Beyond Attention: The Role of Deictic Gesture in Intention Recognition in Multimodal Conversational Interfaces

Shaolin Qu **Joyce Y. Chai**
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
{qushaoli, jchai}@cse.msu.edu

ABSTRACT

In a multimodal conversational interface supporting speech and deictic gesture, deictic gestures on the graphical display have been traditionally used to identify user attention, for example, through reference resolution. Since the context of the identified attention can potentially constrain the associated intention, our hypothesis is that deictic gestures can go beyond attention and apply to intention recognition. Driven by this assumption, this paper systematically investigates the role of deictic gestures in intention recognition. We experiment with different model-based methods and instance-based methods to incorporate gestural information for intention recognition. We examine the effects of utilizing gestural information in two different processing stages: speech recognition stage and language understanding stage. Our empirical results have shown that utilizing gestural information improves intention recognition. The performance is further improved when gestures are incorporated in both speech recognition and language understanding stages compared to either stage alone.

Author Keywords

Multimodal interface, language understanding, speech, gesture.

ACM Classification Keywords

H5.2 [Information interfaces and presentation]: User Interfaces - Theory and methods, Natural language.

INTRODUCTION

In multimodal conversational systems, multiple input modalities (e.g., speech, gesture, and eye gaze) are utilized to facilitate more natural and efficient human machine conversation [23]. Many systems with different combinations of modalities haven been developed in the last two decades [7, 13, 17, 27]. In this paper, we focus on a speech-gesture system where user speech inputs are accompanied by deictic gestures on the graphical display. Since human speech is the most natural communication mode, this type of system is easier and more intuitive for users to interact with.

A key component in multimodal conversational systems is semantic interpretation, which is to identify semantic meanings from user input. In conversational systems, the "meaning" from user input can be generally categorized into *intention* and *attention* [11]. Intention indicates the user's motivation and action. Attention reflects focus of the conversation, in other words, what has been talked about. In the speech-gesture system where speech is the dominant mode of communication, the user intention (such as asking for price of an object) is generally expressed by spoken language and attention (e.g., the specific object) is indicated by the deictic gesture on the graphical display.

Based on such observations, many speech-gesture systems apply a semantic fusion approach where speech is used to mainly identify intention and deictic gestures are used to identify attention [1, 18, 9]. It is not clear whether deictic gestures can be used at all in recognizing intention and how to use those gestures. In our view, deictic gestures not only indicate users' attention, but also can activate the relevant context (e.g., domain context and visual context). This context can constrain the type of intention associated to attention and thus provide useful information for intention prediction.

Driven by this assumption, this paper presents an empirical investigation on the role of deictic gestures in intention recognition. We apply model-based methods and instance-based methods to incorporate gestural information to recognize users' intention. We examine the effects of using gestural information for user intention recognition in two stages – speech recognition stage and language understanding stage. Our empirical results have shown that utilizing gestural information improves intention recognition and the performance is further improved when gestures are incorporated in both speech recognition and language understanding stages compared to either stage alone.

In the following sections, we first give an introduction to multimodal input interpretation, then describe in detail intention recognition with gestural information, and finally present results from empirical evaluations.

IUI'08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain.

RELATED WORK

Intention has been addressed in a number of ways to serve different purposes in different research areas. For example, in the field of Robotics, intention has been used to represent an idea or a mental state associated with human actions [32], a sequence of human behaviors or robot actions [16, 20], and a specific robot action [30]. In dialog systems, intention is used generally to refer to discourse purpose [11], which is related to our definition of intention here. More specifically, intention recognition in our work is to identify specific actions from user utterances that constitute a part of conversation discourse. Rather than identifying high level intentions such as various beliefs, doubt, promise, and compliment, we are only dealing with specific intention that is relevant to the type of conversational systems we are developing.

Besides intention, another important type of information from user utterances is attention [11]. In multimoal conversational systems, deictic gestures have been mainly used for attention identification in previous work. Many approaches have been developed to incorporate gestural information to resolve referring expressions (e.g., using gesture information to resolve what *this* refers to in the utterance "*how much does this cost?*") [15, 33, 22, 19, 5, 3]. For example, a salience-based approach was designed for reference resolution in a multimodal system supporting deictic gesture [15]. In this approach, each potential referent is assigned a salience value decided by discourse and visual contextual factors (e.g., whether objects are visible on screen or selected by deictic gestures). To determine the referent of a multimodal referring expression, the system retrieves the most salient referent that satisfies the semantic restrictions of the referring expression. In a multimodal map application [19], a decision list was designed for multimodal reference resolution based on the theory of Givenness Hierarchy. The decision list determines the referents based on whether there are objects being gestured to or objects visible on the screen and whether these objects satisfy semantic restrictions of the referring expressions. Another graph-based approach was designed for reference resolution in a map-based real estate domain [5]. By representing spoken referring expressions and gesture selections as Attribute Relation Graphs (ARGs), the graph-based approach determines referents for the referring expressions by matching the ARGs in a way that achieves maximum semantic and temporal compatibility. Most recently, a greedy algorithm was designed to resolve referents according to semantic, discourse context and temporal constraints of deictic gestures and referring expressions based on cognitive principles [3]. Different from these earlier works, this paper focuses on how to take gesture beyond attention identification to help intention recognition.

