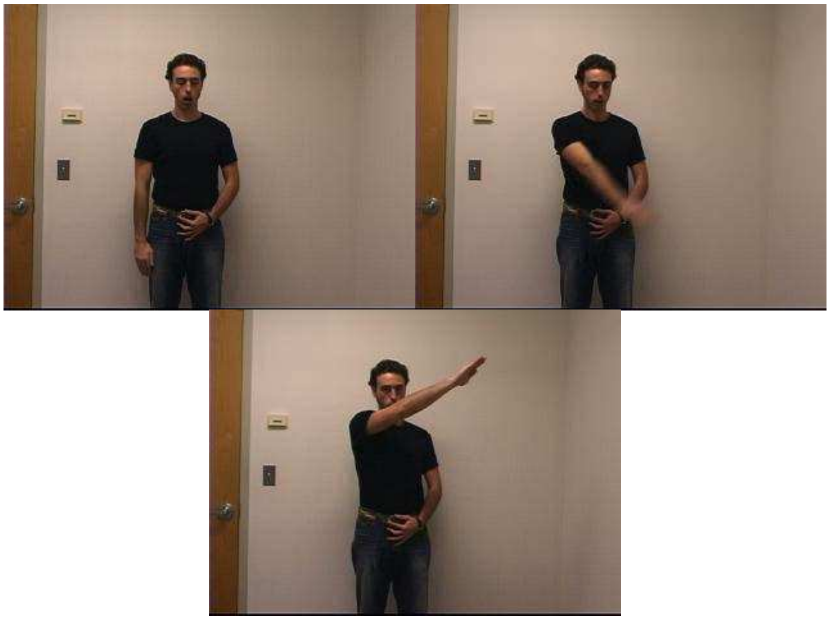
inward – down;



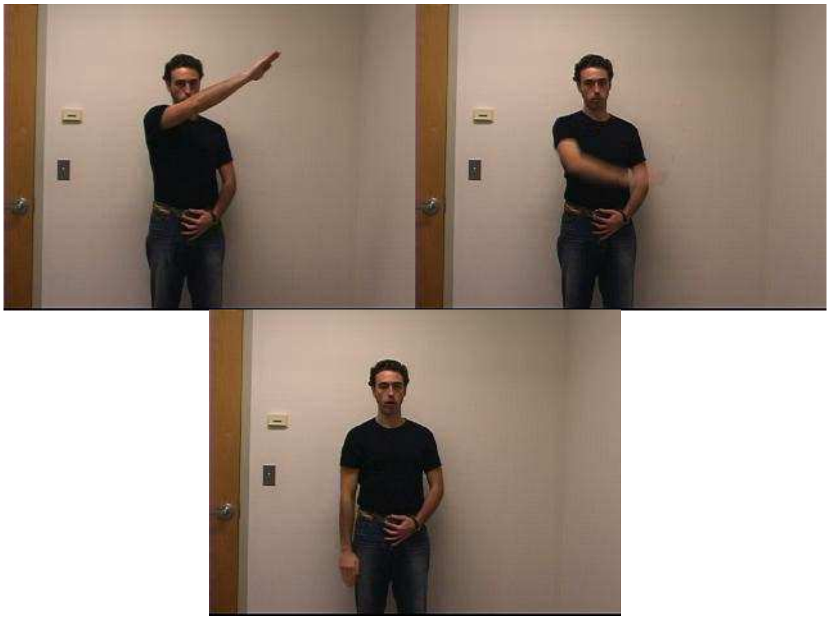
away – up;



towards – down;

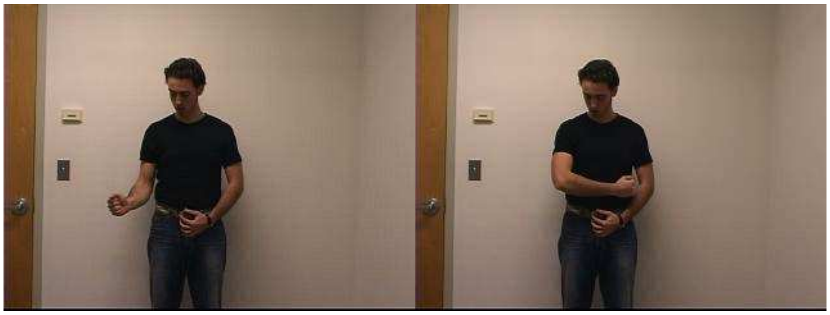


up – inward – away;

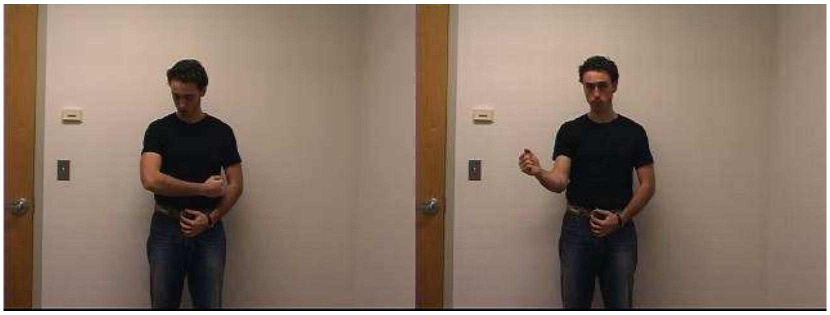


and down – outward – toward.

The “upper arm rotation” is measured in 45-degree increments, and the rotation direction is either:

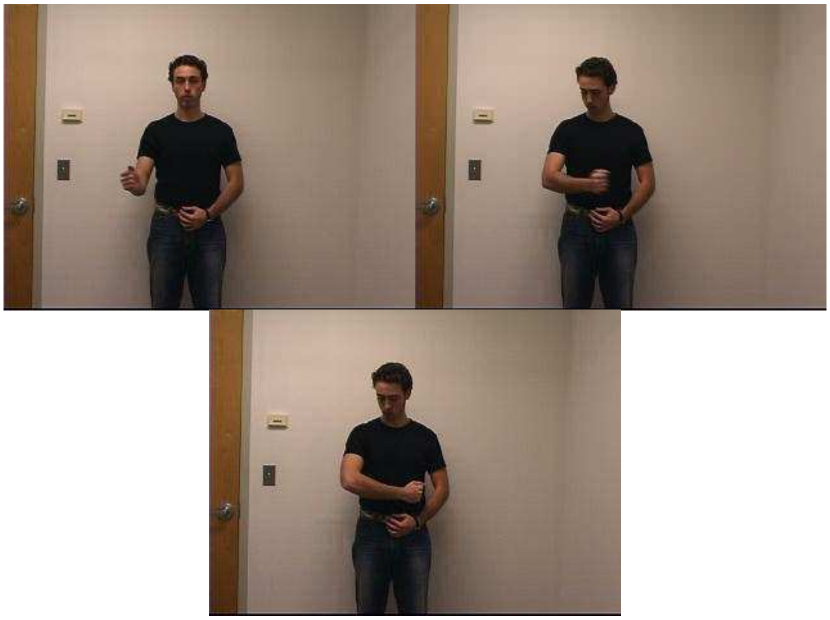


inward;

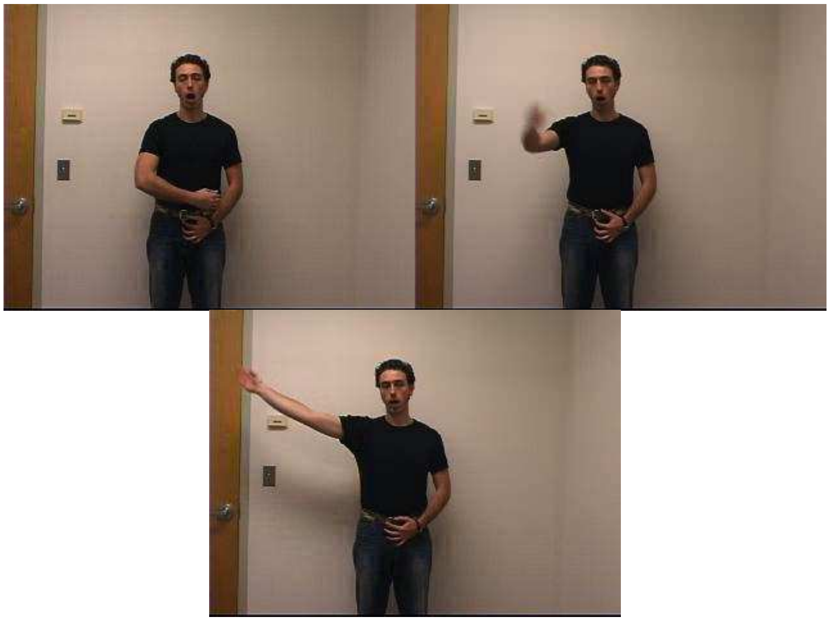


or outward.

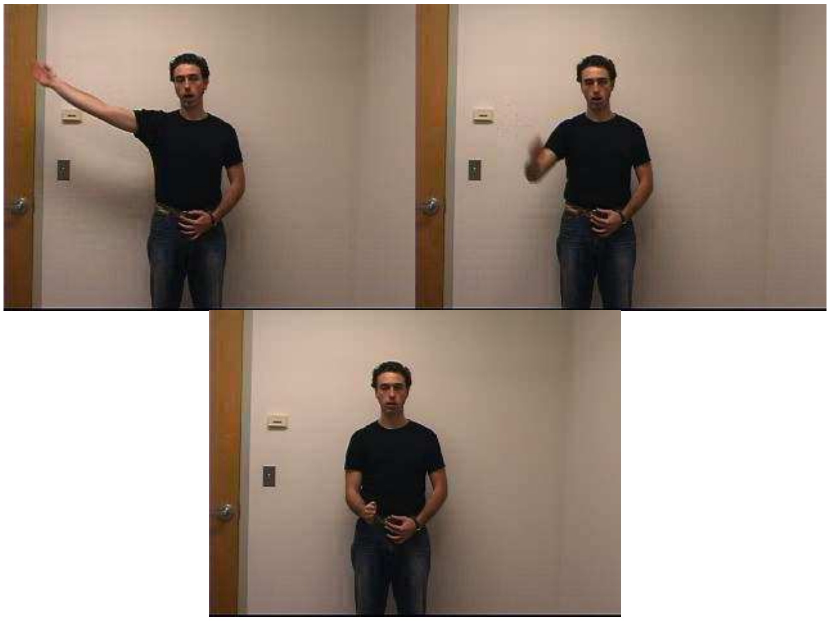
For example, the following is a 90-degree inward rotation.



These two rotation attributes can be used in conjunction with the 3 linear movement attributes to describe more complex gestures such as the following:

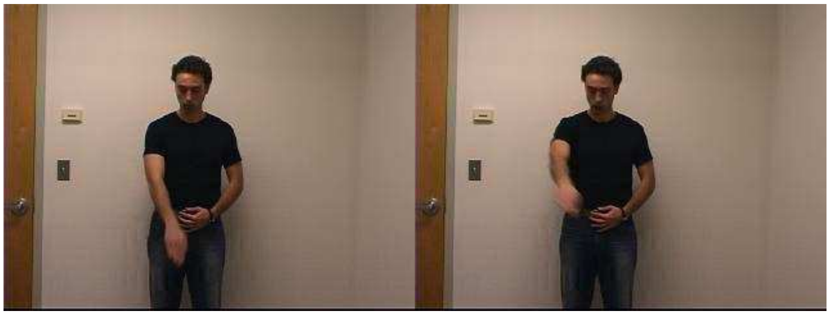


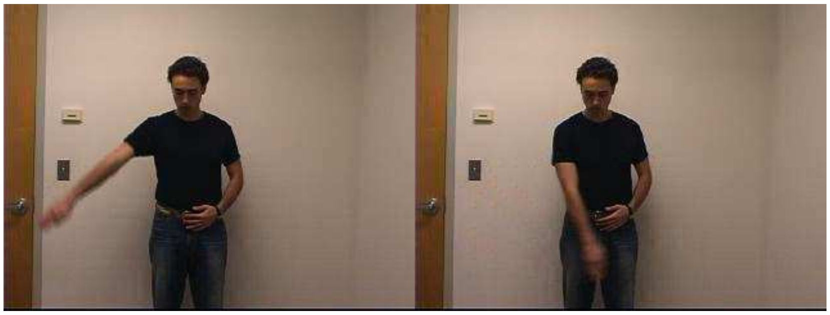
up – outward – 180-degree-rotation outward;



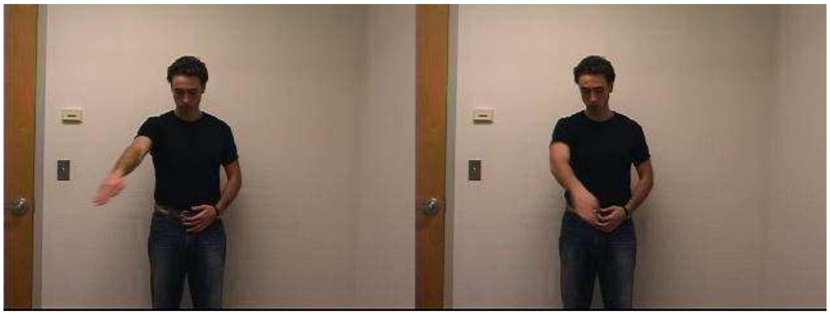
or down – inward – 90-degree-rotation inward.

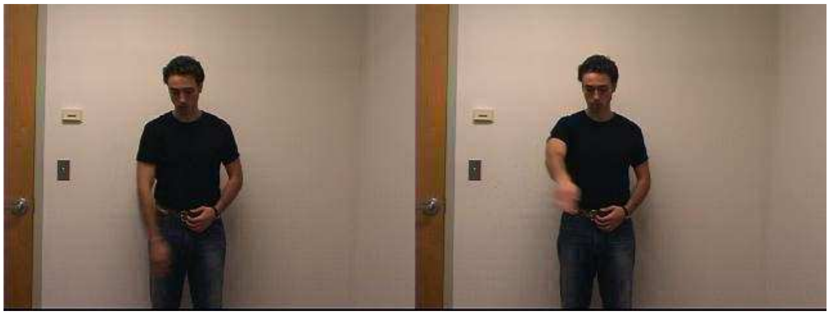
The values for the circular movement attribute are: parallel to the horizontal plane—the plane that divides the body into top and bottom halves—both





clockwise and

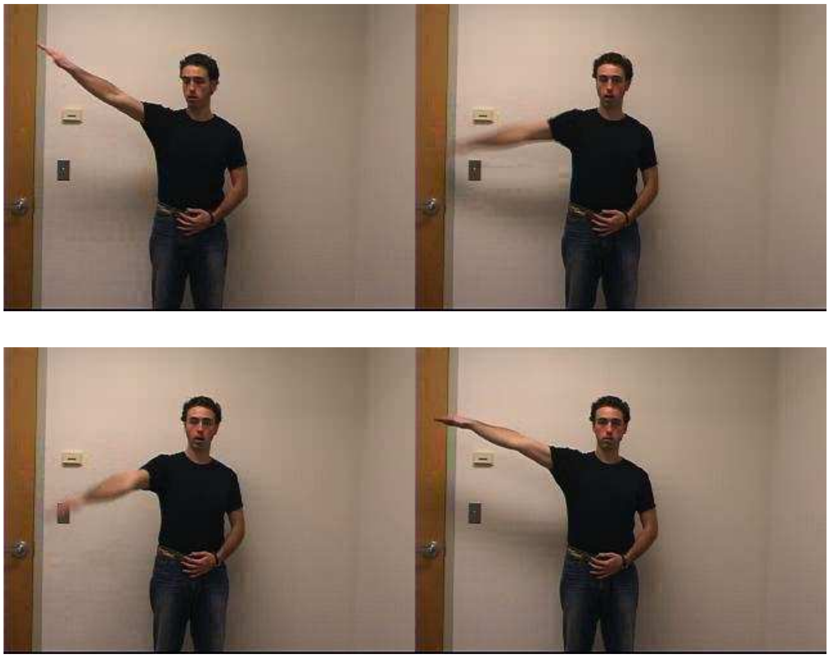




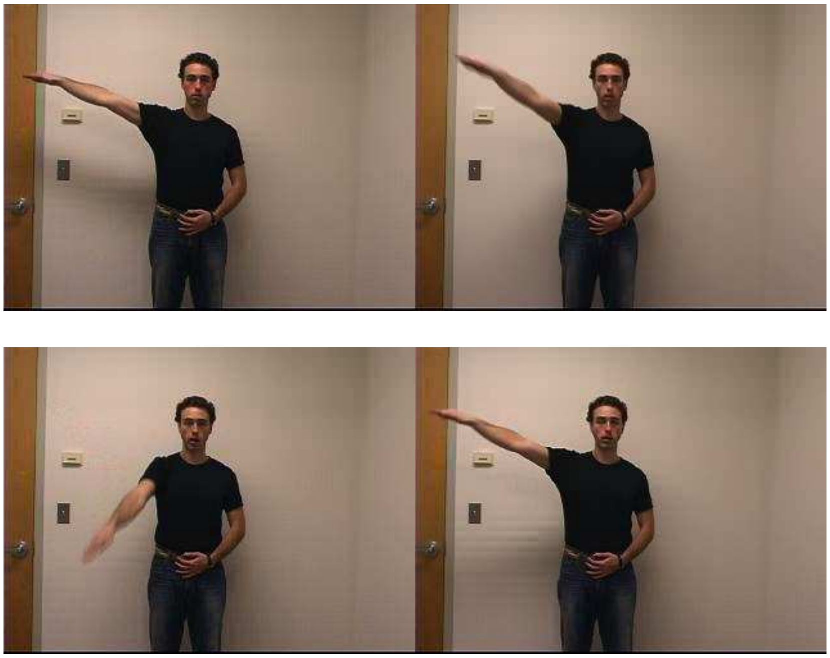
and counter clock-wise;

parallel to the median plane—the plane that divides the body into left and right halves—

both

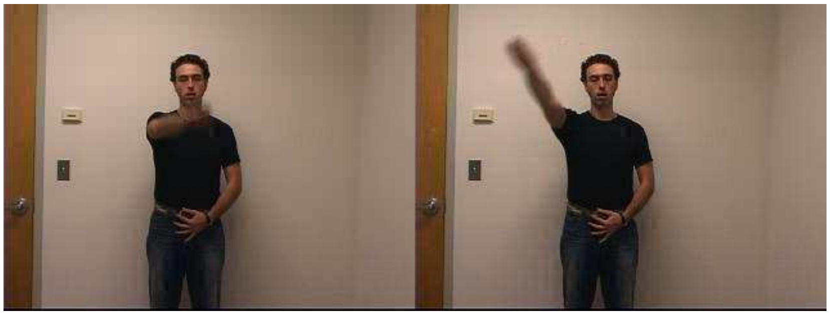


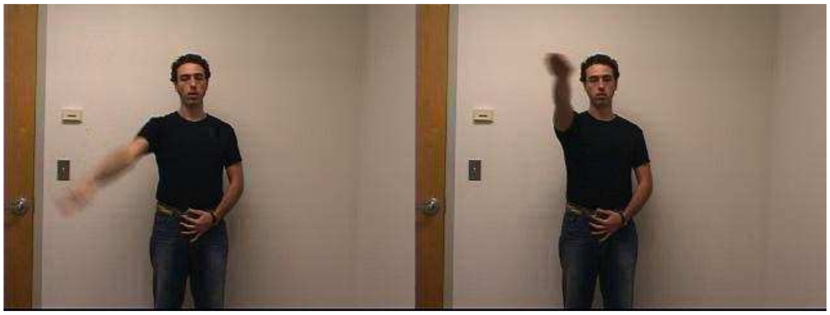
clockwise and



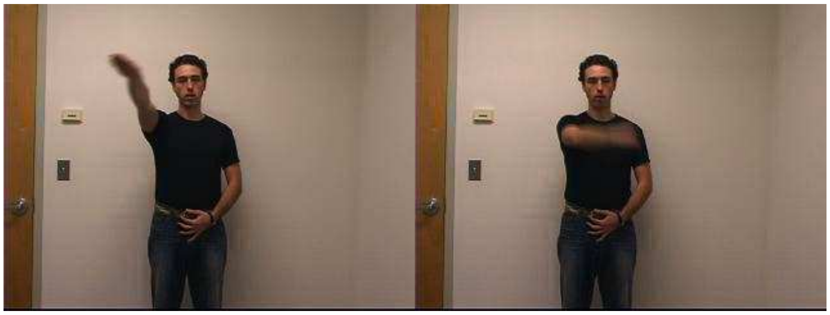
counter-clockwise;

