Visual and Linguistic Information in Gesture Classification

Jacob Eisenstein and Randall Davis
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar St, Cambridge, MA 02139 USA
jacobe@csail.mit.edu
davis@csail.mit.edu

ABSTRACT

Classification of natural hand gestures is usually approached by applying pattern recognition to the movements of the hand. However, the gesture categories most frequently cited in the psychology literature are fundamentally multimodal; the definitions make reference to the surrounding linguistic context. We address the question of whether gestures are naturally multimodal, or whether they can be classified from hand-movement data alone. First, we describe an empirical study showing that the removal of auditory information significantly impairs the ability of human raters to classify gestures. Then we present an automatic gesture classification system based solely on an n-gram model of linguistic context; the system is intended to supplement a visual classifier, but achieves 66% accuracy on a three-class classification problem on its own. This represents higher accuracy than human raters achieve when presented with the same information.

Categories and Subject Descriptors

H.1.2 [**User-Machine Systems**]: Human information processing; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*Artifical, augmented, and virtual realities*; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology, Natural language, Theory and methods, Voice I/O*

General Terms

Human Factors, Reliability, Experimentation

Keywords

Gesture Recognition, Gesture Taxonomies, Multimodal Disambiguation, Validity

ICMI'04, October 13–15, 2004, State College, Pennsylvania, USA.

1. INTRODUCTION

A number of multimodal user interfaces afford interaction through the use of communicative free hand gestures [6, 7, 9, 11]. Since hand gestures can be used for a number of different communicative purposes–e.g., pointing at an object to indicate reference, or tracing a path of motion–*classification* of hand gestures is an important problem.

One class of systems focuses on artificial gestures, such as waving, closed fist, or "thumbs up" (e.g., [6]). These are not intended to correspond to the natural gestures that spontaneously arise during speech. For such systems, the goal is to maximize ease and speed of recognition, rather than the naturalness of the user interface. With such artificial gestures, gesture classes are distinguished purely on the basis of the dynamics of hand motion; however, mutual disambiguation with speech [16] could be used to improve recognition.

There is, however, a growing set of user interfaces that attempt to allow users to communicate using more natural gestures [7, 9, 11]. Here too, gesture classification has been taken to be primarily a problem for computer vision [9] or pattern recognition applied to glove input devices [7, 11]. Mutual disambiguation has been applied to improve recognition by constraining the gesture recognition candidates based on a set of possible semantic frames [7]. But the idea that gesture classes themselves are fundamentally multimodal entities – *defined* not only by the hand motion but also by the role of gesture within the linguistic context – has not yet been given full consideration.

We begin with a brief summary of the most frequently cited gesture taxonomy from the psychology literature; there has been some work on automatic classification for subsets of this taxonomy. Next, we present an empirical study of the ability of naïve raters to classify gestures according to this taxonomy, evaluating the effect of removing either the visual or auditory modalities. Then we present a gesture classification system that uses only the linguistic context; no hand-movement information is used.

2. TYPES OF GESTURES

Kendon describes a spectrum of gesturing behavior [8]. On one end are artificial and highly structured gestural languages, such as American Sign Language. In the middle, there are artificial but culturally shared *emblems*, such as the "thumbs-up" sign. At the far end is *gesticulation*, gestures that naturally and unconsciously co-occur with speech. Gesticulation is of particular interest for HCI since it is completely natural; speakers do not need to be taught how to

do it. However, gesticulation is challenging because of the potential for variety in gesturing behavior across speakers, particularly across cultures.

Linguists have created a taxonomy of gesticulation, and gestures that naturally co-occur with speech are now typically divided into several classes: deictic, iconic, metaphoric, beat [13]. McNeill notes that these types should not be thought of as discrete, mutually exclusive bins, but rather, as features that may be present in varying degrees, possibly in combination. Thus, identification of the extent to which each feature is present would be the ultimate goal, rather than gesture classification. For the moment, however, implemented systems have focused on classification [3, 7, 9, 11].

The following definitions are quoted and summarized from Cassell [2].

- "**Deictics** spatialize, or locate in physical space..." Deictics can refer to actual physical entities and locations, or to spaces that have previously been marked as relating to some idea or concept.

- "**Iconic** gestures depict by the form of the gesture some features of the action or event being described." For example, a speaker might say "we were speeding all over town," while tracing an erratic path of motion with one hand.