Our work in this paper is based on the hypothesis that the context associated with gestured objects constrains intention and thus can be used to help predict intention. Using contextual knowledge to help language understanding has been addressed in previous work [6, 29, 10, 21, 12, 28, 4, 25]. For example, dialog contextual information is combined with the syntactic and acoustic information of the user's utterance to improve speech recognition by re-ranking the n-best speech hypothesis [6]. In a synthetic visual scene description domain [28], visual context, constituted by the visual features of objects such as color and shape, is used to tailor a class-based bigram language model for recognizing users' utterances describing objects in a visual scene. In our previous work, domain context has been incorporated into speech processing under a salience-based framework [4, 25]. In this framework, deictic gestures are used to indicate the salience of domain contextual knowledge to help speech recognition. Extending this earlier work, we investigate utilization of domain context signaled by the user's deictic gesture in understanding the user utterances, specifically, recognizing the user intention.

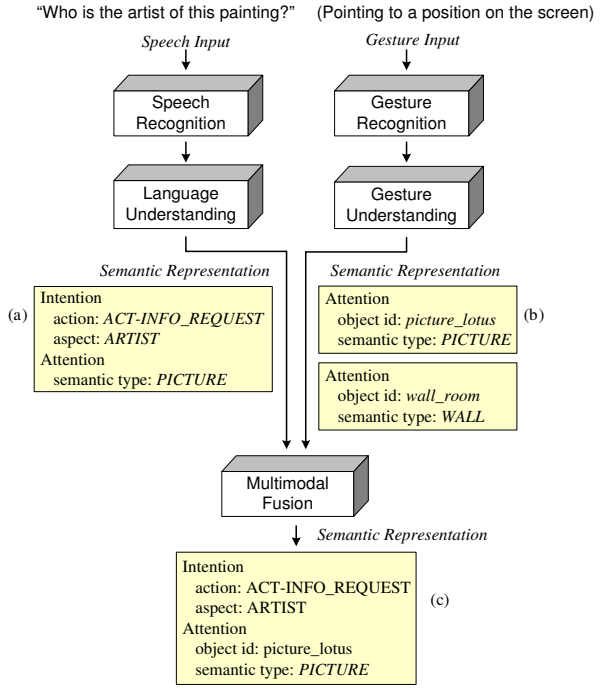


Figure 1. Multimodal interpretation based on semantic fusion

MULTIMODAL INTERPRETATION

Multimodal interpretation is a process of identifying semantic meaning from user inputs. For the system where speech is the main mode of communication and accompanied by deictic gestures, a semantic fusion approach is widely adopted for multimodal interpretation. Figure 1 shows an example of multimodal interpretation by a semantic fusion approach. In the example, the user says "*who is the artist of this painting?*" and at the same time points to a position on the screen. The system first creates partial meaning representations independently from speech and gesture modalities. For speech input, the system first converts acoustic speech signal to word sequence by speech recognition, then, in language understanding, the system infers what the user wants to do from the recognized speech. For instance, in this example, the system identifies that the user intends to request information about some artist (represented in the intention structure in Figure 1-(a)) and the object of interest is some kind of painting (represented in the attention structure). For

gesture input, the system first obtains the location where the deictic gesture takes place by gesture recognition, then identifies what object the gesture points to in gesture understanding. For example, here the deictic gesture would result in two possible objects as shown in Figure 1-(b). The partial meaning representations from speech and gesture input mutually disambiguate each other and the compatible ones are fused together to form the overall semantic representation as shown in Figure 1-(c).

As seen from this example, because of the nature of pointing, deictic gestures can be most conveniently used to identify objects in focus. However, we believe that deictic gestures should also help intention recognition. Therefore, we proposed a new architecture as shown in Figure 2. In this architecture, gesture can be incorporated in two stages to help intention recognition. As illustrated by (a) in Figure 2, gesture can be used together with recognized speech hypotheses in language understanding (LU) stage for intention recognition. Since speech recognition is not perfect, gesture can also be used in speech recognition (SR) stage to improve speech recognition hypotheses and thus improve intention recognition as shown in Figure 2-(b).

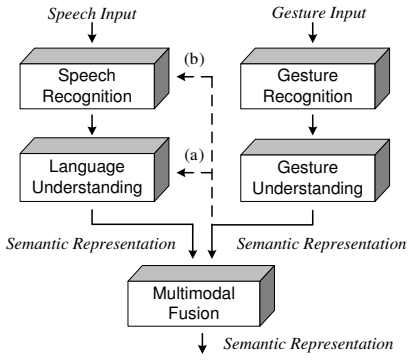


Figure 2. Using Gesture for intention recognition

In the following section, we describe how intention is automatically inferred. In particular, we explain how deictic gesture can be combined with speech to recognize user intention.