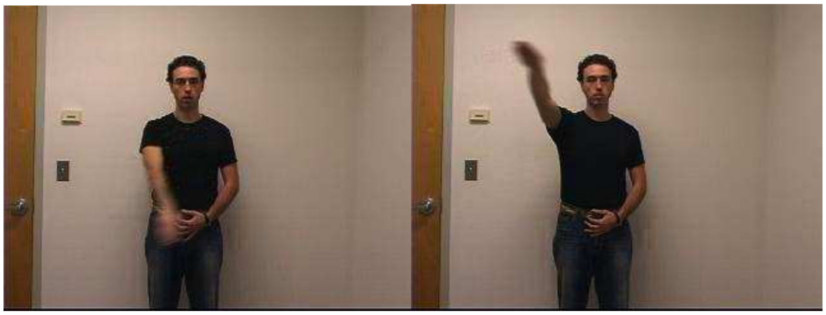
and parallel to the frontal plane—the plane that divides the body into front and back halves—both





clockwise and

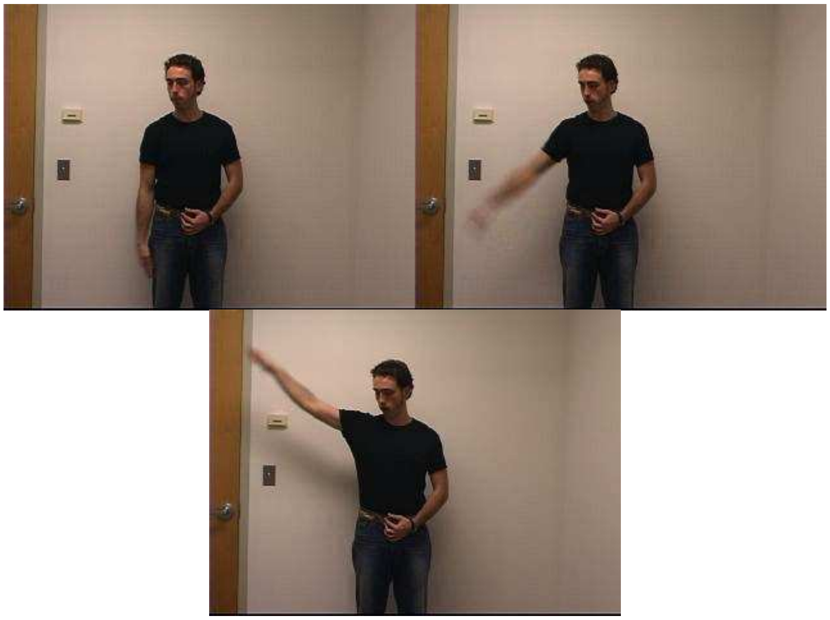


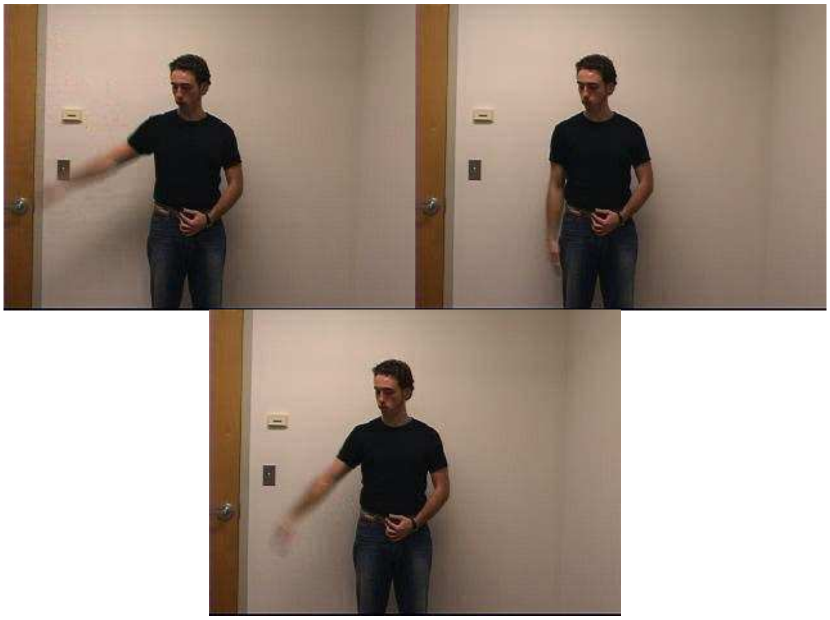


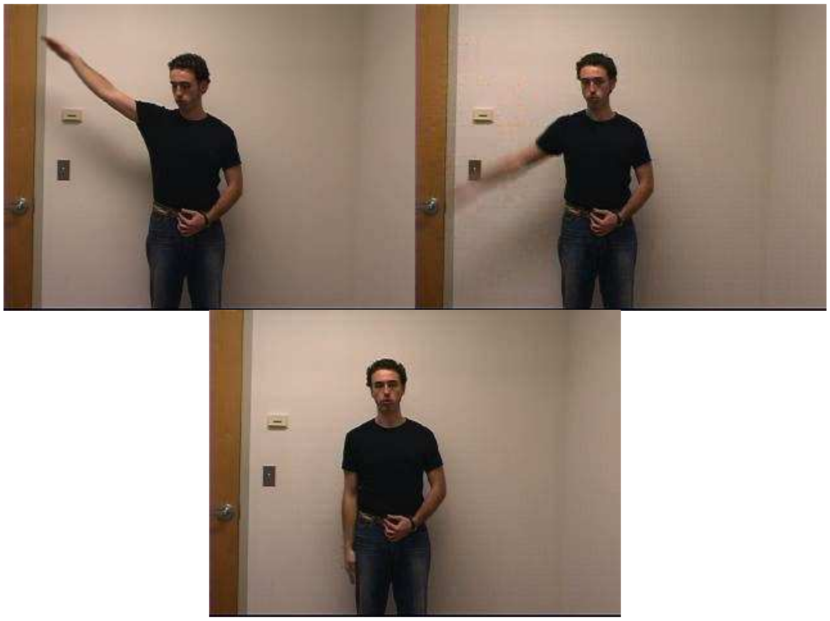
counter-clockwise.

The “effort” attribute is intended to capture the relative amount of effort, on a scale of one to five, exerted during a movement. This attribute requires some level of subjective judgment. A value of *3* should be used to represent an “average” amount of effort.

The “strokes” attribute is used to capture repetitious movement. For example, the following seven stills demonstrate a movement which would be labeled as up-down, outward-inward, 2 strokes.







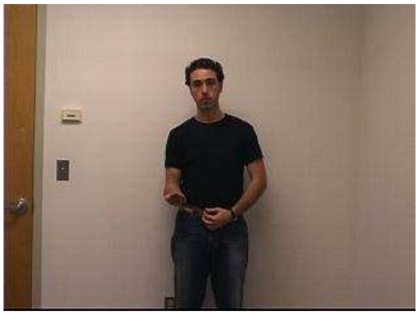
The “effort” and “strokes” attributes are similar for all movement tracks. Again note the ability to mark the track “obscured”.

Forearm

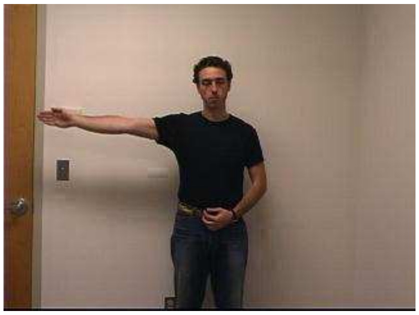
The forearm, like the upper arm, has a location track and a movement track. These are used to record gesture information pertaining to the part of the arm between the elbow and the wrist.

The “Forearm Location” track contains attributes for “elbow flexion and “forearm orientation.”

Elbow flexion refers to the angle formed by the forearm and upper arm. It is measured in 45-degree increments. For example, the following would be

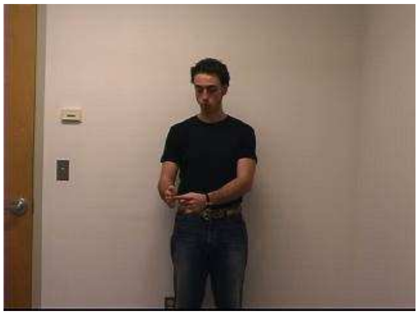


90-degree elbow flexion,

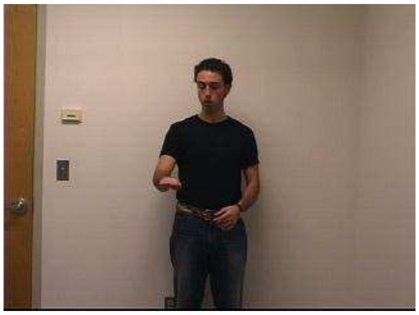


while a straight arm would be 180.

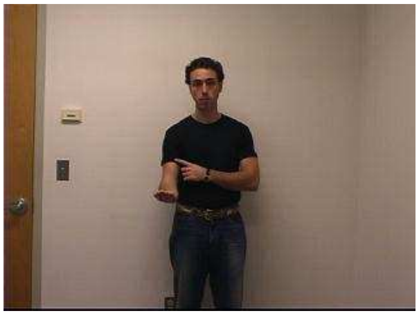
The values for the “forearm orientation” attribute are:



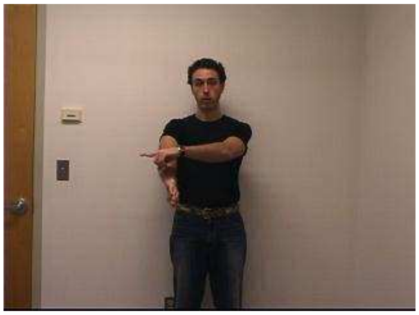
neutral (the palm faces inward);



prone (the palm is oriented away from the bicep);



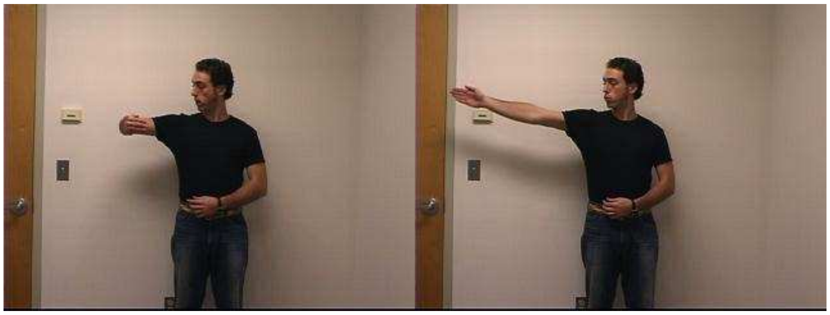
supine (where the palm is oriented towards the bicep);



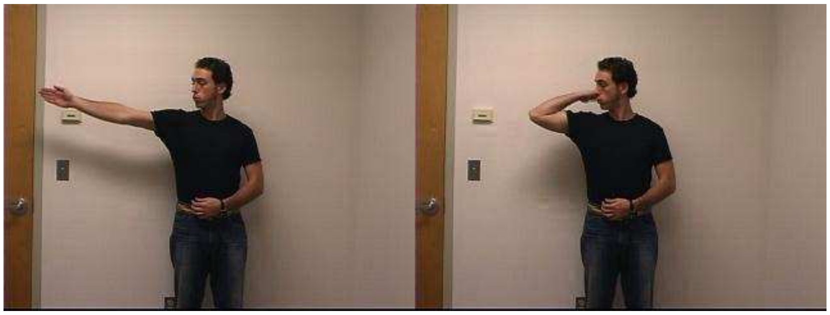
and, inverse (the palm faces outward);

Neutral/Prone, Neutral/Supine, and Prone/Inverse are used when the palm does not face fully in one direction. Again, note the ability to mark this location obscured.

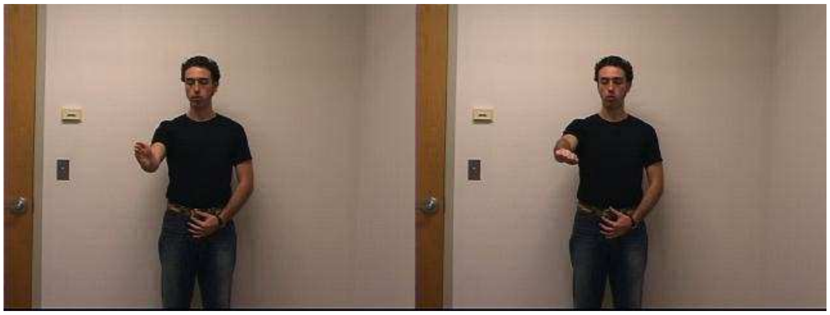
Changes in elbow flexion and forearm orientation are annotated using the "Forearm Movement" track. Elbow flexion changes are recorded in 45-degree increments. For example, changing from 90 degrees to 180 degrees would be a 90-degree change in elbow flexion.



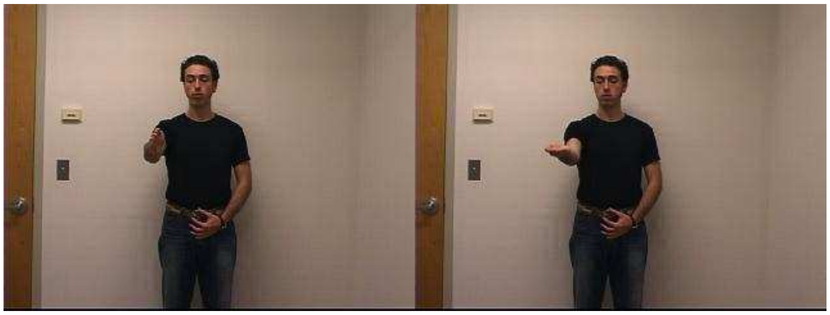
A change from 180 degrees to less than 45 degrees would be a change of “135 to 180.”



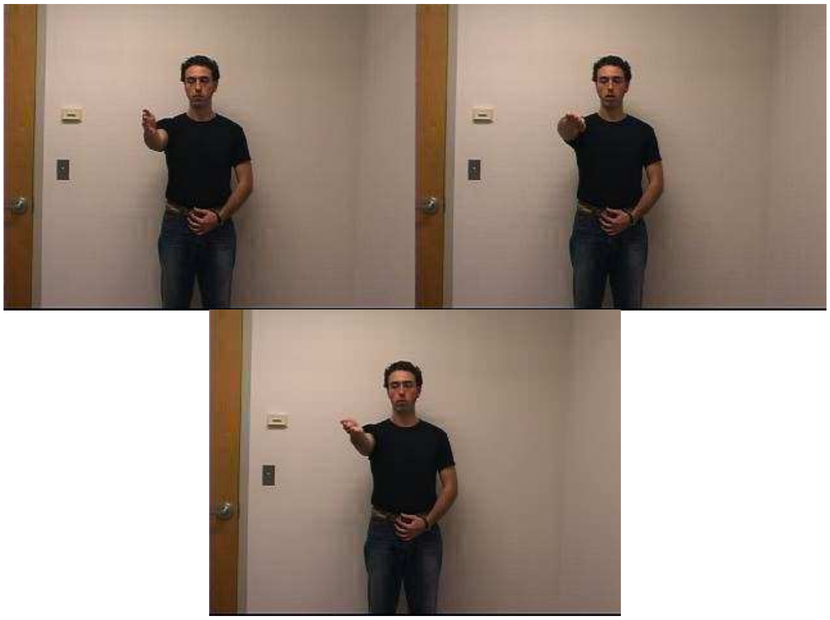
A “change in forearm orientation” from neutral to prone would be annotated as inward forearm rotation.