- "**Metaphoric** gestures are also representational, but the concept they represent has no physical form; instead the form of the gesture comes from a common metaphor." For example, a speaker might say, "it happened over and over again," while repeatedly tracing a circle.

- "**Beat** gestures are small baton-like movements that do not change in form with the content of the accompanying speech. They serve a pragmatic function, occurring with comments on one's own linguistic contribution, speech repairs and reported speech." Speakers that emphasize important points with a downward motion of the hand are utilizing beat gestures.

2.1 Vision and Speech

One thing to notice about the definitions of the gesture types is that they are *linguistic* in nature. That is, gesture types are defined in terms of the role they play in the discourse, rather than in terms of a specific hand trajectory or class of trajectories. Indeed, researchers have found that there is no canonical set of hand trajectories that define each gesture class. For example, Cassell states, "Deictics do not have to be pointing index fingers." [2] For non-deictic gestures, it is even harder to characterize a "typical" set of hand shapes or trajectories; there are perhaps an infinite variety of possible iconic and metaphoric gestures [18]. Clearly, some amount of linguistic evidence – prosodic, lexical, or semantic – is necessary to classify gestures.

The remainder of this paper will seek to answer two questions.

1. To what extent does our perception of gesture types depend on a visual analysis of the hand motion, and to what extent does linguistic evidence come into play?
2. Can we build an accurate gesture classification system using linguistic data? What linguistic features are most informative for this purpose?

This paper describes two experiments aimed at answering these questions. In the first, naïve participants were trained to classify gestures according to the taxonomy described above. We assessed the level of interrater agreement to show that the taxonomy presents meaningful categories. We then removed the auditory and visual modalities separately, and found that participants make significantly different ratings in the absence of either modality. In other words, neither modality alone is sufficient to classify gestures.

Next we describe a gesture classification system that considers only the text surrounding the gesture. This system is trained using the majority classifications from the human raters as ground truth. Our classifier achieves a 66% agreement on a cross-validated evaluation; this is higher than the human coders achieved when they were denied access to the visual modality.

3. CLASSIFICATION BY HUMAN RATERS

In a previous study, nine speakers were videotaped while describing the behavior of three different mechanical devices [5]. These monologues were transcribed and the gesture phrases were segmented by the experimenter. Speakers ranged in age from 22 to 28; eight were native English speakers; four were women. The devices they described were: a latchbox, a piston, and a pinball machine. None of the participants had any special expertise in physics or mechanical engineering.

A second group of participants was then asked to classify the gestures from this corpus of videos, using the categorization scheme described above. There were four types of conditions: both video and audio (VA) were available, video only (V), audio only (A), and a textual transcription of the audio with no video (T). The VA condition was presented twice, and the ordering of conditions was identical for all participants: VA, V, A, T, VA.

A permutation of the videos was used so that no participant saw the same video in more than one condition, and so that each video was used in each condition nearly an equal number of times. The ninth video, of a male native English speaker, was used for training examples, as discussed below. Only videos of the explanations of the piston device were used. Overall, each video was annotated by eight or nine different participants in the VA condition, and by four or five participants in every other condition.

In each condition, participants were required to classify every gesture in the video. The videos ranged in length from 10 to 90 seconds, and included as few as four and as many as 53 distinct gesture phrases.

The entire study was performed using automated software that required no intervention from the experimenter. Participants were able to play each gesture segment from the video whenever and as frequently as they desired. Radio buttons were used to indicate the gesture classes in a fixed order, and were not preset to any value; participants were required to classify each gesture before moving on to the next condition. The video was presented in a separate window, 300 by 400 pixels in size. Each video segment ran from the beginning to the end of the gesture phrase, as segmented by the experimenter. In the audio-only condition, a beep was used

