Using FORM Data to Predict Phase Labels

Craig Martell
Computer Science Department
Naval Postgraduate School
One University Circle
Monterey, California 93943
cmartell@nps.edu

Joshua Kroll
Mathematics Department
Harvard University
One Oxford Street
Cambridge, Massachusetts 02138
jkroll@fas.harvard.edu

ABSTRACT
In this paper we present an augmentation of the FORM gesture corpus and describe experiments using FORM to predict gesture phase, i.e. preparation, stroke, and retraction. We compare these results to experiments using motion-captured data to predict the same. Interestingly, the FORM data, which is gathered via annotation, does significantly better than the motion-captured data.

Author Keywords
Multimodal Corpora, Gesture, Machine Learning

ACM Classification Keywords
H.5.2 Information Interfaces & Presentation: User Interfaces

INTRODUCTION
FORM was developed as a fine-grained, gesture coding scheme that allows annotators to capture exhaustively the constituent parts of the gestures of video-recorded speakers.

FORM represents gesture data as a collection of 4-tuples, <*startTime, endTime, attribute, value*>. The attribute/value pair represents some change during the specified interval. For example, if there was upper-arm rotation during an interval, the attribute would be Upper Arm: Rotation, and the value would be the degree of rotation. All of the possible attribute/value pairs are described extensively in [5]. It is useful to think of these 4-tuples as labeled arcs in a graph, the nodes of which are the timestamps. In FORM, gestural movement is segmented visually. That is, the annotators would focus first on one attribute in order to mark the timestamps of changes, and then replay the video to focus on the next attribute.

The total FORM dataset is approximately 22 minutes long. There are approximately 3500 arcs/minute, for a total of roughly 77000 arcs.

In [4], we presented preliminary results and discussed future research directions. In this paper, we describe refinements to the FORM annotation scheme and present the results of new inter– annotator-agreement studies and machine-learning experiments using the FORM dataset to predict gesture phases.

OVERCOMING AMBIGUITIES IN FORM
There are known ambiguities in the FORM system as described in [4] and in greater detail in [5]. One concerns the *Upper Arm*: *Location* attributes that specify biceps direction. While anatomically it seems accurate to describe the upper arm rotation by degrees of rotation rather than by the direction of the biceps in free space—as is done in FORM—a problem arises when defining the neutral position of the arm rotation.

In light of this ambiguity, we have 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, with (3, 3, 3) being the speaker's sternum. 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 the upper-arm lift—at key points along the movement. This should suffice to recreate the motion.

INTER-ANNOTATOR AGREEMENT: THE BAG OF ARCS METHOD
Our experiments with FORM-annotation show that with sufficient training, agreement among the annotators can be very high. Table 1 shows inter-annotator agreement results for two annotators annotating a file of four gesture excursions. The results were generated by the bag-of-arcs algorithm, as given in [5]. Essentially, given an annotation graph, we combine all the arcs for each annotator into a bag. Then all the bags are combined and 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.

We calculate the intersection with tolerances for time and value chosen, as described below. Each of the annotators agreed that there were four gesture excursions. The *Precision* column gives the number of frames (at 29.97045 fps) that the annotators can be off from one another by and still be counted as having agreed. A precision of 0 frames

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 1. Inter-Annotator Agreement on Jan24-09.mov

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, we first loosened this restriction to within 7 frames (or approximately ± .25 seconds) and then to 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. As examples, let *arc1* be (428, 446, ForearmRotation, 1(and *arc2* be (427, 451, ForearmRotation, 2). Then *arc1* will match *arc2* if the tolerances are set to *frames* = 15 and *Off-by-one-or-less* = *True*. However, they will not match if *frames* = 0 or if, instead of *Off-by-one-or-less*, *Exact Match* = *True*. The Bag-of-Arcs method is similar to the one used by the IBM BLEU project to judge quality of a machine translation [7]. The FACS project also used a similar metric. They let two facial encodings match along a particular dimension if the first choice of one annotator was the first or second choice of another annotator [1].

The most tolerant measure, then, is given by the (15 *frames, Off-by-one-or-less*) cells. 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 intra-annotator agreement was 81.29%).

GESTURE SEGMENTATION: PHASES

In this section, we present experiments which use the FORM representation of gesture—which is fairly *low-level*—to predict the *medium-level* phenomenon of gesture-phase. We have purposively avoided defining what constitute 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. This is done with surprising consistency. Similarly, 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* of that gesture, and *PSR theory* the theory that says excursions can be so divided.

Inter-Annotator Agreement: Phases

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 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

	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 2. Agreement 68.28%

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

Table 3. Agreement 90.42%

the agreement score would be accurate. Tables 2 and 3 present the results of these experiments.

Table 2 is particularly interesting. This presents the result of judging agreement over *all* phase categories, including unsures. Note that the total agreement over all frames was only 68.28%. This low number is largely explained by how Unsures are used, as described above. The annotator represented by the row labels used Unsure much more often. However, we can see that—although there was strong consistency for Preparations, Strokes, and Retractions—there was also more confusion concerning Holds. 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 reveals that in many of these cases the speaker's hand are performing what we call "incidental movement." Incidental movement is movement during a phase that seems cognitively to be a Hold, even though there is some bouncing or jittery movement of the hand. Some annotators paid 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.