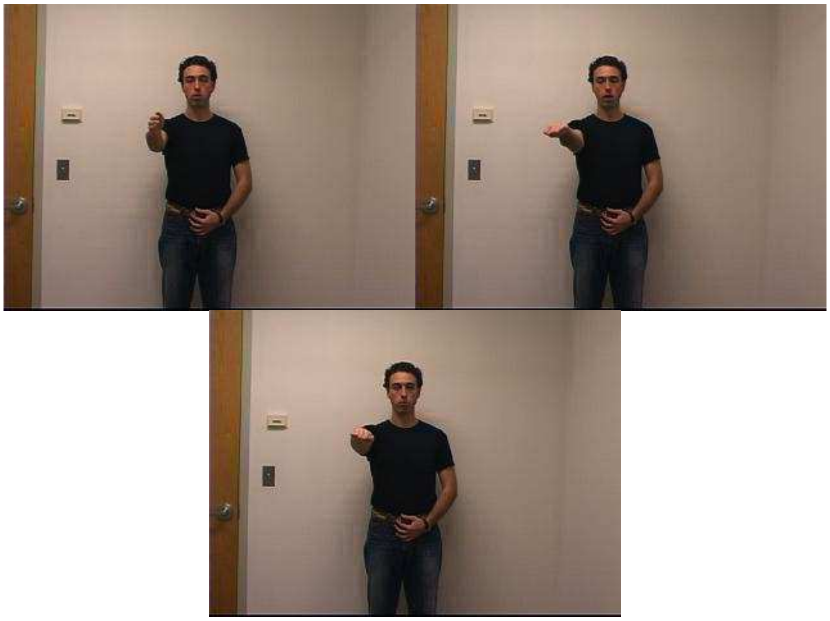
Likewise, a change from neutral to supine is outward rotation.



There are also values for inward-outward and outward-inward which are used to annotate a reversal of direction without a pause such as the following:



inward-outward;



or outward-inward.

Effort, strokes, and obscured are annotated as before.

Hand and Wrist

Instead of location and movement tracks, as the upper arm and forearm, the hand and wrist is divided into a shape track and a movement track.

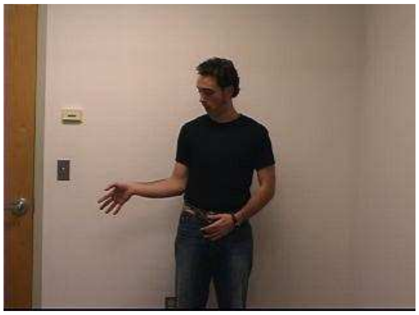
The “shape track” is used to describe the shape of the hand as well as the position and orientation of the wrist.

The handshapes can be found in the catalogue below (* Add hand shape catalogue *). Each handshape is broken into two parts, the handshape group and the handshape letter.

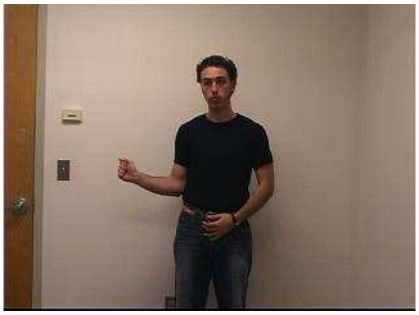
Each group is determined by the number of extended fingers with a closed fist being 0, and ranging f from 1 to 5 fingers extended. Six refers to miscellaneous shapes.

Often, the exact handshape may not be found in the catalogue. In these cases, the closest matching handshape should be chosen.

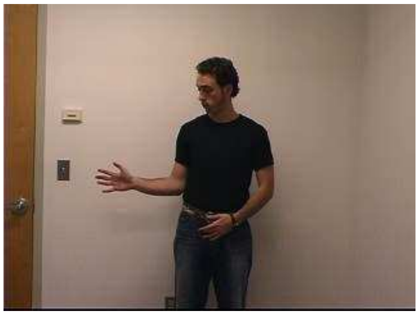
After choosing the handshape, the tension the hand must be annotated. The possible values from:



relaxed, for example a limp wrist;



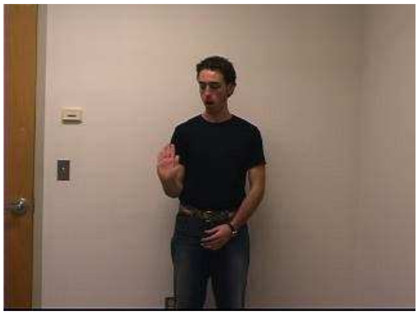
to *very tense*, e.g., for a tightly clenched fist, with



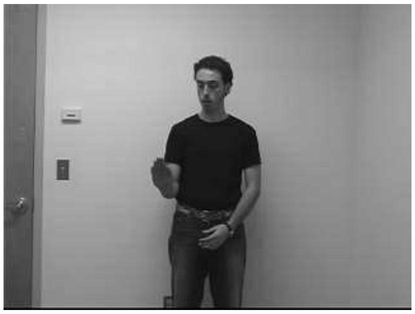
slightly tense used to describe an average amount of exertion to hold the hand shape;

The “Hand and Wrist Shape” track is also used to annotate the orientation of the wrist. There are two attributes to describe the bend of the wrist. The first is “up and down”. This describes the amount of bend toward the upper or under side of the forearm. The values are:

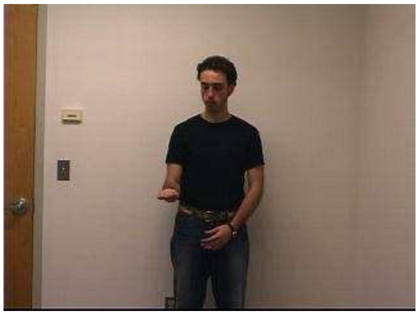
up;



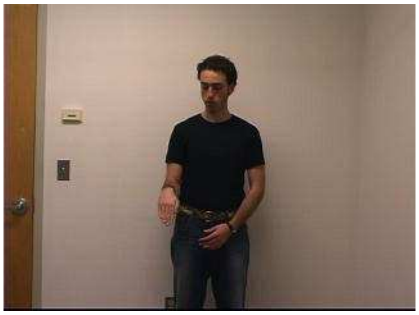
up/neutral;



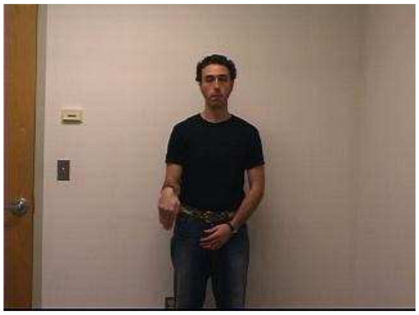
neutral;



down/neutral;

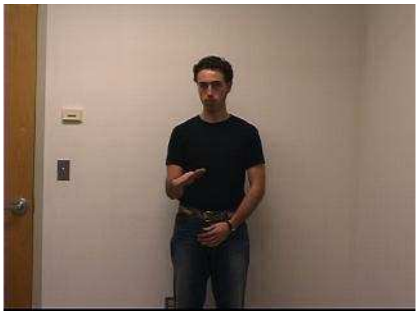


down;

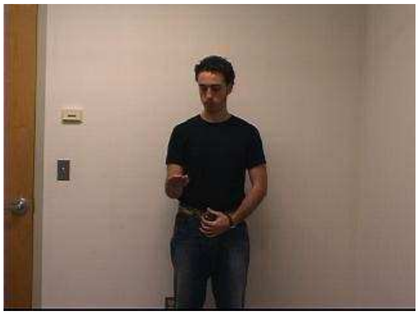


The second attribute of wrist bend is “side to side”. This describes whether the wrist is:

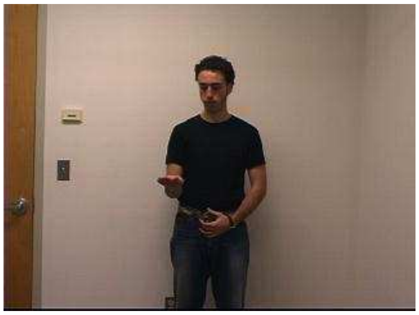
towards the thumb;



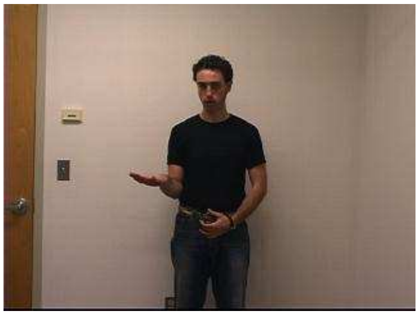
neutral;



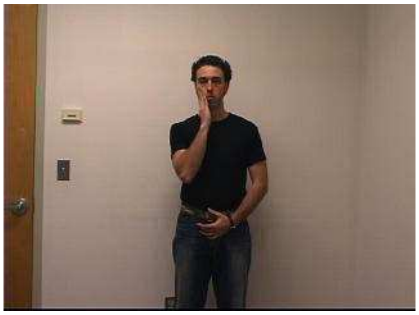
towards the little finger;



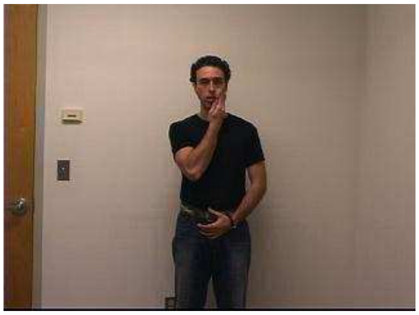
and extremely towards the little finger.



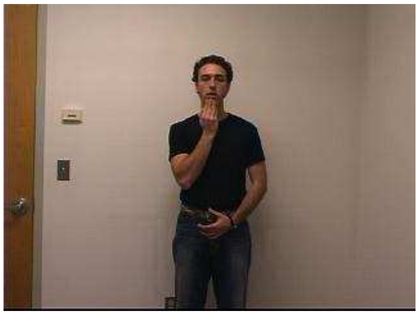
If during a gesture the hand touches another part of the speaker's own body, it can be annotated using the "part of body touched" attribute. The values for this attribute are a list of body parts. Examples include:



cheek on the same;



opposite side;



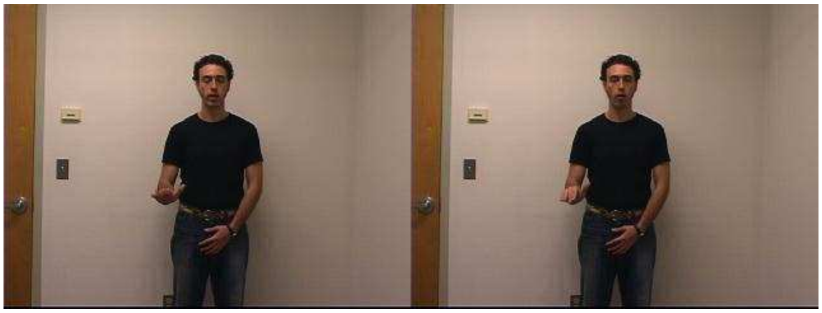
chin;

and chest.

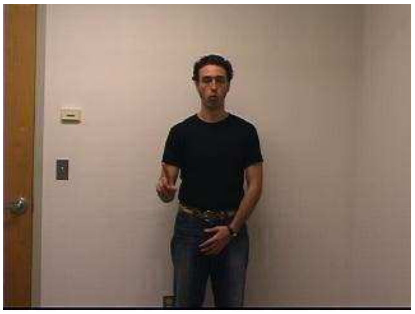
If no part of the body is being touched, then this value is left as “none.” As in previous tracks, it is possible to mark the location obscured.

The “Hand and Wrist Movement” track is used to annotate movement of the hand and fingers and changes in orientation of the wrist.

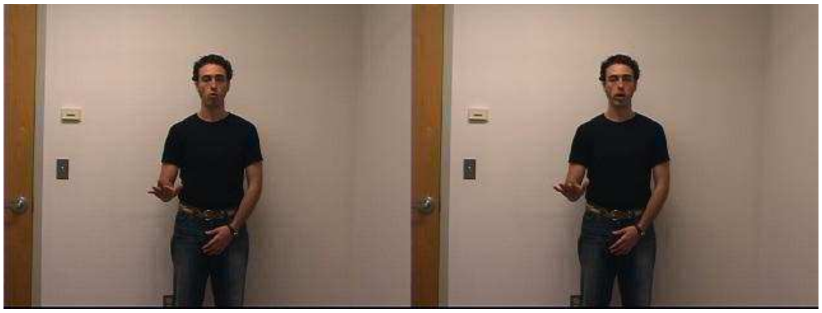
The values of the “hand movement” attribute include three values for finger joint movement. A-joint movement is movement of the joint that joins the finger to the hand. B-joint movement is movement of the middle joint of the finger. A-and-B-joint movement describes simultaneous movement of both joints. These three values are described with greater detail in the “finger coordination” attribute. The values for this attribute describe the finger’s movement in relation to each other. The values are:



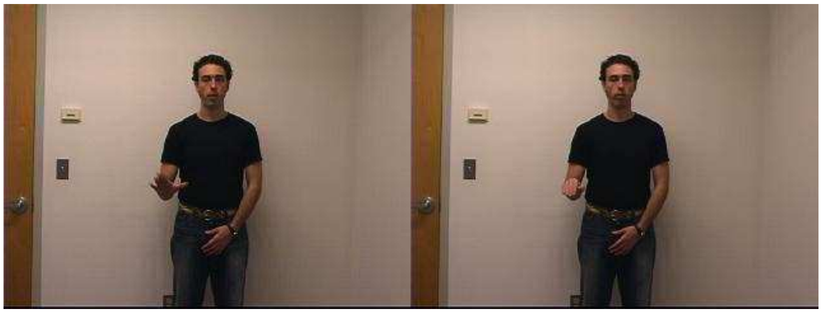
parallel digit movement without thumb



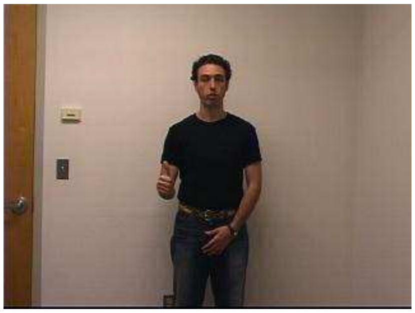
(or one finger);



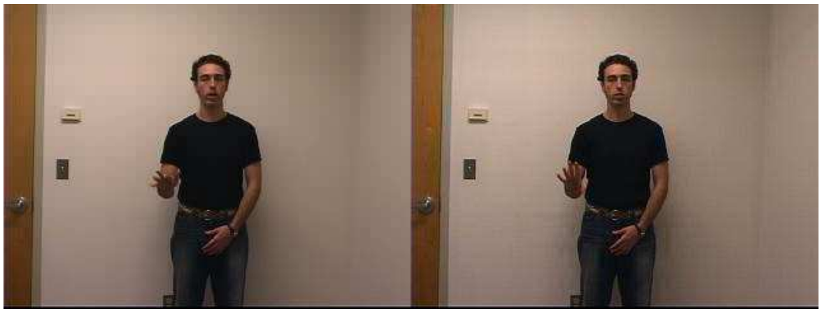
random movement without thumb;



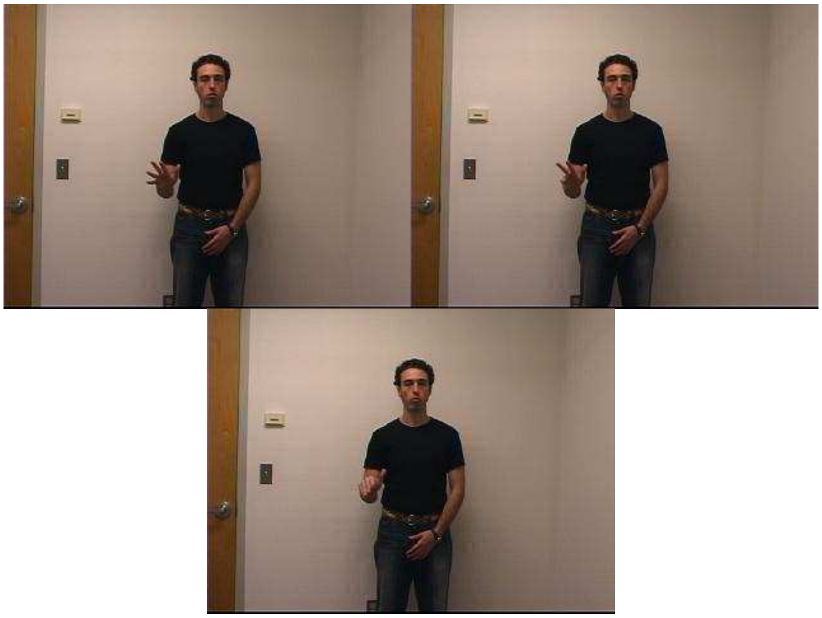
parallel digit movement with thumb



(or just thumb);

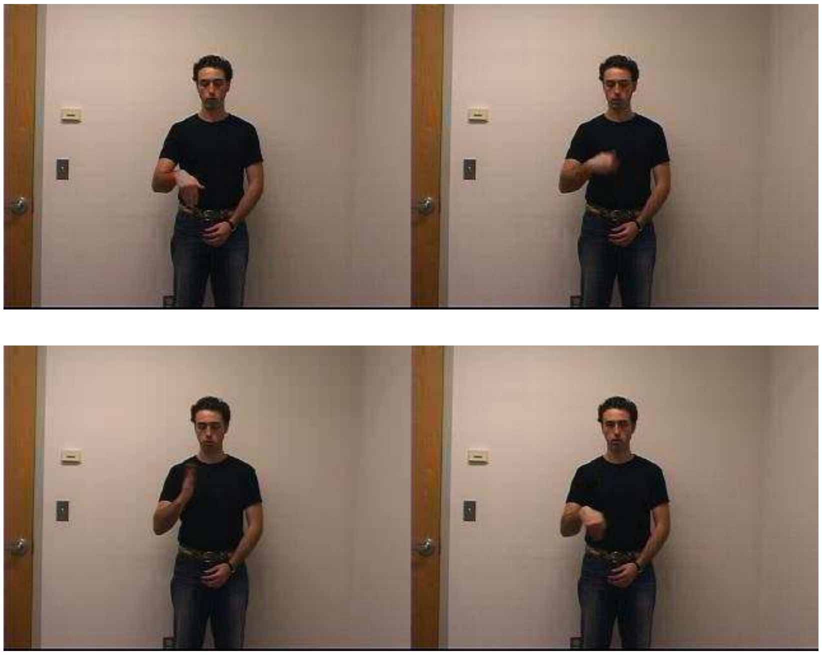


random digit movement with thumb;

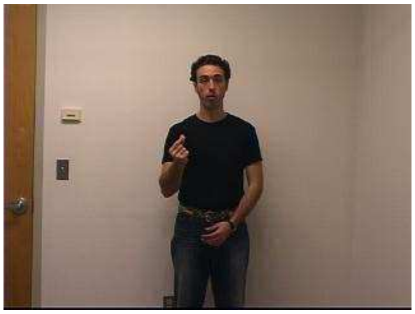


and moving in sequence.