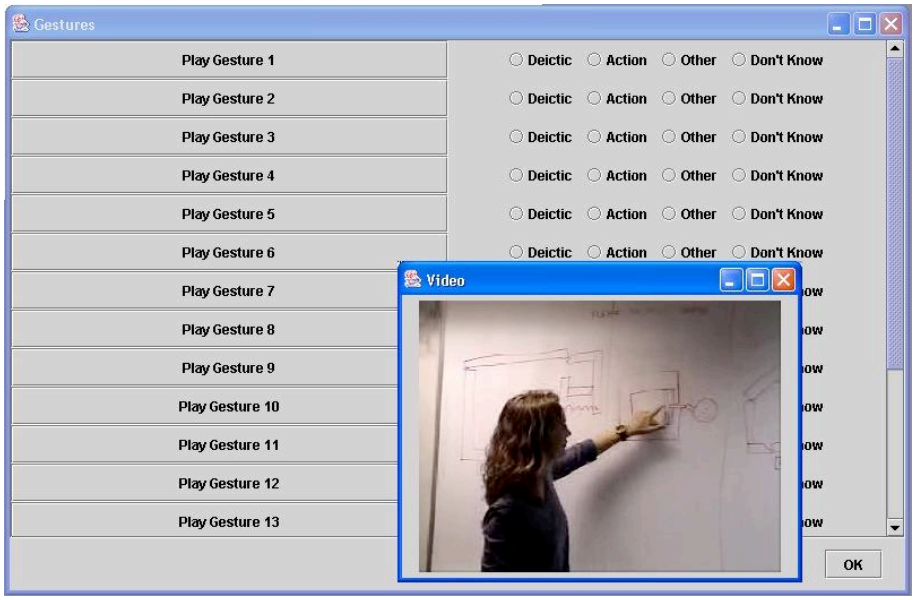

Figure 1: The experimental user interface for the VA, V, and A conditions

to indicate the onset of the stroke phase of the gesture. The user interface for the experimental tool is shown in Figure 1.

A different user interface was used for the text-only condition (Figure 2). Participants were presented with a list of the gestures (at left), while the center of the screen presented a transcript of all of the text used in a 4 second interval surrounding the onset of the stroke phase of the gesture. The location of the onset of the stroke phase was indicated in the transcript as "[GESTURE]". Radio buttons were once again used for the gesture classification.

3.1 Participants

There were 36 participants in this study; 22 men and 14 women. They ranged in age from 18 to 57, with a median of 26 and a mean of 29.3. Ten of the participants self-reported their English as being worse than that of a native speaker. Participants were recruited using posters placed around a university campus, and were compensated with free movie passes for completing the study. None of the participants had any prior experience with gestural or linguistic analysis, and all can be considered "novice" annotators. One participant was excluded because the experimental software crashed.

3.2 Instructions

Text and video examples were used to instruct participants about the gesture classification scheme. The instructions described both the kinetic 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 [5], metaphoric gestures are extremely infrequent in this corpus. A subset of participants were also allowed to classify gestures as "Unknown."

The written instructions given to participants can be found in the appendix.

4. RESULTS

The standard Kappa (κ) metric was used to assess interrater reliability [1]. In the Kappa statistic, a value of zero indicates chance agreement, and a value of one indicates perfect agreement.

A confusion matrix for the second iteration of the video-audio (VA) condition is shown in Table 1. For each condition, a confusion matrix is generated for every pair of raters, and these confusion matrices are then averaged together. Given two raters r_1 and r_2, both pairs $\langle r_1, r_2 \rangle$, and $\langle r_2, r_1 \rangle$ will be included in the average, so the resulting matrix is necessarily symmetric.

The table indicates reasonable agreement for the deictic and action categories: $\kappa = .581$ when isolating the submatrix containing only these categories. However, the labeling of the "other" category is essentially random, lowering the overall Kappa to .449 when this category is included. It is possible to compute the variance of the Kappa statistic; in this case, $\sigma = .033$, yielding better than chance agreement at $p < .01$.

The relatively low Kappa here may reflect McNeill's contention that the gesture types are not truly mutually exclusive. Another possible factor is the limited training for these participants, which typically lasted less than five minutes (see the Appendix for the raters' instructions). Interrater agreement was significantly higher in the second iteration of the VA condition than in the first iteration, where $\kappa = .273$. This suggests that the raters' assessments of the meaning of the gesture categories converged as they gained experience with the rating task. For expert raters, Nakano reports Kappa agreement of .81 using similar categories [15].