Thus, we ran the agreement study again, but only judged agreement on Preparations, Strokes, and Retractions. Overall agreement across these three phases was 90.42% (Table 3). As Holds are presumably important for understanding human gesturing, more work is warranted so that we can consistently annotate Hold phases.

AUTOMATIC PHASE PREDICTION: FORM VS. MOCAP

In this section we describe the results of using hidden Markov models (HMMs) to predict phase labels from the underlying kinetic representation in FORM. We conducted a number of experiments which are described extensively in [5]. In addition, for some experiments, the subject in the video was connected to a ReActor2 infrared motion-capture system. This was done so that we could compare FORM and motion-capture as different methods of gathering human gestural-movement information. Motion capture was chosen for comparison because it is considered "ground truth" for capturing bodily movement information. The best results from each of FORM and motion capture are presented below.

Experimental Overview

As mentioned above, we overcame ambiguity in FORM by adding the end-effector position. This position was given as (x, y, z) coordinates in a $5 \times 5 \times 5$ grid. If we combine these coordinates with the value of the *upperArm-Lift* parameter, we get a vector in $\mathbf{R}^4$ which describes the position of an arm at a particular frame. So, a sequence of these vectors encode the movement of an arm throughout a gesture excursion. By dividing the excursion into subsequences of these vectors such that they are co-extensive with the phase segmentation described above, we created a set of labeled data.

However, 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, we took the R4 vectors for each *Location* point in the gesture excursion and used cubic splines to fill in the values for the intervening frames. This generated a large matrix in R4, the number of columns of which is determined by the number of frames—at 29.97045 fps—in the excursion. We then divided this large matrix in accordance with the phase segmentation to generate bins of matrices representing the different phases. Thus, we produced a bin of preparations, a bin of strokes, and a bin of retractions.

For the motion-capture experiments, we generated vectors with the same parameterization as the FORM vectors from data given by the motion-capture system. However, as the motion-capture system generated vectors for all frames of an excursion, no interpolation was necessary. We simply segmented the sequence of frames according to the human-annotated phase labels to create analogous matrices[1].

For each of these methods, we then ran the following HMM experiment. It is a version of a cross-validation method known as *Leaving-one-out* [6]. 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. 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, and Retraction, generate an HMM representing that phase and train with all the samples for that phase only. Label observation_i after the hidden Markov model, *M*, which maximizes $P(\text{observation}_i|M)$. If

[1] Other methods tried included automatic smoothing of MoCap data (SimulatedFORM) [5]. However, these results were inferior to those presented here.

	Preparation			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Prep	0.35	1.00	0.52	
FORM	0.67	0.50	0.57	+5.8%
MoCap	0.61	0.49	0.54	+3.8%
	Stroke			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Stroke	0.45	1.00	0.62	
FORM	0.72	0.73	0.72	+16%
MoCap	0.69	0.60	0.64	+3.22%
	Retraction			
	Precision	**Recall**	**F-Score**	**±Baseline**
Call-all-Retraction	0.22	1.00	0.33	
FORM	0.61	0.86	0.71	+115%
MoCap	0.46	0.76	0.57	+73%

Table 4. Precision, Recall, and F-Score Results for Various HMM Methods Using the *Craig* Data Set

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*).

Baseline: Call-all-*X*

The baseline used for these experiments was 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.

Results

Table 4 presents the results of these experiments. For all three phases, both FORM and MoCap did better than baseline. Interestingly, though, FORM did significantly better than MoCap, with a *p*-value of 0.05 using a two-tailed McNemar's test. While this obviously satisfied our original goal of being at least almost as good as motion-capture, it begs the question as to why the FORM data produced better results. Further research is needed here, but we believe that the smoothing of the movement curve imposed by FORM removes much of the incidental movement that MoCap faithfully captures. The result of this smoothing is a curve with coarse-grained features which are more easily classifiable.

REFERENCES

1. Ekman, P., Friesen, W. V., and Tomkins, S. Facial affect scoring technique: A first validity study. In *Nonverbal Communication: Readings with Commentary*, Shirley Weitz, (ed.). Oxford University Press, New York, (1974), 34–50.
2. Kendon, A.. *Gesture: Visible Action as Utterance*. Cambridge University Press, Cambridge, (2004).
3. Kita, S., van Gijn, I. and van der Hulst, H. Movement Phases in Signs and Co-speech Gestures, and Their Transcription by Human Coders. In *Gesture and Sign Language in Human-Computer Interaction*, Wachsmuth, I. and Fröhlich, M. (eds.). Springer (1998), 23-35.
4. Martell, C. Form: An extensible, kinematically-based gesture annotation scheme. In *Proceedings of the International Conference on Language Resources and Evaluation.* (2002).
5. Martell, C.. FORM: *An Experiment in the Annotation of the Kinematics of Gesture*. Ph.D. thesis, University of Pennsylvania (2005).
6. Ney, H., Martin, S, and Vessel, F. Statistical language modeling using leaving-one-out. In *Corpus-Based Methods in Language and Speech Processing*, Steve Young and Gerrit Bloothooft (eds.). Kluwer Academic, Dordrecht, (1997), 174–207.
7. Papineni, K., Roukos, S, Ward, T, and Zhu, W. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, (2002), 311–318.