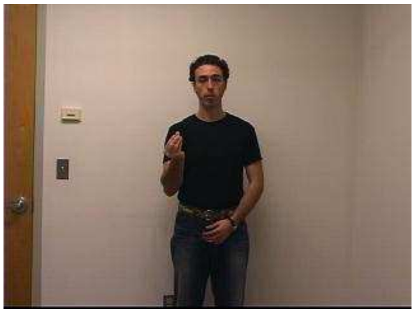
Returning to the “hand movement” attribute, the remaining values are:



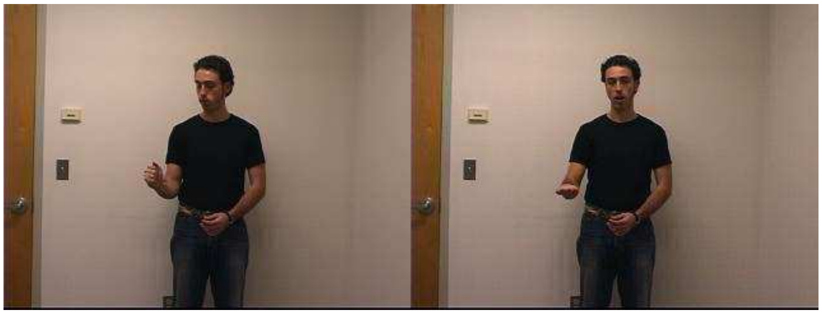
wrist circular;



thumb rubbing index finger;

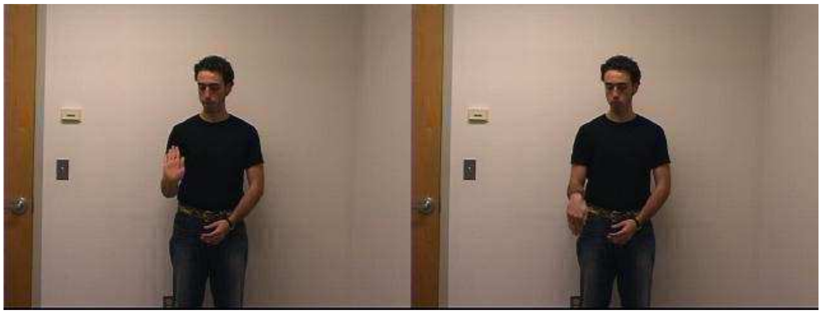


thumb rubbing multiple fingers;

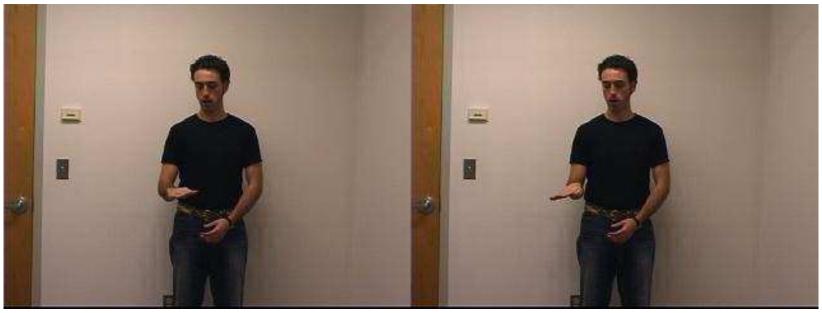


and direct movement between two shapes, e.g., a movement from *0B* to *5A*.

Changes in wrist orientation are annotated using

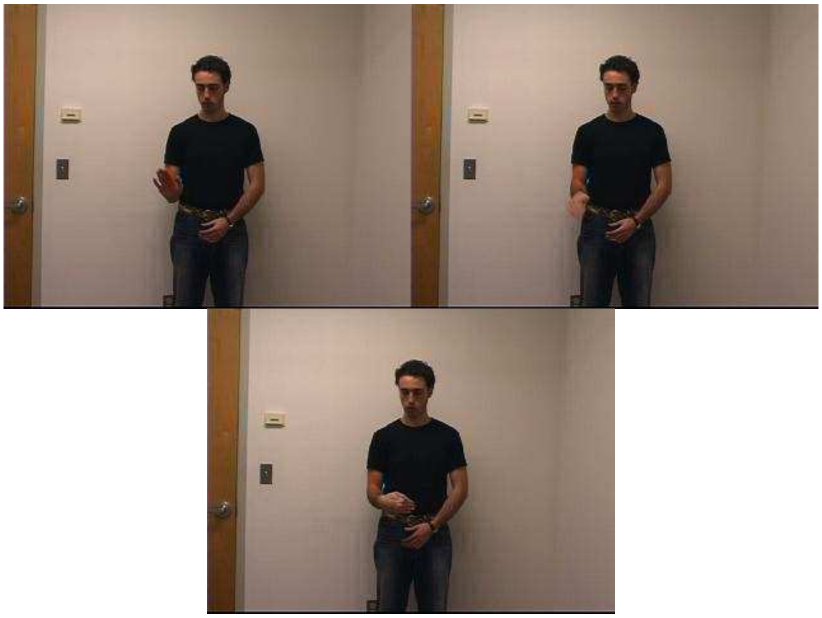


the wrist up-down-movement and

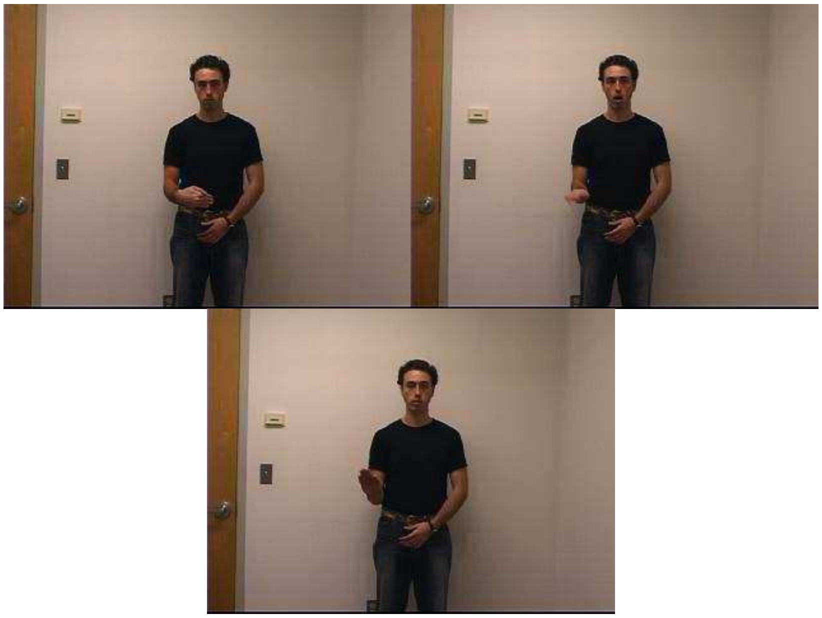


the wrist side-to-side-movement attributes.

These attributes can be used in coordination with others concerning the wrist, for example,



down and towards thumb



or up and towards little finger.

Similar to before, effort, strokes and obscured are also annotated.

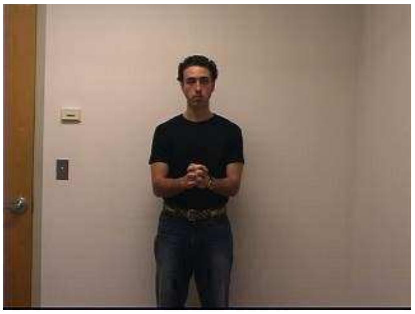
Two-Handed Gestures

The “Two Handed Gesture” track is used to annotate gestures in which both hands are moving in concert. If the hands are touching, the point of contact for each hand is noted.

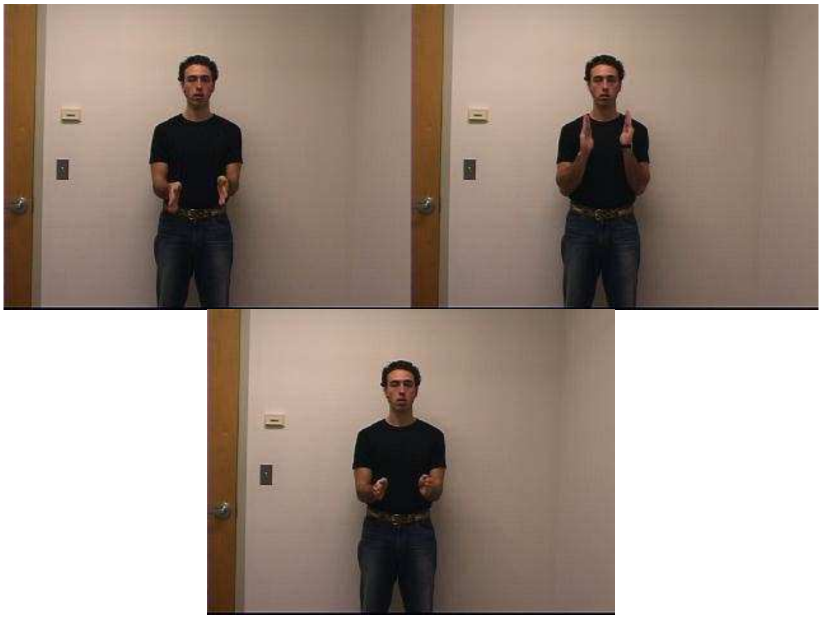
Above, the right hand point of contact is the index finger, and the left hand point of contact is the palm.

Here, the right and left hand points of contact are both multiple fingers.

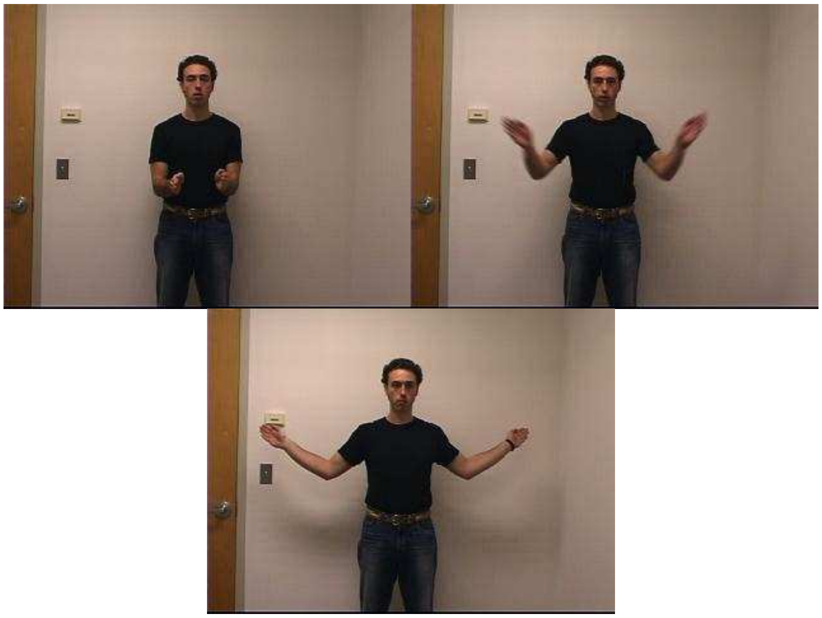
Additionally, there is a list of values that further describes the hands’ relationship to each other. These values are:



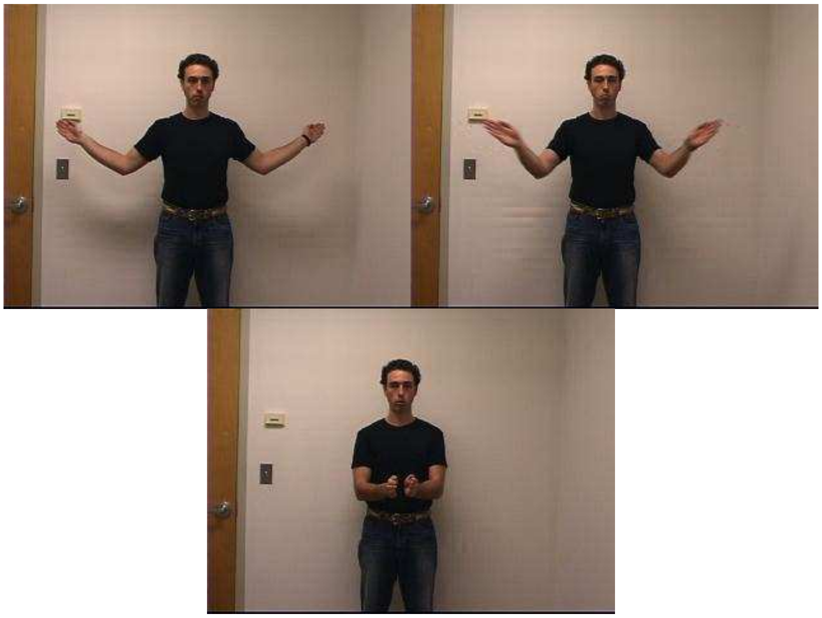
interlaced;



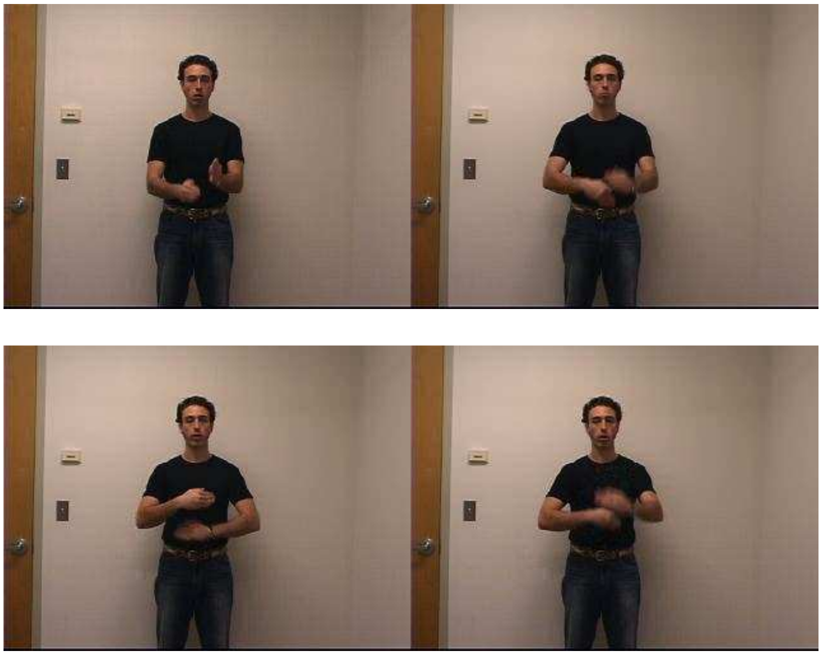
moving in parallel;



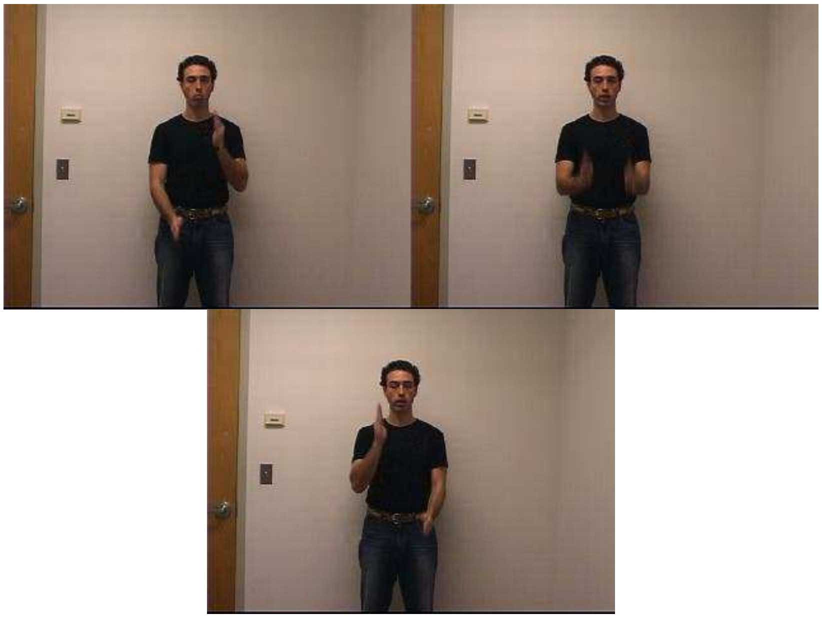
moving apart;



moving towards;



moving around one another;



moving in alternation;

and hands crossed.

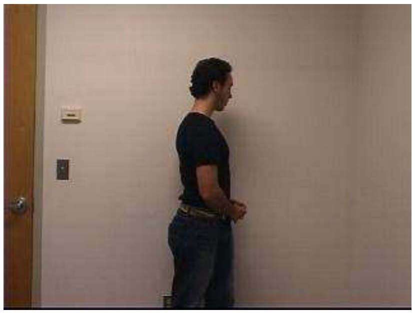
Again, there is the ability to mark the track obscured.

Torso

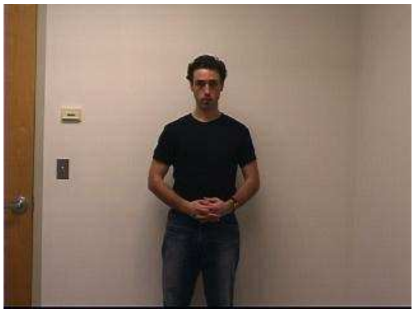
To annotate the torso and head, there are orientation and movement tracks for each.

In the orientation track of the torso there are attributes for "vertical axis orientation", "front-back orientation", and "side-to-side orientation."