The extremely low agreement on the "other" category suggests that some raters may have used "other" whenever they were unable to classify the gesture as either "deictic" or

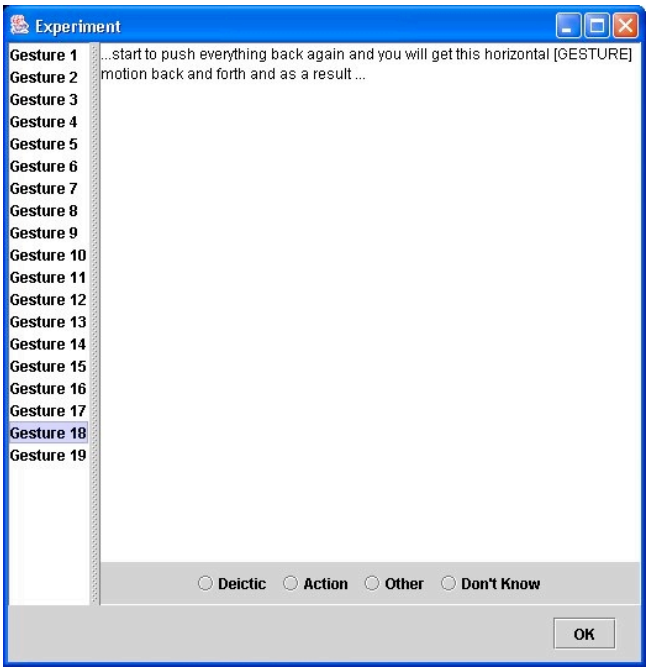

Figure 2: The experimental user interface for the text-only condition

	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 1: Confusion matrix for the second VA condition

"action." Note that this confusion matrix includes results from the sixteen participants who did not have access to the "don't know" option, as well as those who did have access to this option. The "don't know" option increased the Kappa marginally, to 0.451, but this difference is not significant ($p > .05$).

4.1 Conditions

The agreement for the audio-only (A) condition was significantly lower than the VA condition, $\kappa = .337, p < .01$. The same is true of the video-only (V) condition, $\kappa = .276$, $p < .01$, and the text condition (T), $\kappa = .315, p < .01$. However, in all cases, the Kappa value was better than chance, $p < .01$.

It may be somewhat surprising that interrater agreement was lower in the impaired conditions. One conceivable source of disagreement in the VA condition is the choice of which modality to favor when each suggests a different classification. In the impaired conditions, no such choice need be made, so one might predict that agreement within the impaired conditions would be higher. But in fact, the opposite is the case – intra-condition agreement increases when both modalities are available. This suggests that the modalities usually provide complementary cues, and that in many cases, neither modality provides enough information on its own.

Condition	Intra-condition agreement (κ)	Agreement with VA majority
VA	.451	78%
V	.276	59%
A	.335	45%
T	.315	41%

Table 2: Agreement results for each condition

We computed the majority vote classifications for each video in the second VA condition, and took this as ground truth. Then for each condition, we computed the average percentage agreement between ground truth and each rater's annotations. As an upper bound, in the VA condition, the average rater agreed with the majority annotations at a level of 78%, $\sigma = 0.018$. In the audio-only condition (A), the average agreement with the modal classifications from the VA condition is 45%, $\sigma = 0.017$. In the video-only condition (V), the average agreement is better, at 59%, $\sigma = 0.021$. In the text-only condition (T), the average agreement is 41%, $\sigma = 0.016$.

Since the video-only condition had the highest level of agreement with the VA condition, this would suggest that visual information is the primary cue for gesture classification. However, there is a statistically significant drop-off from the VA condition to the video-only condition ($p < .01$), suggesting that audio cues do play a necessary supplementary role.

5. AUTOMATIC CLASSIFICATION FROM TEXT

The previous section shows that human listeners use both vision and audition when recognizing gestures, and that two modalities contain complementary information. In this section, we explore the idea of classifying gestures using only linguistic information. The goal here is to determine what type of linguistic cues are most useful for gesture classification, to get a sense for the classification performance these cues can provide, and to develop a system that could be combined with a vision-based approach in an integrated multimodal gesture classifier. We use the majority classifications from the previous study as ground truth, and evaluate our system's ability to replicate these classifications using only textual information.

5.1 Features

For each gesture, a feature vector was constructed using the words that appear within a series of windows surrounding the onset of the stroke phase of the gesture. According to the psychology literature, the stroke phase usually overlaps the most prosodically prominent part of the associated speech [13]. We used two windows to differentiate words that appear during the stroke phase from words that appear at any point during the whole gesture phrase (see Figure 3). The windows were buffered by 133 milliseconds at the front and 83 milliseconds at the back. Ideally, these parameters should be estimated by cross-validation, but the results are not overly sensitive to their settings.