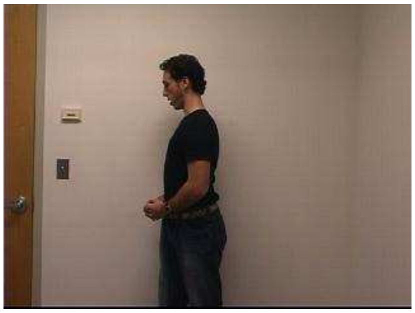
For vertical axis, values range from



left

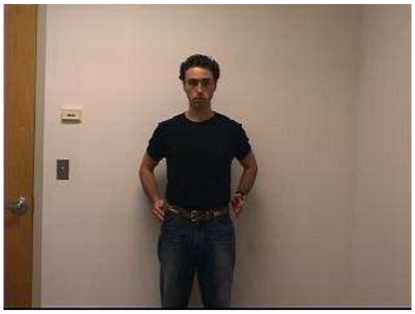


to center

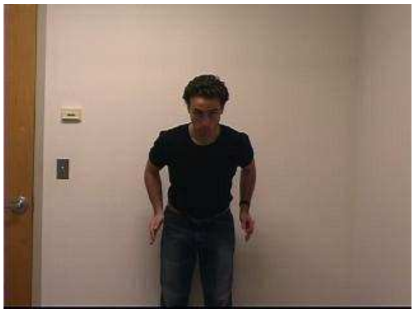


to right.

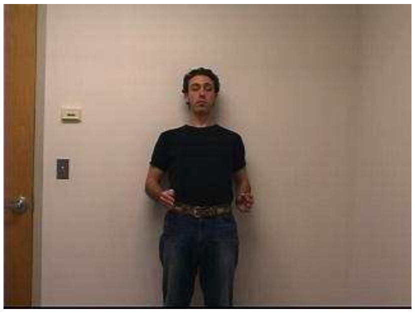
Front-back values are:



center;

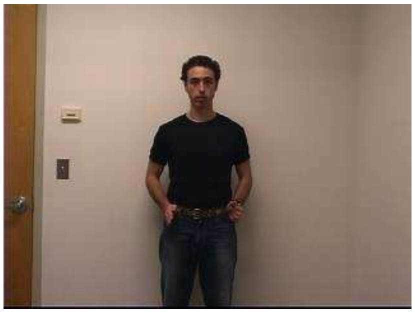


forward;

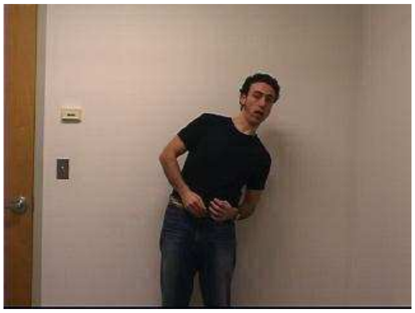


and backward.

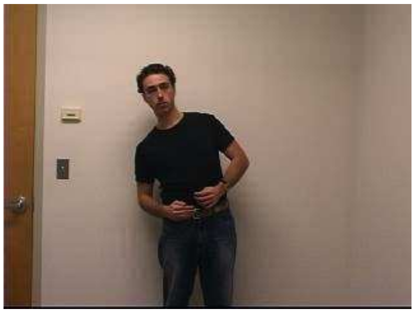
Side-to-side values are:



center;

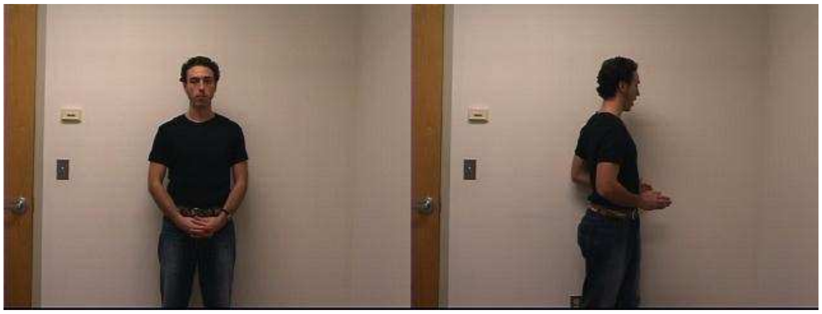


left;

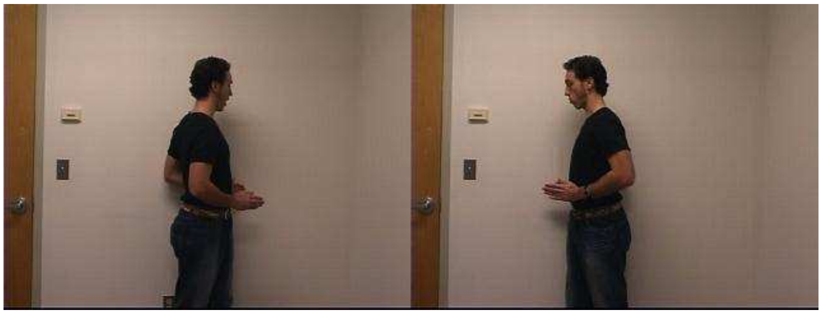


and right;

The “Torso Movement” track has attributes for “vertical axis rotation” in 45 degree increments. For example:

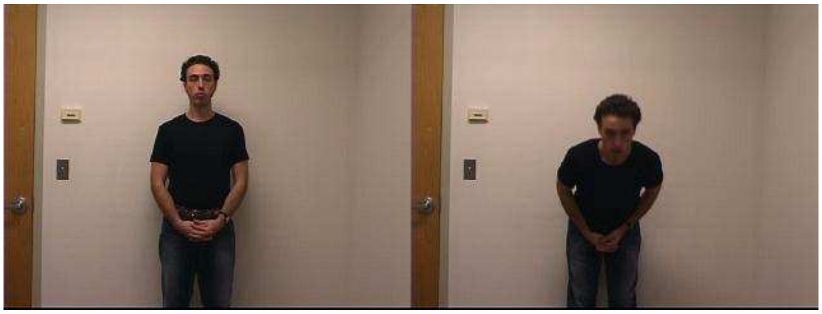


90-degree rotation left;

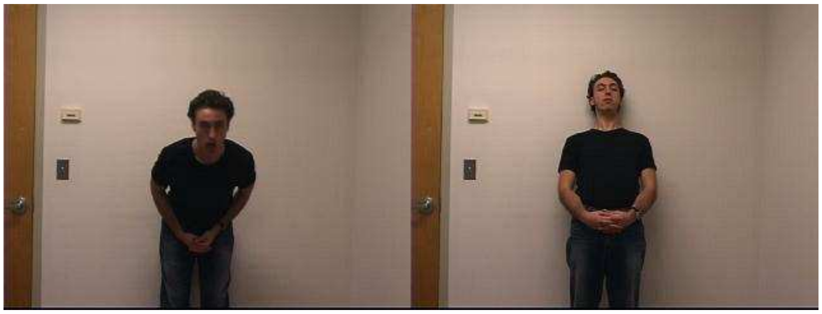


or 180-degree rotation right.

The values for “front-back movement” are:

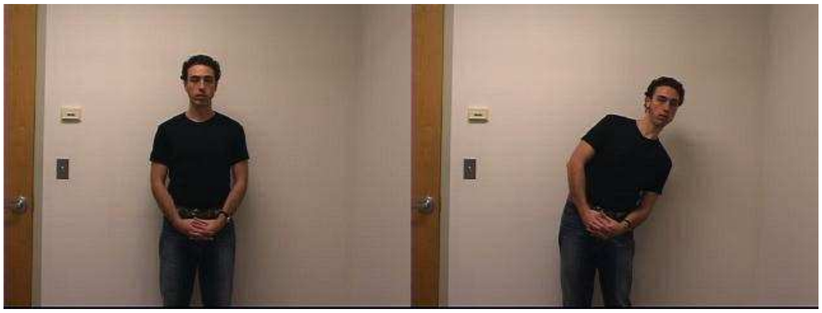


forward;

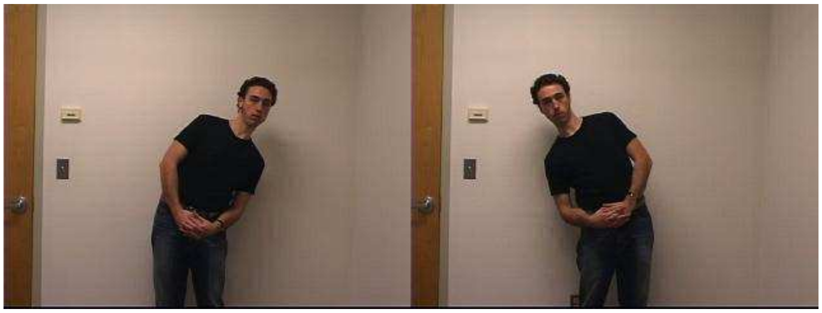


and backward;

The values for “side-to-side movement” are

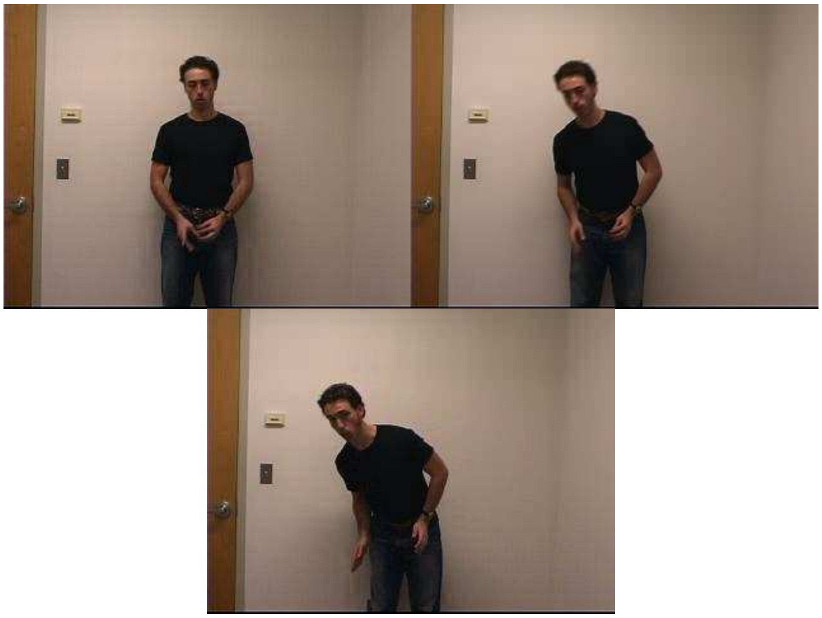


left;

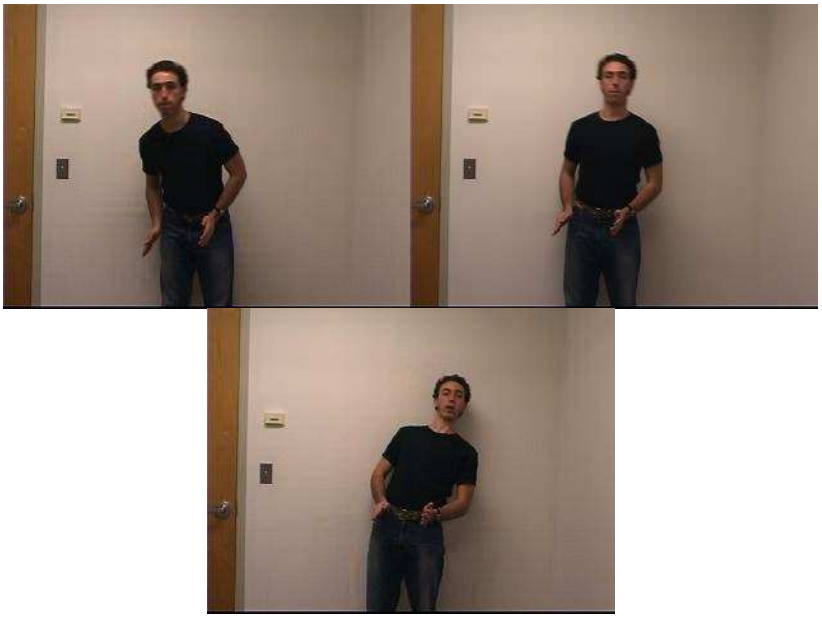


and right.

These values can also be used in conjunction with one another. For example:



forward and to the right;



and backward and to the left.

Head movement is done analogously to torso movement.

Appendix C

Inter-Annotator Agreement Code (in Perl)

```
#!/usr/bin/perl
# Inter-Annotator Agreement
#
# This script takes two .anvil files of the same
# timeline as arugments and calculates the amount of
# agreement between them.
#
# Craig Martell
#
# FORM Project, Linguistic Data Consortium
# University of Pennsylvania
#
# Joshua Kroll
#
# Computer Science Department
# Naval Postgraduate School
#
# http://www.ldc.upenn.edu/Projects/FORM/
#

use Graph;

$fps = 29.97045;

print "Usage is: \"iaa [StartTime] [EndTime] [Window] [File1] [File2]\n";

$timeStart = $ARGV[0]; ### This is to insure that all files are read to the same point.
```

```
$timeLimit = $ARGV[1]; ### Set to the ceiling of the highest time stamp of the shortest file.

$window = $ARGV[2]; ### This specifies the frame range within which matches are allowed.

### This reads in the files and create the annotation graphs for each annotator.
$graphCounter = 0;
for ($m = 3; $m <= $#ARGV; $m++) {
    open (FILE, "./$ARGV[$m]");
    print "$ARGV[$m]\n";
    $annotationGraph[$graphCounter] = Graph->new();
    $unique = 0;
    while (<FILE>) {
        $line = $_;
        $edgeAdded=0;

### The following extracts the trackname, which will be part of the edge name
        if ($line =~ /track name/) {
            ($trackName) = ($line =~ /track name="(.*?)"/);
            $trackName =~ s/ //g;
            ($highLevel,$midLevel,$lowLevel) = (split(/\./, $trackName));
        }

### This sets the nodes
        if ($line =~ /<el/){
            ($index,$start,$end)=($line =~ /index="(.*?)" start="(.*?)" end="(.*?)">/);
            $startFrame = $start * $fps;
            $endFrame = $end * $fps;
            if ($start >= $timeStart && $end <= $timeLimit){
                $annotationGraph[$graphCounter]->add_edge( "$startFrame", "$endFrame");
            }
        }

### This sets the names of the edges
        if ($line =~ /<attribute name/){
            ($attribute, $value) = ($line =~ /name="(.*?)">(.*?)</);
            if ($value =~ /unsure/){next;}
            if ($value =~ /Only select if/){$value = "Only select if . . .";}
            $totalAttribute = "$unique" . "::$trackName" . "::$attribute";
            $uncountedAttribute = "$trackName" . "$attribute";
            if ($annotationGraph[$graphCounter]->has_edge($startFrame,$endFrame)) {
                $annotationGraph[$graphCounter]->set_attribute($totalAttribute,$startFrame,
                    $endFrame,$value);
                $unique++;
            }
        }
    }
    close (FILE);
    $graphCounter++;
}

### This creates the labeled-edge arrays for each file
```

```
$arcCounter = 0;
for ($graphIndex = 0; $graphIndex <= 1; $graphIndex++) {
    my $u, $v;
    my @vertices = sort {$a <=> $b} $annotationGraph[$graphIndex]->vertices;
    print "Number of nodes (Graph $graphIndex): " . ($#vertices+1) . "\n";
    foreach $u (@vertices){
        my @successors = $annotationGraph[$graphIndex]->successors($u);
        undef %singular;
        foreach $v (@successors){
            if (!$singular{$v}){
                $singular{$v} = 1;
                %attributes = $annotationGraph[$graphIndex]->get_attributes($u, $v);
                foreach $totalAttribute (sort keys %attributes){
                    ($readUnique) = ($totalAttribute =~ /^(.*?)::/);
                    $unfixedAttribute = $totalAttribute;
                    $totalAttribute =~ s/^.*?:://;
                    $arcValue = "$u" . "::$v" . "::$totalAttribute" .
                        "::$attributes{$unfixedAttribute}";
                    $arcCounter++;
                    if ($graphIndex == 0) {
                        $attributeArray0[$readUnique] = $arcValue;
                    } elsif ($graphIndex) {
                        $attributeArray1[$readUnique] = $arcValue;
                    }
                }
            }
        }
    }
}

###Find the intersection and mark it in each graph
for ($i = 0; $i <= $#attributeArray0; $i++){
    for ($j = 0; $j <= $#attributeArray1; $j++){
        ($startFrame0, $endFrame0, $remainder0) = ($attributeArray0[$i] =~ /^(.*?)::(.*?)::(.*)$/);
        ($startFrame1, $endFrame1, $remainder1) = ($attributeArray1[$j] =~ /^(.*?)::(.*?)::(.*)$/);
        if (!defined($used1Array[$i]) && !defined($used2Array[$j]) && (abs($startFrame0 -
               $startFrame1) <= $window) && (abs($endFrame0 - $endFrame1) <= $window)
               && ($remainder0 eq $remainder1)){
            $intersection++;
            $used1Array[$i]=1;
            $used2Array[$j]=1;
        }
    }
}

print "$arcCounter arcs.\n\n";
$intersectionPercent = ((2*($intersection))/$arcCounter)*100;
$intersectionPercent = sprintf("%.2f", $intersectionPercent);
print "Exact Agreement $intersectionPercent\%\t\t\t$intersection intersections out
   of $arcCounter arcs.\n" .
   "===========================================================================\n";
```

```
### Now find and remove the intersection if we allow for a tolerance of
### +/- 1 in the value of the attribute.
### This checks to see if we are getting the same properties, and times,
### and are just disagreeing on the value.

for ($i = 0; $i <= $#attributeArray0; $i++){
    ($startFrame0, $endFrame0, $key0, $value0) = ($attributeArray0[$i] =~
                                                   /^(.*?)::(.*?)::(.*)::(.*?)$/);
    for ($j = 0; $j <= $#attributeArray1; $j++){
        ($startFrame1, $endFrame1, $key1, $value1) = ($attributeArray1[$j] =~
                                                       /^(.*?)::(.*?)::(.*)::(.*?)$/);
        if ((abs($startFrame0 - $startFrame1) <= $window) && (abs($endFrame0 - $endFrame1)
               <= $window) && $key0 eq $key1 && !defined($used2NoValueArray[$j]) &&
               !defined($used1NoValueArray[$i])){
            $used1NoValueArray[$i]=1;
            $used2NoValueArray[$j]=1;
            if ($key0 =~ /Handshape/) {
                if ($key0 =~ /Handshape letter/) {
                    $handShapeCounter++;
                    $handShapeFlag++;
                }
            } else {
                if ($value0 =~ /Only select if/) {
                    $onlySelects++;
                    $onlySelectFlag++;
                } elsif ($value0 =~ /^\d+/) {
                    ($value0) = ($value0 =~ /^(\d+)/);
                } elsif ($key0 =~ /Obscured/){
                    $obscuredFlag++;
                    $obscuredCounter++;
                }
            }
            if ($key1 =~ /Handshape/) {
                if ($key1 =~ /Handshape letter/ && $key0 !~ /Handshape letter/) {
                    $handShapeCounter++;
                    $handShapeFlag++;
                } elsif ($key1 =~ /Handshape letter/ && $key0 =~ /Handshape letter/) {
                    $variance[0]++;
                }
            } else {
                if ($value1 =~ /Only select if/ && $value0 !~ /Only select if/) {
                    $onlySelects++;
                    $onlySelectFlag++;
                } elsif ($value1 =~ /Only select if/ && $value0 =~ /Only select if/) {
                    $variance[0]++;
                } elsif ($value1 =~ /^\d+/) {
                    ($value1) = ($value1 =~ /^(\d+)/);
                } elsif ($key1 =~ /Obscured/ && $key0 !~ /Obscured/) {
                    $obscuredFlag++;
                    $obscuredCounter++;
            } elsif ($key1 =~ /Obscured/ && $key0 =~ /Obscured/) {
                $variance[0]++;
            }
                if ($handShapeFlag) {
                    undef $handShapeFlag;
                } elsif ($onlySelectFlag) {
                    undef $onlySelectFlag;
                } elsif ($obscuredFlag) {
                    undef $obscuredFlag;
                    if ($value0 eq $value1) {
                        $variance[0]++;
                    } else {
                        $obscuredCounterMissed++;
                    }
                } elsif ($value0 =~ /^\d+/ && $value1 =~ /^\d+/ && (abs($value1 − $value0)
                        == 1)) {
                    $NoValueIntersection++;
                }
            }
        }
    }
}

$intersectionNoValuePercent = ((2*($NoValueIntersection))/$arcCounter)*100;
$intersectionNoValuePercent = sprintf("%.2f", $intersectionNoValuePercent);
print "\nNo-Value Agreement $intersectionNoValuePercent\%\t\t$NoValueIntersection
  intersections out of $arcCounter arcs\n" .
  "==========================================================================\n";

if (!$obscuredCounterMissed) {
    $obscuredCounterMissed = 0;
}

print "\nNumber of unmatched \"Only select if . . .\" results is $onlySelects.\n";
print "Number of unmatched handshapes is $handShapeCounter.\n";
print "There were $obscuredCounter results in the \"Obscured\" category, of which
  $obscuredCounterMissed missed.\n";
$totalResults = $totalVariance + $onlySelects + $handShapeCounter + $obscuredCounter;
$totalNonNumeric = $onlySelects + $handShapeCounter + $obscuredCounter;
print "I am reporting $totalResults results, including $totalVariance matches and
  $totalNonNumeric non-numeric misses.\n";
print "==========================================================================\n";
```

Appendix D

Confusion Matricies

D.1 Annotation Confusion Matrices (Annotator B vs Annotator D)

D B	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1 (0-45)	6	4	3	2	1	1	0	0
2 (approx. 45)	2	8	4	1	3	0	2	0
3 (45-90)	2	4	4	0	2	0	0	0
4 (approx. 90)	1	2	2	3	8	0	3	0
5 (90-135)	2	2	4	2	4	1	0	0
6 (approx. 135)	1	0	0	1	0	4	0	2
7 (135-180)	0	3	0	0	4	2	4	2
8 (straight)	0	0	0	0	0	0	0	0

Table D.1: RightArm Forearm Location: Elbow flexion

D B	**2**	**3**	**4**	**5**
2 (supine/neutral)	2	0	0	0
3 (neutral)	0	9	5	0
4 (neutral/prone)	0	1	9	2
5 (prone)	0	0	3	0

Table D.2: RightArm Forearm Location: Forearm orientation

D **B**	**2**	**3**	**4**
2	4	5	2
3	1	11	2
4	1	0	4

Table D.3: RightArm Forearm Movement: Effort

D **B**	**1**	**2**	**3**	**4**	**5**	**6**
1 (0-45)	3	3	3	0	0	0
2 (approx. 45)	0	1	0	0	0	0
3 (45-90)	2	4	6	0	0	0
4 (approx. 90)	0	0	2	0	0	0
5 (90-135)	0	0	0	0	2	0
6 (approx. 135)	0	0	0	0	0	1

Table D.4: RightArm Forearm Movement: Elbow flexion change

D **B**	**1**	**2**	**3**	**4**
1 (inward)	4	5	0	1
2 (outward)	2	7	0	1
3 (inward-outward)	0	1	0	0
4 (outward-inward)	0	0	0	0

Table D.5: RightArm Forearm Movement: Forearm rotation

D **B**	**1**	**2**	**3**	**4**
1	0	1	0	0
2	0	3	3	0
3	1	7	16	3
4	0	0	0	0

Table D.6: RightArm HandandWrist Movement: Effort

D **B**	**3**	**5**	**Only Select If . . .**
3 (fingers moving in parallel)	0	0	2
5 (movement in sequence)	0	2	1
Only select if selected 1, 2, or 3	0	2	33

Table D.7: RightArm HandandWrist Movement: Finger coordination

D **B**	**3**	**7**
3 (A and B joint movement)	1	3
7 (direct movement between two shapes)	0	16

Table D.8: RightArm HandandWrist Movement: Hand movement

D **B**	**1**	**2**
1 (towards little finger)	1	2
2 (towards thumb)	2	4

Table D.9: RightArm HandandWrist Movement: Wrist side-to-side movement

D **B**	**1**	**2**	**4**
1 (up)	9	2	1
2 (down)	3	8	1
4 (down-up)	0	0	0

Table D.10: RightArm HandandWrist Movement: Wrist up-down movement

D **B**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	2	9	0	0	0	4	0
1	0	12	5	0	0	7	0
2	0	1	2	3	0	6	0
3	0	0	0	0	0	0	0
4	2	0	0	0	0	2	0
5	1	6	6	3	0	16	2
6	0	0	0	0	0	0	0

Table D.11: RightArm HandandWrist Shape: Handshape group

D **B**	**A**	**B**	**C**	**D**	**F**	**G**	**J**
A	3	5	3	3	0	0	0
B	0	5	0	0	0	0	0
C	0	7	0	0	0	0	0
D	1	1	0	1	0	1	0
F	1	0	0	0	0	0	0
G	1	0	0	0	0	0	0
J	4	1	0	0	0	0	0

Table D.12: RightArm HandandWrist Shape: Handshape letter

D **B**	**1**	**2**	**3**
1 (relaxed)	4	4	0
2 (slightly tense)	6	20	3
3 (very tense)	0	0	0

Table D.13: RightArm HandandWrist Shape: Tension

D B	1	2	3	4
1 (towards thumb)	1	1	0	1
2 (neutral)	8	19	3	1
3 (towards little finger)	1	1	0	1
4 (extremely towards little finger)	0	0	0	0

Table D.14: RightArm HandandWrist Shape: Wrist bend: side to side

D B	1	2	3	4
1 (up)	1	5	0	0
2 (up-neutral)	1	7	2	0
3 (neutral)	0	7	10	3
4 (down-neutral)	0	0	0	1

Table D.15: RightArm HandandWrist Shape: Wrist bend: up and down

D B	1	2
1 (forward)	25	0
2 (backward)	1	0

Table D.16: RightArm UpperArm Location: Bicep: Forward/Backward

D B	**1**	**2**
1 (inward)	20	4
2 (outward)	2	4

Table D.17: RightArm UpperArm Location: Bicep: Inward/Outward

D B	**1**	**2**
1 (upward)	21	0
2 (downward)	1	5

Table D.18: RightArm UpperArm Location: Bicep: Upward/Downward

D B	**2**	**3**	**4**	**5**	**6**
2 (inward)	0	1	0	0	0
3 (front)	0	6	1	0	0
4 (front-outward)	0	4	10	0	0
5 (outward (in frontal plane))	0	0	1	4	0
6 (behind)	0	0	0	0	5

Table D.19: RightArm UpperArm Location: Relative elbow position

D B	**1**	**2**	**3**	**4**	**5**	**6**
1 (no lift)	1	1	1	0	1	0
2 (0-45)	0	7	4	4	1	0
3 (approx. 45)	0	3	6	4	0	0
4 (45-90)	0	2	4	6	3	0
5 (approx. 90)	0	1	0	0	6	1
6 (90-135)	0	1	0	0	2	3

Table D.20: RightArm UpperArm Location: Upper arm lift

D **B**	**1**	**2**	**3**	**4**
1	0	1	0	0
2	0	2	4	0
3	0	2	18	0
4	0	0	0	2

Table D.21: RightArm UpperArm Movement: Effort

D **B**	**1**	**2**
1 (inward)	4	2
2 (outward)	2	8

Table D.22: RightArm UpperArm Movement: Linear movement (X plane)

D **B**	**1**	**2**	**4**
1 (up)	9	0	1
2 (down)	0	6	0
4 (down-up)	2	0	4

Table D.23: RightArm UpperArm Movement: Linear movement (Y plane)

D B	1	2	3
1 (towards)	4	0	0
2 (away)	2	9	0
3 (towards-away)	2	4	0

Table D.24: RightArm UpperArm Movement: Linear movement (Z plane)

D B	1	2
1 (inward)	5	1
2 (outward)	1	7

Table D.25: RightArm UpperArm Movement: Rotation direction

D B	1	2	3	4	5	7
1 (0-45)	3	0	0	0	0	0
2 (approx. 45)	0	1	0	0	0	0
3 (45-90)	1	0	3	1	0	0
4 (approx. 90)	0	1	0	1	1	0
5 (90-135)	1	0	0	1	0	0
7 (135-180)	0	0	0	0	0	1

Table D.26: RightArm UpperArm Movement: Upper arm rotation

D B	3	4	5
3 (center)	4	0	1
4 (right)	1	0	1
5 (far right)	0	0	0

Table D.27: Torso Orientation: Vertical axis orientation (to trans. channel)

D.2 Intra-Annotator Agreement Confusion Matrices

For all tables in this section, annotator A is the same as annotator D. The D annotation was done about a year after the A annotation.

A \ D	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1 (0-45)	6	6	3	2	2	0	1	0
2 (approx. 45)	5	6	4	1	2	1	1	0
3 (45-90)	5	6	5	6	10	3	1	1
4 (approx. 90)	1	0	2	0	5	0	3	0
5 (90-135)	1	3	2	4	3	2	0	1
6 (approx. 135)	0	0	0	0	0	0	0	0
7 (135-180)	0	0	0	0	0	0	0	0
8 (straight)	0	2	0	0	4	2	5	2

Table D.28: RightArm Forearm Location: Elbow flexion

A \ D	**2**	**3**	**4**	**5**
2 (supine/neutral)	2	0	0	0
3 (neutral)	1	4	1	0
4 (neutral/prone)	0	5	11	3
5 (prone)	0	1	6	0

Table D.29: RightArm Forearm Location: Forearm orientation

A \ D	**2**	**3**	**4**
2	0	1	0
3	7	11	6
4	0	3	1

Table D.30: RightArm Forearm Movement: Effort

D A	3	4
3 (45-90)	0	2
4 (approx. 90)	0	0

Table D.31: RightArm Forearm Movement: Elbow flexion

D A	1	2	3	4	5	6	7
1 (0-45)	3	1	2	0	1	0	0
2 (approx. 45)	3	2	4	0	0	0	0
3 (45-90)	1	2	2	0	0	0	0
4 (approx. 90)	0	0	2	0	0	0	0
5 (90-135)	0	1	1	0	1	0	0
6 (approx. 135)	0	0	0	0	0	1	0
7 (135-180)	0	0	0	0	1	0	0

Table D.32: RightArm Forearm Movement: Elbow flexion change

D A	1	2	4
1 (inward)	6	4	1
2 (outward)	2	6	1
4 (outward-inward)	0	0	0

Table D.33: RightArm Forearm Movement: Forearm rotation

D A	1	2	3	4
1	0	0	2	0
2	1	1	5	1
3	1	7	14	2
4	0	4	3	1

Table D.34: RightArm HandandWrist Movement: Effort

D A	**5**	**Only Select If . . .**
5 (movement in sequence)	2	1
Only select if selected 1, 2, or 3	1	41

Table D.35: RightArm HandandWrist Movement: Finger coordination

D A	**3**	**7**
3 (A and B joint movement)	0	1
7 (direct movement between two shapes)	1	16

Table D.36: RightArm HandandWrist Movement: Hand movement

D A	**1**	**2**
1	41	0
2	1	0

Table D.37: RightArm HandandWrist Movement: Strokes

D A	**1**	**2**
1 (towards little finger)	6	2
2 (towards thumb)	1	4

Table D.38: RightArm HandandWrist Movement: Wrist side-to-side movement

D **A**	**1**	**2**	**4**
1 (up)	12	5	1
2 (down)	6	8	0
4 (down-up)	0	0	0

Table D.39: RightArm HandandWrist Movement: Wrist up-down movement

D **A**	**0**	**1**	**2**	**3**	**5**	**6**
0	2	0	0	0	1	0
1	0	6	8	0	9	0
2	0	10	4	0	4	0
3	0	0	0	0	0	0
5	1	3	8	2	18	2
6	0	0	0	0	3	4

Table D.40: RightArm HandandWrist Shape: Handshape group

D **A**	**A**	**B**	**C**	**D**	**G**
A	7	8	0	2	0
B	0	3	3	0	0
C	1	6	1	0	0
D	1	3	0	3	0
G	0	1	0	0	1

Table D.41: RightArm HandandWrist Shape: Handshape letter

D **A**	**1**	**2**	**3**
1 (relaxed)	8	7	0
2 (slightly tense)	4	16	2
3 (very tense)	0	2	1

Table D.42: RightArm HandandWrist Shape: Tension

D **A**	**1**	**2**	**3**	**4**
1 (towards thumb)	8	8	0	1
2 (neutral)	3	9	0	1
3 (towards little finger)	0	4	3	0
4 (extremely towards little finger)	1	1	0	1

Table D.43: RightArm HandandWrist Shape: Wrist bend: side to side

D **A**	**1**	**2**	**3**	**4**
1 (up)	2	0	0	0
2 (up-neutral)	0	13	3	1
3 (neutral)	0	2	6	3
4 (down-neutral)	0	5	2	3

Table D.44: RightArm HandandWrist Shape: Wrist bend: up and down

D **A**	**1**	**2**
1 (inward)	30	11
2 (outward)	0	0

Table D.45: RightArm UpperArm Location: Bicep: Inward/Outward

D **A**	**1**	**2**
1 (upward)	27	1
2 (downward)	0	5

Table D.46: RightArm UpperArm Location: Bicep: Upward/Downward

D A	2	3	4	5	6
2 (inward)	0	1	0	0	0
3 (front)	0	7	3	1	0
4 (front-outward)	0	5	11	0	1
5 (outward (in frontal plane))	0	1	1	2	2
6 (behind)	0	0	0	0	5

Table D.47: RightArm UpperArm Location: Relative elbow position

D A	1	2	3	4	5	6	7
1 (no lift)	0	2	1	0	0	0	0
2 (0-45)	1	5	1	0	2	0	0
3 (approx. 45)	0	5	7	5	0	0	0
4 (45-90)	0	2	4	7	0	0	0
5 (approx. 90)	0	0	0	1	2	0	0
6 (90-135)	0	0	0	1	5	2	0
7 (approx. 135)	0	0	0	0	0	2	0

Table D.48: RightArm UpperArm Location: Upper arm lift

D A	2	3	4
2	2	5	0
3	5	21	2
4	0	0	0

Table D.49: RightArm UpperArm Movement: Effort

D A	1	2	4
1 (inward)	6	3	0
2 (outward)	2	10	0
4 (outward-inward)	1	0	0

Table D.50: RightArm UpperArm Movement: Linear movement (X plane)

D **A**	**1**	**2**	**4**
1 (up)	11	0	4
2 (down)	3	4	3
4 (down-up)	0	0	1

Table D.51: RightArm UpperArm Movement: Linear movement (Y plane)

D **A**	**1**	**2**	**4**
1 (towards)	9	4	0
2 (away)	2	13	0
4 (away-towards)	0	1	0

Table D.52: RightArm UpperArm Movement: Linear movement (Z plane)

D **A**	**1**	**2**
1 (inward)	6	3
2 (outward)	1	6

Table D.53: RightArm UpperArm Movement: Rotation direction

D **A**	**1**	**2**	**3**	**4**	**7**	**8**
1 (0-45)	4	3	0	2	0	0
2 (approx. 45)	2	0	0	1	0	0
3 (45-90)	0	0	1	1	0	0
4 (approx. 90)	0	0	2	0	0	0
7 (135-180)	0	0	0	0	0	0
8 (approx. 180)	0	0	0	0	1	0

Table D.54: RightArm UpperArm Movement: Upper arm rotation

D A	**2**	**3**	**4**
2	0	0	0
3	1	2	1
4	0	0	0

Table D.55: Torso Movement: Effort

D A	**1**	**2**
1 (forward)	2	2
2 (backward)	1	2

Table D.56: Torso Movement: Front-back movement

D A	**1**	**2**
1 (left)	1	0
2 (right)	0	1

Table D.57: Torso Movement: Side-to-side movement

D **A**	**1**	**2**	**3**
1 (0-45)	2	0	0
2 (approx. 45)	0	0	1
3 (45-90)	0	0	1

Table D.58: Torso Movement: Vertical axis rotation

D **A**	**2**	**4**
2 (center)	6	0
4 (far forward)	0	1

Table D.59: Torso Orientation: Front-back orientation

D **A**	**2**	**3**
2 (left)	1	0
3 (center)	0	6

Table D.60: Torso Orientation: Side-to-side orientation

D **A**	**3**	**4**	**5**
3 (center)	7	2	1
4 (right)	3	2	1
5 (far right)	0	0	0

Table D.61: Torso Orientation: Vertical axis orientation (to trans. channel)

Appendix E

McNemar's Test Results

Prep	**Correct**	**Error**	**p-Value**
Correct	62	4	0.39
Error	8	30	
Stroke	**Correct**	**Error**	**p-Value**
Correct	115	9	1.0
Error	8	75	
Retraction	**Correct**	**Error**	**p-Value**
Correct	56	6	0.50
Error	3	10	
All	**Correct**	**Error**	**p-Value**
Correct	233	19	0.87
Error	19	115	

Table E.1: L1000 (along the top) vs. FixedGrid (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	59	7	1.0
Error	8	30	
Stroke	**Correct**	**Error**	**p-Value**
Correct	118	6	0.60
Error	9	74	
Retraction	**Correct**	**Error**	**p-Value**
Correct	56	6	1.0
Error	5	8	
All	**Correct**	**Error**	**p-Value**
Correct	233	19	0.75
Error	22	112	

Table E.2: L500 (along the top) vs. FixedGrid (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	62	4	0.26
Error	9	29	
Stroke	**Correct**	**Error**	**p-Value**
Correct	114	10	0.68
Error	13	70	
Retraction	**Correct**	**Error**	**p-Value**
Correct	56	6	1.0
Error	5	8	
All	**Correct**	**Error**	**p-Value**
Correct	232	20	0.38
Error	27	107	

Table E.3: S1000 (along the top) vs. FixedGrid (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	56	10	1.0
Error	9	29	
Stroke	**Correct**	**Error**	**p-Value**
Correct	112	12	0.66
Error	9	74	
Retraction	**Correct**	**Error**	**p-Value**
Correct	55	7	0.77
Error	5	8	
All	**Correct**	**Error**	**p-Value**
Correct	223	29	0.48
Error	23	111	

Table E.4: S500 (along the top) vs. FixedGrid (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	65	5	0.46
Error	2	32	
Stroke	**Correct**	**Error**	**p-Value**
Correct	118	5	0.42
Error	9	75	
Retraction	**Correct**	**Error**	**p-Value**
Correct	58	1	0.62
Error	3	13	
All	**Correct**	**Error**	**p-Value**
Correct	241	11	0.70
Error	14	120	

Table E.5: L500 (along the top) vs. L1000 (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	66	4	1.0
Error	5	29	
Stroke	**Correct**	**Error**	**p-Value**
Correct	117	6	0.46
Error	10	74	
Retraction	**Correct**	**Error**	**p-Value**
Correct	56	3	0.72
Error	5	11	
All	**Correct**	**Error**	**p-Value**
Correct	239	13	0.30
Error	20	114	

Table E.6: S1000 (along the top) vs. L1000 (down the side).