Since strokes are a component of gesture phrases, the

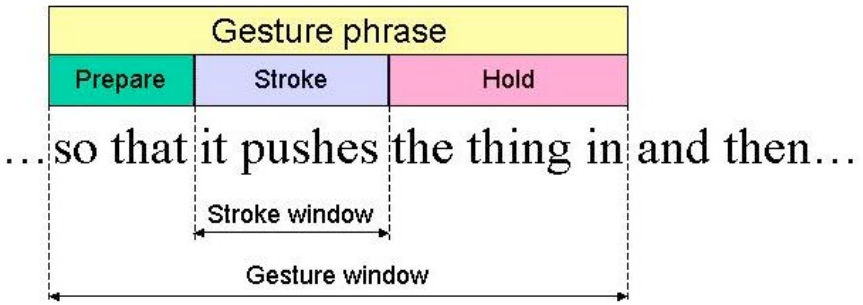

Figure 3: Separate windows are used to capture stroke and gesture phrase features

stroke window is a subset of the gesture phrase window. By including the stroke window, we are heeding McNeill's advice that the words overlapping the stroke phase are the most important for determining the semantic content of the multimodal utterance [13]. This did in fact improve performance; from 61.5% using the only the gesture phrase window, to 65.9% when using both windows. Using the stroke phase window alone produced performance of 58.7%; the multiple-window technique was significantly better than both alternatives.

The stroke window contained n-grams that were highly informative but sparse. For example, consider the part-of-speech unigram "VBZ", indicating a verb in the 3rd person singular, present tense. This feature is somewhat informative when appears in a gesture phrase window:

$$p(\text{VBZ} \in \text{GP window} \mid \text{Deictic}) = .38 \quad (1)$$

$$p(\text{VBZ} \in \text{GP window} \mid \text{Iconic}) = .52 \quad (2)$$

This feature is more informative if it appears in the stroke window:

$$p(\text{VBZ} \in \text{Stroke window} \mid \text{Deictic}) = .21 \quad (3)$$

$$p(\text{VBZ} \in \text{Stroke window} \mid \text{Iconic}) = .44 \quad (4)$$

Put another way, if the VBZ feature appears during the gesture phrase window of an iconic gesture, it is almost always during the stroke phase. For deictic gestures, it could appear with equal likelihood anywhere throughout the gesture phrase.

5.1.1 *Linguistic Analysis*

Each word was stemmed, using a lexically-based stemmer, and tagged, using a Java implementation of the Brill tagger [12]. Stemming had no appreciable affect on performance. Each word stem was included as a feature. We also tried some coarse word-sense disambiguation by appending the part-of-speech tag to each word, and including each type of usage as an independent feature (e.g., "fish/NN" and "fish/VB") – this decreased performance from 65.9% to 64.2%. POS tags were used as features on their own; without them, performance decreased to 58.6%. Both differences were significant.

For both words and POS tags, n-grams of size 1 to 3 were used. All n-grams were simply thrown into the feature vector together; in the future we may use backoff models to combine the different size n-grams in a more intelligent way. Unigrams alone provided a performance of 55.1%; adding bigrams improved performance to 60.0%; adding trigrams improved performance to 65.9%; adding 4-grams decreased performance to 65.7%, an insignificant change (all other changes were significant).The mean number of words in each gesture phrase window was 5.0 (median = 4, σ = 3.7), and the mean for the stroke window was 2.8 (median = 2, σ = 2.0). Thus it is unsurprising that larger n-grams afforded no improvement. In total, when using unigrams, bigrams, and trigrams, there were 2746 features.

5.2 Classifier Performance

Table 3 compares the performance of various classifiers on this task. For all classifiers except TWCNB, the Weka [19] implementation was used.

HyperPipes is a simple, fast classifier for situations with a large number of attributes (there are 2746 in this case). HyperPipes records the attribute bounds for each category, and then classifies each test instance according to the category that most contains the instance. As shown in the table, HyperPipes significantly outperforms all other classifiers on this task.

TWCNB is a modification of the Naive Bayes classifier designed by Rennie et. al [17] to better suit text-classification problems. It includes a complement-class formulation which is useful when the number of examples is poorly balanced across classes, as is the case here. It also implements term-frequency transformations, addressing the fact that the multinomial distribution is a poor model of text. Our own implementation of this classifier is used in these experiments.