Prep	Correct	Error	p-Value
Correct	60	10	0.30
Error	5	29	
Stroke	**Correct**	**Error**	**p-Value**
Correct	112	11	0.82
Error	9	75	
Retraction	**Correct**	**Error**	**p-Value**
Correct	55	4	1.0
Error	5	11	
All	**Correct**	**Error**	**p-Value**
Correct	227	25	0.46
Error	19	115	

Table E.7: S500 (along the top) vs. L1000 (down the side).

Prep	Correct	Error	p-Value
Correct	64	3	0.38
Error	7	30	
Stroke	**Correct**	**Error**	**p-Value**
Correct	119	8	0.80
Error	8	72	
Retraction	**Correct**	**Error**	**p-Value**
Correct	57	4	0.72
Error	4	10	
All	**Correct**	**Error**	**p-Value**
Correct	240	15	0.60
Error	19	112	

Table E.8: S1000 (along the top) vs. L500 (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	59	8	0.80
Error	6	31	
Stroke	**Correct**	**Error**	**p-Value**
Correct	113	14	0.28
Error	8	72	
Retraction	**Correct**	**Error**	**p-Value**
Correct	57	4	1
Error	3	11	
All	**Correct**	**Error**	**p-Value**
Correct	229	26	0.22
Error	17	114	

Table E.9: S500 (along the top) vs. L500 (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	64	7	0.07
Error	1	32	
Stroke	**Correct**	**Error**	**p-Value**
Correct	112	15	0.30
Error	9	71	
Retraction	**Correct**	**Error**	**p-Value**
Correct	58	3	1.0
Error	2	12	
All	**Correct**	**Error**	**p-Value**
Correct	234	25	0.049
Error	12	115	

Table E.10: S500 (along the top) vs. S1000 (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	63	9	0.14
Error	3	29	
Stroke	**Correct**	**Error**	**p-Value**
Correct	115	7	0.80
Error	9	76	
Retraction	**Correct**	**Error**	**p-Value**
Correct	57	4	1.0
Error	5	9	
All	**Correct**	**Error**	**p-Value**
Correct	235	20	0.74
Error	17	114	

Table E.11: FixedGrid (along the top) vs. SnoVQ (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	65	7	0.78
Error	5	27	
Stroke	**Correct**	**Error**	**p-Value**
Correct	113	9	1.0
Error	10	75	
Retraction	**Correct**	**Error**	**p-Value**
Correct	57	4	0.68
Error	2	12	
All	**Correct**	**Error**	**p-Value**
Correct	235	20	0.74
Error	17	114	

Table E.12: L1000 (along the top) vs. SnoVQ (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	63	9	0.27
Error	4	28	
Stroke	**Correct**	**Error**	**p-Value**
Correct	115	7	0.36
Error	12	73	
Retraction	**Correct**	**Error**	**p-Value**
Correct	59	2	0.62
Error	2	12	
All	**Correct**	**Error**	**p-Value**
Correct	237	18	0.87
Error	18	113	

Table E.13: L500 (along the top) vs. SnoVQ (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	67	5	1.0
Error	4	28	
Stroke	**Correct**	**Error**	**p-Value**
Correct	115	7	0.36
Error	12	73	
Retraction	**Correct**	**Error**	**p-Value**
Correct	59	2	0.62
Error	2	12	
All	**Correct**	**Error**	**p-Value**
Correct	241	14	0.60
Error	18	113	

Table E.14: S1000 (along the top) vs. SnoVQ (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	63	9	0.07
Error	2	30	
Stroke	**Correct**	**Error**	**p-Value**
Correct	110	12	1.0
Error	11	74	
Retraction	**Correct**	**Error**	**p-Value**
Correct	58	3	1.0
Error	2	12	
All	**Correct**	**Error**	**p-Value**
Correct	231	24	0.20
Error	15	116	

Table E.15: S500 (along the top) vs. SnoVQ (down the side).

Prep	**Correct**	**Error**	**p-Value**
Correct	35	16	1.0
Error	15	35	
Stroke	**Correct**	**Error**	**p-Value**
Correct	57	31	0.09
Error	18	22	
Retraction	**Correct**	**Error**	**p-Value**
Correct	41	10	0.18
Error	4	4	
All	**Correct**	**Error**	**p-Value**
Correct	133	57	0.05
Error	37	61	

Table E.16: mocapNoVQ (along the top) vs. SnoVQ-Craig (down the side).

Appendix F

Hidden Markov Models

Hidden Markov models have become one standard method in the toolbox of the statistical modeler, particularly for speech recognition. [1], [50], [24] [93], [12] and [47] all have excellent overview presentations. [11] gives probably the earliest presentation of their usefulness for speech and natural-language processing. Although all of the above have influenced this section, the presentation will most closely follow that given in [69][1].

To understand a hidden Markov model, we should first understand a Markov model. A Markov model starts with a is a series of *non*-independent random variables such that the value of each variable is dependent solely on the value of the previous variable in the sequence.

> For example, if the random variables measure the number of books in the university library, then, knowing how many books were in the library today might be an adequate predictor of how many books there will be in the library tomorrow . . . [W]e don't really need to additionally know how many books the library had last week, let alone last year. That is, future elements are conditionally independent of past elements, given the present element ([69], p. 318).

This set of random variables, then, constitute a Markov chain since they abide by the Markov assumption:

[1] In particular, see pages 317–336.

$$P(X_{n+1} = s_i | X_1, \ldots, X_n) = P(X_{n+1} = s_{n+1} | X_n), \tag{F.1}$$

where the set $S = \{s_1, \ldots, s_k\}$ defines some finite state space.

We can fully specify a Markov model, then, by two data structures:

- A, a state transition matrix, the entries of which, a_{ij}, represent

  $P(X_{n+1} = s_j | X_n = s_i)$,

  such that, $\forall_{ij} a_{ij} \geq 0$, and $\forall_i \sum_{j=1}^{N} a_{ij} = 1$; and

- Π, a table of representing that probability of different initial states of the Markov chain such that $\pi_i = P(X_1 = s_i)$ and $\sum_{i=1}^{N} \pi_i = 1$.

It should be clear by the above that a Markov model is simply a non-deterministic finite state machine such that there are probabilistic arcs between the states—given by A—and such that each state is a potential start state—as determined by Π. Additionally, this is a *visible* Markov model since the sequence of states can be known and the labels of the states can be considered an output of the machine. In a hidden Markov model (HMM), the state sequence is not known. Additionally, the emissions that constitute the output of the machine are distinct from the states. At each state, there is a probability distribution over all output emissions. That is, each state is capable of emitting each of the possible outputs according to a given probability distribution.

An HMM, then, is specified by a five-tuple $\langle S, K, \Pi, A, B \rangle$, where S defines a finite state space, K defines an output alphabet, A and Π are as above, and B is a symbol emission matrix such that b_{ijk} is the probability of emitting symbol k when transitioning from state i to state j.

F.1 Three Problems

Given an HMM as specified above, we can define three problems concerning it, the Evaluation Problem, the Decoding Problem, and the Learning Problem[2].

- **The Evaluation Problem** is concerned with determining the probability of an observation sequence, **O**, given a fully specified hidden Markov model, **M**. That is, we want to know $P(\mathbf{O}|\mathbf{M})$.
- **The Decoding Problem** is concerned with determining the most likely sequence of states, **S**, given a fully-specified hidden Markov model, **M**, and an observation sequence, **O**.
- **The Learning Problem** is concerned with determining settings for A, B, and Π, given S, K and a corpus of observation sequences, O_i.

For our experiments, we are concerned only with the Evaluation Problem and the Learning Problem. As such, we will only present the these algorithms here.

F.2 The Evaluation Problem

Again, our goal for this problem is to efficiently compute how likely an observation sequence **O** is given a model M with parameters $\mu = (A, B, \Pi)$. That is we want to know $P(O|\mu)$. To do this we first note that **O** is a sequence of observations: $\mathbf{O} = (o_1, \ldots, o_T)$. Next we note that for any sequence of states $\mathbf{X} = (X_1, \ldots, X_{T+1})$,

$$P(\mathbf{O}|\mathbf{X}, \mu) = \prod_{t=1}^{T} P(o_t|X_t, X_{t+1}, \mu) = b_{X_1X_2o_1} b_{X_2X_3o_2} \ldots b_{X_TX_{T+1}o_T} \tag{F.2}$$

and,

$$P(\mathbf{X}|\mu) = \pi_{X_1} a_{X_1X_2} a_{X_2X_3} \ldots a_{X_TX_{T+1}}. \tag{F.3}$$

[2] This section is based on the presentation given in [24], pp. 130–138.

By Bayes Rule,

$$P(\mathbf{O},\mathbf{X}|\mu) = P(\mathbf{O}|\mathbf{X},\mu)P(\mathbf{X}|\mu). \tag{F.4}$$

Therefore, we get

$$P(\mathbf{O}|\mu) = \sum_{X} P(\mathbf{O}|\mathbf{X},\mu)P(\mathbf{X}|\mu) = \sum_{X_1 \ldots X_{T+1}} \pi_{X_1} \prod_{t=1}^{T} a_{X_t X_{t+1}} b_{X_t X_{t+1} o_t}. \tag{F.5}$$

Direct evaluation this expression takes exponential time. However, since its evaluation entails computing the same probabilities many times over, we can reduce the complexity by using dynamic programming. This is done via a square array—called a lattice—where the rows are labeled with the names of states and the columns are labeled for time-steps. So, for each node in the lattice, we compute the probability of being in that state at that time step. These probabilities are computed only with reference to the probabilities of being in each state in the prior time-step.

F.2.1 The Forward Procedure

The forward procedure utilizes this lattice as follows. For each state s_i and time-step t, we compute

$$\alpha_i(t) = P(o_1 o_2 \ldots o_{t-1}, X_t = i|\mu). \tag{F.6}$$

This *forward variable* is stored at lattice location (s_i, t). It expresses the total probability of ending up in state s_i at time t given that the sequence $o_1 \ldots o_{t-1}$ was observed. This is calculated by summing over all the values of all nodes in the column $t-1$ along with the probabilities that s_i will be the next state for each. We do this left to right as follows. We first initialize the column $t = 1$:

$$\alpha_i(1) = \pi_i, 1 \leq i \leq N \tag{F.7}$$

For all subsequent columns we compute

$$\alpha_j(t+1) = \sum_{i=1}^{N} \alpha_i(t) a_{ij} b_{ijo_t}, 1 \leq t \leq T, 1 \leq j \leq N. \tag{F.8}$$

Finally, we compute our original goal as follows:

$$P(\mathbf{O}|\mu) = \sum_{i=1}^{N} \alpha_i(T+1). \tag{F.9}$$

F.2.2 The Backward Procedure

We need not only store results from left to right. We could also compute from right to left. This would give us, for each node, the probability of seeing the rest of the observation sequence from our current state at our current time, viz. (s_i, t). To do this, we, similarly, define a *backward variable*:

$$\beta_i(t) = P(o_t \ldots o_T | X_t = i, \mu). \tag{F.10}$$

We calculate these backward variables from right to left as follows. We first initialize the column $t = T + 1$:

$$\beta_i(T+1) = 1, 1 \leq i \leq N. \tag{F.11}$$

Remember, that $\beta_i(t)$ is the probability of seeing the rest of the observation from our current state at our current time. But, for any state in the last column of the lattice, we have have already seen the whole sequence. So the probability of seeing the rest of the sequence is 1.

Next, moving to the left, we compute the prior columns as follows:

$$\beta_i(t) = \sum_{j=1}^{N} a_{ij} b_{ijo_t} \beta_j(t+1), 1 \leq t \leq T, 1 \leq i \leq N. \qquad \text{(F.12)}$$

Finally, we get

$$P(\mathbf{O}|\mu) = \sum_{i=1}^{N} \pi_i \beta_i(1). \qquad \text{(F.13)}$$

F.2.3 Putting the Two Together: Forward-Backward

Looking again at our two results, Equations F.9 and F.13, we can see that they are both just special cases of the following of

$$P(\mathbf{O}|\mu) = \sum_{i=1}^{N} \alpha_i(t) \beta_i(t). \qquad \text{(F.14)}$$

For F.9, we have simply assumed that $\beta_i(t) = 1$. That is, we have only looked at the probability of *getting* to S_i at time t and assumed that is guaranteed that we will finish fine. For F.13, we assumed that opposite, that $\alpha_i(t) = 1$. That is, we assumed that it was guaranteed that we got to S_i at time t and that we were are only concerned with the probability of making it to the end. Equation F.14 then represents a much better estimate of our original goal.

F.3 The Learning Problem

Now that we know how to estimate the probability of an observation given a model, we can utilize the forward and backward procedures to help with our goal of automatically estimating model parameters, $\mu = (A, B, \Pi)$. There is no known method for guaranteeing optimal settings of these parameters, but the forward-backward[3] algorithm described here can nearly always determine a good solution. It uses iterative hill-climbing to find the values that maximizes $P(\{O_1, \ldots, O_i\}|\mu)$, or

$$argmax_{\mu} P(O_{training}|\mu). \tag{F.15}$$

If our training data is representative, then maximizing here should improve the HMM's abilities on other sequences as well.

So, to begin, we don't know what our model $\mu = (A, B, \Pi)$ is. However, given some model, we can figure out the probability of some observation given that model. We can, then, look to see what state transitions and symbol emissions were used the most. By increasing the probability of these, we can revise the model to give a higher probability to these transitions and emissions.

So, let $p_t(i,j)$ be the probability of moving from state s_i to state s_j at time t given our observation sequence (for $1 \leq t \leq T, 1 \leq i,j \leq N$). It is defined as follows:

$$\begin{aligned} p_t(i,j) &= P(X_t = i, X_{t+1} = j|\mathbf{O}, \mu) \\ &= \frac{P(X_t = i, X_{t+1} = j, \mathbf{O}|\mu)}{P(\mathbf{O}, \mu)} \\ &= \frac{\alpha_i(t) a_{ij} b_{ijo_t} \beta_j(t+1)}{\sum_{m=1}^{N} \alpha_m(t) \beta_m(t)} \\ &= \frac{\alpha_i(t) a_{ij} b_{ijo_t} \beta j(t+1)}{\sum_{m=1}^{N} \sum_{n=1}^{N} \alpha_m(t) a_{mn} b_{mno_t} \beta_n(t+1)}. \end{aligned} \tag{F.16}$$

Now, if we sum over the time indexes we can represent the expected number of transitions from state s_i to state s_j in $\mathbf{O}$ as

[3] The forward-backward algorithm is a special case of Expectation Maximization and is also known as the Baum-Welsh algorithm.

$$\sum_{t=1}^{T} p_t(i,j) \tag{F.17}$$

and the total expected number of transitions out of state s_i as

$$\sum_{t=1}^{T}\sum_{j=1}^{N} p_t(i,j). \tag{F.18}$$

With these, we can choose our new model, $\hat{\mu} = (\hat{A}, \hat{B}, \hat{\Pi})$, from our old model, $\mu = (A, B, \Pi)$, as follows.

$$\hat{\pi}_i = \sum_{j=1}^{N} p_t(i,j). \tag{F.19}$$

That is, $\hat{\pi}_i$ is the expected frequency of being in state s_i at time $t = 1$.

$$\hat{a_{ij}} = \frac{\sum_{t=1}^{T} p_t(i,j)}{\sum_{t=1}^{T}\sum_{j=1}^{N} p_t(i,j)}. \tag{F.20}$$

That is, $\hat{a_{ij}}$ is the expected number of transitions from state s_i to state s_j normalized by the total number of expected transitions out of state s_i.

$$\hat{b_{ijk}} = \frac{\sum_{\{t|o_t=k, 1\leq t\leq T\}} p_t(i,j)}{\sum_{t=1}^{T} p_t(i,j)} \tag{F.21}$$

That is, $\hat{b_{ijk}}$ is the expected number of transitions from state s_i to state s_j where k was observed normalized by the total expected number of transitions between s_i and s_j.