The NaiveBayes, SVM, and C4.5 classifiers are used "as is" from the Weka library; default settings are used for all parameters. While any one of these classifiers might perform substantially better given an optimal choice of parameters, our purpose is to show the range of performance on this task achieved by some commonly-used techniques, rather than to offer a comprehensive comparison of classifiers.

Table 3 compares the performance of each classifier on the gesture classification task. The results were the average of one hundred experiments, each of which involved randomizing the dataset and then performing a stratified ten-fold cross-validation. All classification accuracy differences were significant, except for SVM versus C4.5, where the difference was not significant.

The "always deictic" classifier chooses the "deictic" class every time. All classifiers significantly outperformed this

	Accuracy	σ
HyperPipes	65.9%	1.47
TWCNB	63.5%	1.66
Naive Bayes	58.9%	1.10
C4.5	56.0%	2.17
SVM	55.9%	2.17
Always deictic	48.7%	N/A
Humans: audio-only	45%	2.7
Humans: audio-video	78%	2.8

Table 3: Comparison of classifier performance, averaged over 100 stratified, ten-fold cross-validation experiments

baseline. Another baseline is the performance of human raters who had access to the same information, the audio surrounding the gesture. The performance of human raters in the audio-only condition was actually worse than the "always deictic" baseline. This suggests that while the linguistic context surrounding the gesture clearly does provide cues for classification, human raters were unable to use these cues in any meaningful way when the video was not also present.

As an upper bound, we consider the performance of the human raters who had access to both the audio and video; the majority opinion of these raters forms the ground truth for this experiment. As shown in the table, the average rater agreed with the majority 78% of the time. This appears to be a reasonable upper bound for a multimodal gesture classification system; it seems unlikely that using the text only, we could achieve higher performance than human raters who had access to visual and prosodic information.

5.3 Discussion

Table 4 lists the ten features that were found to be carry the highest information gain. Capital letters indicate part-of-speech tags, which are defined according to the Penn Treebank set. "UH" indicates an interjection, e.g., "um", "ah", "uh"; "VB*" is a verb, with the last character indicating case and tense; "PRP" is a personal pronoun.

The features correlate with gesture categories in a way that accords well with linguistic theory about the role of speech and gesture as part of an integrated communicative system [13]. For deictics, the word "here" is a good predictor, since it is typically accompanied by a gestural reference to a location in space. The class of "other" gestures is primarily composed of beats, which serve the same turn-keeping function as interjections such as "uh." The "VBZ" tag – indicating a verb in the third-person singular – is a good predictor of iconic gestures, as are the more domain-specific cue words, "back" and "push." These words were used by several speakers to describe the motion of the piston, and were typically accompanied by an iconic gesture describing that motion.

6. RELATED WORK

For a more detailed discussion of the gesture classes described in this paper, see [13]; for an analysis specifically geared towards multimodal user interfaces, see [2].

Computational analysis of unconstrained, natural gesture is relatively unexplored territory, but one exception is the research of Quek and Xiong et al. They have applied McNeill's catchment model [14] to completely unconstrained dialogues, extracting discourse structure information from a number of different hand movement cues, such as gestural oscillations [20].

Pattern-recognition approaches to recognizing some of these gesture classes have been reported in a few publications. Kaiser et al. [7] describe a system that recognizes deictic pointing gestures and a set of manipulative gestures: point, push, and twist. Kettebekov and Sharma [9] present a map-control user interface that distinguishes between deixis and "motion" gestures that are a subset of the class of iconic gestures in the taxonomy that we have used. Kettebekov, Yeasin, and Sharma also applied prosodic information to improve gesture segmentation and the recognition of movement phrases and various types of deictic gestures [10].

Perhaps the most closely related research topic is mutual disambiguation [16], which views speech and gesture as co-expressive streams of evidence for the underlying semantics. If the speech modality suggests a given semantic frame with very high probability, then the probabilities on gestures that are appropriate to that frame are increased; the converse is also possible, with gesture disambiguating speech. While most of the work on mutual disambiguation involves pen/speech interfaces [4], it has more recently been applied to free hand gestures as well [7].

Mutual disambiguation relies on having a constrained domain in which the semantics for every utterance can be understood within the context of a formal model of the topic of discourse. Our approach gives up some of the power of mutual disambiguation, in that semantic information may provide tighter constraints on gesture than the linguistic cues that we use. Our approach is more appropriate to situations in which a formal model of the domain is not available.

7. FUTURE WORK

The ultimate goal of this research is multimodal gesture recognition: a combination of linguistic priors of gesture classes with vision-based recognition. Consequently, the most pressing future work is to combine the textual classifier developed here with traditional pattern-recognition techniques. Hopefully this will show that linguistic context does indeed improve classification performance, as it does for humans.

In addition, there are a number of other ways in which both the empirical study and the automatic classifier can be extended.

7.1 Prosodic versus lexical cues

The experiment involving human raters showed that auditory cues significantly improve visual classification of gestures. However, this experiment does not disambiguate the role of prosody versus lexical and higher-order linguistic features. We can remove prosody by transcribing the speech and feeding it to a text-to-speech engine. If the results using this audio and the original video are indistinguishable from the video-audio condition with human speech, then we could conclude that prosody plays no role in gesture classification. Alternatively, we can remove lexical and higher-order linguistic cues by having speakers communicate in a language unknown to the listeners, but with similar prosodic conventions. If the results prove to be indistinguishable from the video-audio condition in which the listener understands the

Feature	Window	Information	$p(w\|$ Deictic)	$p(w\|$ Iconic)	$p(w\|$ Other)
back	phrase	0.088	.013	.17	.04
UH	stroke	0.064	.051	.017	.24
push	stroke	0.058	.013	.12	.04
VBZ	stroke	0.056	.21	.44	.16
back	stroke	0.055	.026	.15	.04
here	phrase	0.054	.23	.051	.24
as	phrase	0.053	.064	.20	.04
uh	stroke	0.051	.039	.017	.20
as	stroke	0.044	.039	.15	.04
PRP-VBP	phrase	0.044	.12	.017	.08

Table 4: The top ten features by information gain

speaker, then lexical and higher-order linguistic cues are irrelevant to gesture classification.

7.2 Domain generality

All of the test and training data in this corpus is drawn from an experiment within a single domain: engineering mechanical devices. Another experiment could help determine whether the language model learned here is general beyond that domain. The absence of obviously domain-specific terms in the set of more informative features described in the previous section is encouraging.

7.3 Recognized speech and gesture boundaries

The current evaluation is performed using transcriptions, rather than automatically recognized speech. Thus, this system does not have to deal with word errors. In the future, we hope to demonstrate that this classifier is still accurate, even when presented with errorful speech. In addition, we would like to segment gestures automatically, possibly with the aid of prosodic cues as in [10].

7.4 Feature fusion

The classification system as implemented uses classes of features varying on several dimensions: gesture phrase window versus stroke window; word versus part of speech tag; n-gram size. Currently, all features are combined into a single vector and sent to a classifier. A more sophisticated approach might be to interpolate between multiple classifiers and use backoff models to combine the different size n-grams.

8. CONCLUSIONS

Natural, communicative gesture is well described by gesture classes that are fundamentally multimodal in nature, pertaining to both the hand motion and the role played by the gesture in the surrounding linguistic context. Humans rely on auditory as well as visual cues to classify gestures; without auditory cues, performance decreases significantly. This suggests that automatic classification of gestures should make use of both hand movement trajectories and linguistic cues. We have developed a gesture classifier that uses only linguistic features and achieves 66% accuracy on a corpus of unconstrained, communicative gestures.

Acknowledgements

We thank Aaron Adler, Christine Alvarado, Sonya Cates, Tracy Hammond, Michael Oltmans, Sharon Oviatt, Metin Sezgin, Vineet Sinha, and Özlem Uzuner for their helpful suggestions about this work. We also thank Jason Rennie and Kai Shih for their help with the TWCNB classifier. This research is supported by the Intel Corporation, the Microsoft iCampus project and the sponsors of MIT Project Oxygen.

9. REFERENCES

[1] Carletta, J. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics 22*, 2 (1996), 249–254.

[2] Cassell, J. A framework for gesture generation and interpretation. *Computer Vision in Human-Machine Interaction*. Cambridge University Press (1998), 191–215.