To do the hill-climbing aspects, then, of the forward-backward algorithm, we iterate through the above procedure until growth has stopped, or until some predetermined threshold has been reached.

Appendix G

Maximum Likelihood Baselines

In this Appendix we present the results of comparing our experiments against the Maximum-Likelihood baseline. As mentioned in Section 6.2.2, the trends in these results are the same as the trends in the results described in Chapter 6. It is important to note that using this baseline always produces a recall of one. This will push the f-score unnaturally high. The results of this is that our experiments measured against this baseline have lower f-scores. However, the precision is, with one exception, significantly higher.

Method	Precision	Recall	F-Score
Baseline	**.54**	**1**	**.70**
Fixed-Grid	.65	.44	.52
L500	.66	.44	.53
L1000	.67	.45	.54
S500	.63	.43	.51
S1000	.67	.45	.54
SnoVQ	.66	.44	.53

Table G.1: Balanced Experiments

Method	Precision	Recall	F-Score
Baseline	**.45**	**1**	**.62**
S1000-Craig	.66	.49	.56
SnoVQ-Craig	.67	.50	.57
mocap1000	.55	.42	.43
mocapNoVQ	.59	.45	.51
simulatedFORM	.37	.31	.34

Table G.2: FORM vs Motion Capture Experiments

Method	Precision	Recall	F-Score
Baseline	**.36**	**1**	**.53**
1zcKeyFramesXYZ	.51	.17	.25
1zcKeyFramesXYZTh	.54	.18	.26

Table G.3: Motion-based Experiments

Method	Precision	Recall	F-Score
Baseline	**.39**	**1**	**.56**
2zcKeyFramesXYZ	.55	.36	.43
2zcKeyFramesXYZTh	.54	.35	.42

Table G.4: Motion-based Experiments

Method	Precision	Recall	F-Score
Baseline	**.45**	**1**	**.62**
6deltas	.50	.38	.43
9deltas	.52	.40	.45

Table G.5: Motion-based Experiments

Method	Precision	Recall	F-Score
Baseline	**.44**	**1**	**.61**
S1000-Paul-Lec	.70	.62	.66

Table G.6: Paul: Lecture

Method	Precision	Recall	F-Score
Baseline	**.43**	**1**	**.60**
S1000-Paul-Con	.63	.40	.49

Table G.7: Paul: Conversation

Bibliography

[1] James Allen. *Natural Language Understanding, Second Edition.* Benjamin/Cummings Publishing Company, Redwood City, CA, 1995.

[2] Roger Bakeman and John M. Gottman. *Observing Interaction: An Introduction to Sequential Analysis.* Cambridge University Press, Cambridge, 1986.

[3] Irmgard Bartenieff. *Body Movement: Coping with the Environment.* Toronto Press, Toronto, 1980.

[4] Rama Bindiganavale and Norm Badler. Motion abstraction and mapping with spatial constraints. In *Modelling and Motion Capture Techniques for Virtual Environments, International Workshop (CAPTECH'98)*, pages 70–82. Springer-Verlag, 1998.

[5] S. Bird and P. Buneman. Towards a query language for annotation graphs. In *International Conference on Language Resources and Evaluation.*, Paris, 2000. European Language Resources Association. http://citeseer.nj.nec.com/298297.html.

[6] Steven Bird and Mark Liberman. A formal framework for linguistic annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, Pennsylvania, 1999. http://citeseer.nj.nec.com/article/bird99formal.html.

[7] Cynthia Butcher and Susan Goldin-Meadow. Gesture and the transition from one- to two-word speech: When hand and mouth come together. In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[8] Terry Caelli, Adnan Amin, Robert P. W. Duin, Mohamed Kamel, and Dick de Ridder, editors. *Structural, Syntactic, and Statistical Pattern Recognition.* Springer, August 2002.

[9] Justine Cassell. Embodied conversational agents: A new paradigm for the study of gesture and for human-computer interface. In Lynn Messing and

Ruth Campbel, editors, *Gesture, Speech, and Sign.* Oxford University Press, Oxford, 1999.

[10] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. BEAT: The behavior expression animation toolkit. In Eugene Fiume, editor, *SIGGRAPH 2001, Computer Graphics Proceedings*, pages 477–486. ACM Press / ACM SIGGRAPH, 2001.

[11] Robert L. Cave and Lee P. Neuwirth. Hidden markov models for english. In John D. Ferguson, editor, *Symposium on the Application of Hidden Markov Models to Text and Speech*, pages 57–87, Princeton, New Jersey, October 1980. IDA-CRD.

[12] Eugene Charniak. *Statistical Language Learning.* MIT Press, Cambridge, Massachusetts, 1993.

[13] Stanley F. Chen and Joshua Goodman. An empirical study of smoothing techniques fo language modeling. *ACL*, 34:310–318, 1996.

[14] Diane Chi, Monica Costa, Liwei Zhao, and Norm Badler. The emote model for effort and shape. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 173–182, New York, 2000. ACM Press/Addison-Wesley Publishing Co.

[15] Diane M. Chi. *A Motion Control Scheme for ANimating Expressive Figure Movements.* PhD thesis, University of Pennsylvania, 1999.

[16] Childes/clan. http://childes.psy.cmu.edu/.

[17] Noam Chomsky. *Syntactic Structures.* Mouton, The Hague, 1957.

[18] Noam Chomsky. *The Minimalist Program.* MIT Press, Cambridge, MA, 1995.

[19] Herbert H. Clark. *Using Language.* Cambridge University Press, Cambridge, 1996.

[20] P.R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, J. Smith, I. Chen, and J. Clow. Quickset: Multimodal interaction for distributed applications. In *ACM Multimedia '97*, pages 31–40. ACM Press, 1997.

[21] Ascension Technology Corporation. Reactor2. www.ascension-tech.com.

[22] Eden Davies. *Beyond Dance: Laban's Legacy of Movement Analysis.* Brechin Books, London, 2001.

[23] Cecily Dell. *Primer for Movement Description.* Princeton University Press, Princeton, 1970.

[24] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, New York, second edition, 2001.

[25] David Efron. *Gesture and Environment*. King's Crown Press, Morningside Heights, NY, 1941.

[26] David Efron. *Gesture, Race, and Culture: A Tentative Study of Some of the Spatio-Temporal and "Linguistic" Aspects of the Gestural Behavior of Eastern Jews and Southern Italians in New York City, Living Under Similar as well as Different Environmental Conditions*. Mouton, The Hague, 1972.

[27] Jacob Eisenstein and Randall Davis. Visual and linguistic information in gesture classification. In *Presentation at the Sixth International Conference on Multimodal Interfaces*, State College, PA, October 2004.

[28] Paul Ekman and Wallace Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1:49–98, 1969.

[29] Paul Ekman and Wallace V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32:88–105, 1969.

[30] Paul Ekman and Wallace V. Friesen. *Manual for the Facial Action Coding System*. Consulting Psychologists Press, Palo Alto, 1978.

[31] Paul Ekman, Wallace V. Friesen, and Maureen O'Sullivan. Smiles when lying. *Journal of Personality and Social Psychology*, 54:414–420, 1988. Reprinted in [33], pp. 201-214.

[32] Paul Ekman, Wallace V. Friesen, and Silvan S. Tomkins. Facial affect scoring technique: A first validity study. In Shirley Weitz, editor, *Nonverbal Communication: Readins with Commentary*, pages 34–50. Oxford University Press, New York, 1974.

[33] Paul Ekman and Erika Rosenberg, editors. *What the face reveals*. Oxford University Press, New York and Oxford, 1997.

[34] Charles Elkam. Using the triangle inequality to accelerate k-means. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML'03)*, 2003.

[35] Karen Emmorey. Do signers gesture? In Lynn Messing and Ruth Campbel, editors, *Gesture, Speech, and Sign*. Oxford University Press, Oxford, 1999.

[36] Irfan Essa and Alex Pentland. A vision system for observing and extracting facial action parameters. In *Proceedings of 1994 IEEE Computer Vision and Pattern Recognition Conference*, pages 76–83, Seattle, Washington, June 1994.

[37] Christine Fellbaum, editor. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, Massachusetts, 1998.

[38] W. N. Francis and H. Kŭcera. *Manual of Information to Accompany a Standard Corpus of Present-Day Edited American English, For Use with Digital Computers.* Department of Linguistics, Brown University, Providence, RI, 1964.

[39] Michael Gleicher. Retargeting motion to new characters. In *SIGGRAPH 98 Conference Proceedings*, pages 33–42, 1998.

[40] Michael Gleicher and Peter Litwinowicz. Constraint-based motion adaption. *The Journal of Visualization and Computer Animation*, 9(2):65–94, 1998.

[41] Susan Goldin-Meadow. The development of gesture with and without speech in hearing and deaf children. In Lynn Messing and Ruth Campbel, editors, *Gesture, Speech, and Sign.* Oxford University Press, Oxford, 1999.

[42] Charles Goodwin. Gesture, aphasia, and interaction. In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[43] Ann Hutchinson Guest and Joukje Kolff. *Spatial Variations.* Number 9 in Advanced Labanotation. Dance Books, Hampshire, 2002.

[44] Zellig Harris. *Mathematical Structures of Language.* Interscience Publishers, New York, 1968.

[45] John Haviland. Pointing, gesture spaces, and mental maps. In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[46] John Hodgson. *Mastering Movement: The Life and Work of Rudolf Laban.* Routledge, New York, 2001.

[47] X.D. Huang, Y. Ariki, and M.A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh Information Technology Series. Edinburgh University Press, Edinburgh, 1990.

[48] Ann Hutchinson. *Labanotation.* James Laughlin, New York, 1954.

[49] Growth Points in Thinking-for Speaking. David mcneill and susan d. duncan. In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[50] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, Upper Saddle River, NJ, 2000.

[51] Ashish Kapoor, Yuan Qi, and Rosalind W. Picard. Fully automatic upper facial action recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, October 2003.

[52] Adam Kendon. Some relationships between body motion and speech: An analysis of an example. In Aron Wolfe Siegman and Benjamin Pope, editors, *Studies in Dyadic Communication*, pages 177–210. Pergamon Press, 1972.

[53] Adam Kendon. Movement coordination in social interaction: Some examples described. In Shirley Weitz, editor, *Nonverbal Communication: Readings with Commentary*, pages 150–168. Oxford University Press, New York, 1974.

[54] Adam Kendon. Differential perception and attentional froma in face-to-face interaction: Two problems for investigation. *Semiotica*, 24:304–315, 1978.

[55] Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance. In M. R. Key, editor, *Nonverbal Communication and Language.* Mouton, The Hague, 1980.

[56] Adam Kendon. Gesture and speech: How they interact. In John M. Wiemann and Randall P. Harrison, editors, *Nonverbal Interaction.* Sage, Beverly Hills, 1983.

[57] Adam Kendon. How gestures can become like words. In Fernando Poyatos, editor, *Cross-Cultural Perspectives in Nonverbal Communication.* C. J. Hogrefe, Toronto, 1988.

[58] Adam Kendon. An agenda for gesture studies. *Semiotic Review of Books*, 7(3):8–12, 1996.

[59] Adam Kendon. Language and gesture: Unity or duality? In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[60] Adam Kendon. Suggestions for a descriptive notation for manual gestures. Unpublished, 2000.

[61] Michael Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech 2001*, pages 1367–1370, 2001.

[62] Sotaro Kita. How representational gestures help speaking. In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[63] Robert Krauss and Uri Hadar. The role of speech-related arm/hand gestures in word retrival. In Lynn Messing and Ruth Campbell, editors, *Gesture, Speech, and Sign.* Oxford University Press, Oxford, 1999.

[64] Curtis Lebaron and Jürgen Streeck. Gestures, knowledge, and the world. In David McNeill, editor, *Language and Gesture*. Cambridge University Press, Cambridge, 2000.

[65] David Lewis. *Counterfactuals*. Blackwell, Oxford, 1973.

[66] David Lewis. *Convention: A Philosophical Study*. Blackwell, Oxford, 2002.

[67] Ilkka Linnankoski, Lea Leinonen, Minna Vihla, Maija-Liisa Laakso, and Synnve Carlson. Conveyance of emotional connotations by a single word in english. In *Speech Communication*, volume 45, pages 27–39, January 2005.

[68] Vera Maletic. *Body–Space–Expression: The Development of Rudolf Laban's Movement and Dance Concepts*. Mouton de Gruyter, Berlin, 1987.

[69] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[70] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19:313–330, 1993.

[71] Craig Martell. Form: An extensible, kinematically-based gesture annotation scheme. In *International Conference on Language Resources and Evaluation*. European Language Resources Association, 2002. http://www.ldc.upenn.edu/Projects/FORM.

[72] Rachel I. Mayberry and Joselynne Jaques. Gesture production during stuttered speech: Insights into the nature of gesture–speech integration. In David McNeill, editor, *Language and Gesture*. Cambridge University Press, Cambridge, 2000.

[73] David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.

[74] David McNeill. Triangulating the growth point—arriving at consciousness. In Lynn Messing and Ruth Campbell, editors, *Gesture, Speech, and Sign*. Oxford University Press, Oxford, 1999.

[75] David McNeill, editor. *Language and Gesture*. Cambridge University Press, Cambridge, 2000.

[76] David McNeill and Elena Levy. Conceptual representations in language activity and gesture. In R. J. Jarvella and W. Klein, editors, *Speech, Place, and Action: Studies in Dexis and Gesture*. Wiley, 1982.

[77] Mcneill lab. http://mcneilllab.uchicago.edu/topics/annotation.html.

[78] Lynn Messing and Ruth Campbell, editors. *Gesture, Speech, and Sign.* Oxford University Press, Oxford, 1999.

[79] Lynn S. Messing. Two modes—two languages? In Lynn Messing and Ruth Campbel, editors, *Gesture, Speech, and Sign.* Oxford University Press, Oxford, 1999.

[80] C. Neidle, S. Sclaroff, and V. Athitsos. Signstream: A tool for linguistic and computer vision research on visual-gestural language data. In *Behavior Research Methods, Instruments, and Computers*, volume 33:3, pages 311–320. Psychonomic Society Publications, 2001.

[81] Bruce E. Nevin, editor. *The Legacy of Zellig Harris: Language and information into the 21st century. Volume 1: Philosophy of science, syntax and semantics.* John Benjamins Publishing Company, Amsterdam/Philadelphia, 2002.

[82] Bruce E. Nevin and Stephen B. Johnson, editors. *The Legacy of Zellig Harris: Language and information into the 21st century. Volume 2: Mathematics and computability of language.* John Benjamins Publishing Company, Amsterdam/Philadelphia, 2002.

[83] Hermann Ney, Sven Martin, and Frank Vessel. Statistical language modeling using leaving-one-out. In Steve Young and Gerrit Bloothooft, editors, *Corpus-Based Methods in Language and Speech Processing*, pages 174–207. Kluwer Academic, Dordrecht, 1997.

[84] NIST. Mcnemar's test.

[85] Shuichi Nobe. Where do *Most* spontaneous representational gestures actually occur with respect to speech. In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[86] Maja Pantic and Leon Rothkrantz. An expert system for recognition of facial actions and their intensity. In *AAAI/IAAI*, pages 1026–1033, 2000.

[87] Maja Pantic, Leon Rothkrantz, and Henk Koppelaar. Automation of nonverbal communication of facial expressions.

[88] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceeding of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, July 2002.

[89] Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. In *Speech Communication*, volume 45, pages 89–95, January 2005.

[90] Kim Plunkett and Jeffrey L. Elman. *Rethinking Innateness: A Connectionist Perspective on Development.* MIT Press, Cambridge, 1996.

[91] Valerie Preston-Dunlop. *Rudolf Laban: An Extraordinary Life.* Dance Books, London, 1998.

[92] Francis Quek et al. Gestural origo and loci-transitions in natural discourse segmentation. Technical Report VISLab-01-12, Department of Computer Science and Engineering, Wright State University, 2001. http://vislab.cs.wright.edu/Publications/QueBMH01.html.

[93] L. Rabiner and B. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 1986.

[94] Philip Resnik. Wordnet and distributional analysis: a class-based approach to lexical discovery. In *Workshop Notes, Statisticall-Based NLP Techniques.* AAAI, 1992.

[95] Jan Peter De Ruiter. The production of gesture and speech. In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[96] Stuart Russel and Peter Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, Upper Saddle River, New Jersey, 2nd edition, 2003.

[97] Andrew Schein. *Active Learning for Logistic Regression.* PhD thesis, University of Pennsylvania, 2005.

[98] Richard Sproat, editor. *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach.* Kluwer Academic Publishers, Boston, 1998.

[99] V. I. Stepanov. *Alphabet of Movements of the Human Body: A Study in Recording the Movements of the Human Body by Means of Musical Signs (Translated by Raymond Lister).* Dance Horizons, Brooklyn, NY, 1969.

[100] William Stokoe. Gesture to sign (language). In David McNeill, editor, *Language and Gesture.* Cambridge University Press, Cambridge, 2000.

[101] William C. Stokoe and Mark Marschark. Signs, gestures, and signs. In Lynn Messing and Ruth Campbel, editors, *Gesture, Speech, and Sign.* Oxford University Press, Oxford, 1999.

[102] Vicon Motion Systems. Vicon mx. www.vicon.com.

[103] Talkbank project. http://www.talkbank.org.

[104] Vislab project. http://vislab.cs.wright.edu/.

[105] Tracy Westeyn, Helene Brashear, Amin Atrash, and Thad Starner. Georgia tech gesture toolkit: Supporting experiments in gesture recognition. In *International Conference on Perceptive and Multimodal User Interfaces*, 2003.

[106] Peter Wittenburg, Stephen C. Levinson, Sotaro Kita, and Hennie Brugman. Multimodal annotations in gesture and sign language studies. In *International Conference on Language Resources and Evaluation*. European Language Resources Association, 2002.

[107] S. Young. The htk hidden markov model toolkit: Design and philosophy, 1993.

[108] Liwei Zhao. *Synthesis and Acquisition of Laban Movement Analysis Qualitative Parameters for Communicative Gestures*. PhD thesis, University of Pennsylvania, 2001.