[3] Cassell, J., Vilhjalmsson, H., and Bickmore, T. Beat: the behavior expression animation toolkit. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*. ACM Press (2001), 477–486.

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

[5] Eisenstein, J., and Davis, R. Natural gesture in descriptive monologues. *UIST'03 Supplemental Proceedings*. ACM Press (2003), 69–70.

[6] Freeman, W. T., and Weissman, C. Television control by hand gestures. *International Workshop on Automatic Face- and Gesture- Recognition*. IEEE Press (1995), M. Bichsel, Ed., 179–183.

[7] Kaiser, E., Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., and Feiner, S. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. *Proceedings of the Fifth International Conference on Multimodal Interfaces*. ACM Press (2003), 12–19.

[8] Kendon, A. *Conducting Interaction*. Cambridge University Press, 1990.

[9] Kettebekov, S., and Sharma, R. Toward natural gesture/speech control of a large display. *Engineering for Human-Computer Interaction (EHCI'01). Lecture Notes in Computer Science*. Springer Verlag (2001).

[10] Kettebekov, S., Yeasin, M., and Sharma, R. Prosody based co-analysis for continuous recognition of coverbal gestures. *Proceedings of the Fourth International Conference on Multimodal Interfaces (ICMI'02)*. IEEE Press (Pittsburgh, USA, 2002), 161–166.

[11] Koons, D. B., Sparrell, C. J., and Thorisson, K. R. Integrating simultaneous input from speech, gaze, and hand gestures. *Intelligent Multimedia Interfaces*. AAAI Press (1993), 257–276.

[12] Liu, H. Montylingua v1.3.1: An end-to-end natural language processor of english for python/java, 2003.

[13] McNeill, D. *Hand and Mind*. The University of Chicago Press, 1992.

[14] McNeill, D., Quek, F., McCullough, K.-E., Duncan, S., Furuyama, N., Bryll, R., Ma, X.-F., and Ansari, R. Catchments, prosody, and discourse. *Gesture 1* (2001), 9–33.

[15] Nakano, Y. I., Okamoto, M., Kawahara, D., Li, Q., and Nishida, T. Converting a text into agent animations: Assigning gestures to a text. *HLT-NAACL 2004: Companion Volume*. ACL Press (2004), 153–156.

[16] Oviatt, S. L. Mutual disambiguation of recognition errors in a multimodel architecture. *Human Factors in Computing Systems (CHI'99)*. ACM Press (1999), 576–583.

[17] Rennie, J. D. M., Shih, L., Teevan, J., and Karger, D. R. Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*. AAAI Press (2003).

[18] Sparrell, C. Coverbal iconic gesture in human-computer interaction. Master's thesis, Massachusetts Institute of Technology, 1993.

[19] Witten, I. H., and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[20] Xiong, Y., Quek, F., and McNeill, D. Hand motion gestural oscillations and multimodal discourse. *Fifth International Conference on Multimodal Interfaces (ICMI'03)*. IEEE Press (2003), 132–139.

Appendix: Instructions for Raters

In this study you will be asked to identify gestures as belonging to one of three classes: deictic, action, or other.

DEICTIC gestures involve pointing at, tracing the outline of, or otherwise indicating a specific object or region of space. For example, a speaker might point at a book and say, "this is the book I read last week." Drag the mouse over the squares below to see short video clips of deictic gestures.

ACTION gestures reenact a physical interaction, trajectory of motion, or some other event. For example, a speaker might describe a bouncing pinball by tracing a path of motion with the hand while saying, "the ball bounces all over the place." Drag the mouse over the square below to see a short video clip of an action gesture.

OTHER gestures include the gesticulation that typically accompanies speech (e.g., creating visual "beats" to emphasize important speaking points) as well as any other gesture that cannot easily be classified in either of the above two categories. Drag the mouse over the square below to see a short video clip of an "other" gesture.

First, you will be presented with a video, and a user-interface window that allows you to play specific clips from the video. Each clip includes a single gesture, which you will be asked to classify using the above framework. You will be presented with four such videos; at times, the audio may be muted, or the video itself may be hidden. Based on whatever information is available, please make your best effort to correctly classify each gesture. Even if you feel that you do not have enough information to correctly classify a gesture, please make your best guess.

Next, you will be presented with a set of textual transcriptions of the speech surrounding each gesture. Based on this text, please make your best effort to correctly classify each gesture.