INTENTION RECOGNITION

We investigate utilizing gesture for intention recognition in a speech-gesture system that is built for a 3D interior decoration domain as shown in Figure 3. Users can interact with the system using both speech and deictic gestures to query information about the 3D objects or arrange the room by adding, removing, moving, and coloring the objects. For example, the user may say "*remove this lamp*" or ask "*what's the power of this lamp?*" while pointing at a lamp in the scene.

As shown by (c) in Figure 1, semantic meanings of users' multimodal inputs are represented by semantic frames (Figure 1-(c)), which consists of two parts: intention and attention. Intention specifies what the user intends to do to an object, whereas attention indicates which object the

Figure 3. A 3D interior decoration domain

user wants to take action on. Specifically, intention is represented by an action and its corresponding aspect. All actions and corresponding aspects in the interior decoration domain are shown in Table 1. Note that for action *ACT-INFO_REQUEST*, the aspect includes different domain properties such as *ARTIST*, *AGE*, and *PRICE*. Instead of putting the properties in the attention structure as in [2], we move these properties to the intention structure since they can be applied to any object of interest within a certain semantic class.

Action	Aspect
ACT-ADD	<null>
ACT-ALTERNATES_SHOW	<null>
ACT-INFO_REQUEST	<domain property> or <null>
ACT-MOVE	<location> or <null>
ACT-PAINT	<color> or <null>
ACT-REMOVE	<null>
ACT-REPLACE	<replacement> or <null>
ACT-ROTATE	<direction> or <null>

Table 1. Intentions in the 3D interior decoration domain

Given this representation, intention recognition can be formulated as a classification problem. Each action-aspect pair can be considered as a particular type of intention. For action *ACT-INFO_REQUEST*, there are 11 possible aspect values, which result in 11 classes. For all other 7 actions, each action is treated as one type of intention despite multiple possible aspect values. During interpretation, additional postprocessing will take place to identify different aspects. For example, for action *ACT-PAINT*, the system will try to identify the <color> value (e.g., *red*, *blue*) from the user's utterance after *ACT-PAINT* is predicted as the user's intended action. In this paper, we only focus on the classification of intention without elaborating on the postprocessing. In total, there are 19 target classes for intention recognition (including class *NOT-UNDERSTOOD* to represent intention not supported in the domain).

Using Gesture in Language Understanding for Intention Recognition

To examine the role of deictic gestures in intention prediction, we apply different approaches to incorporate gesture with recognized speech hypotheses during language understanding stage. Next, we describe how we extract semantic features from users spoken utterances and gestures for predicting user intention.

Semantic Features

The semantic features of users' multimodal input consist of two parts: lexical features extracted from users' spoken utterances, and domain contextual features extracted from users' deictic gestures.

- Lexical Features
  Lexical feature is represented by a binary feature vector that indicates what semantic concepts appear in the user's utterance. The semantic concepts are extracted from the recognized speech hypotheses (could be n-best hypotheses or 1-best hypothesis) based on lexical rules. Currently, we have 18 semantic concepts in the interior decoration domain with 130 lexical rules.
- Domain Contextual Features
  When a deictic gesture takes place, the selected object and its properties as defined in the domain are activated, which forms the domain context of the user's utterance. This context constrains what the user is likely talking about. For example, the user is unlikely to ask the artist of a lamp or the wattage of a picture. Therefore, this domain context can be used to help predict user intention. For each gesture that accompanies the user's utterance, we choose the most likely object selected by the gesture and use the semantic type of the object as the contextual feature. There are 14 semantic types of objects in the domain.

Model-Based Intention Prediction

Given an instance $\mathbf{x}$ that is represented by semantic features, we applied three classifiers to predict user intention.

- Naive Bayes
  The prediction c^* of instance $\mathbf{x}$ is given by
  $$c^* = \arg\max_c p(c|\mathbf{x}) = \arg\max_c p(c|x_1, x_2, \ldots, x_m) \quad (1)$$
  where x_i is the i-th feature of instance $\mathbf{x}$.

  Applying Bayes' theorem and assuming the features are conditionally independent given a class, we have
  $$\begin{aligned} p(c|\mathbf{x}) &= \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} = \frac{p(x_1, x_2, \ldots, x_m|c)p(c)}{p(\mathbf{x})} \\ &= \frac{p(c)\prod_{i=1}^{m} p(x_i|c)}{p(\mathbf{x})} \\ &\propto p(c)\prod_{i=1}^{m} p(x_i|c) \end{aligned} \quad (2)$$
  Estimating $p(c)$ and $p(x_i|c)$ from the training data, we can get the prediction of a testing instance by Equation (1). In our evaluation, add-one smoothing was used in the estimation of $p(c)$ and $p(x_i|c)$ for predicting user intention.
- Decision Tree
  In a decision tree, each root node provides the classification of the instances, each non-leaf node specifies a test of some attribute of the instances, and each branch descending from that node corresponds to one of the possible values for this attribute. Decision trees classify instances by sorting them down the tree from the root node to some leaf node through a list of attribute tests. We used C4.5 algorithm [26] to construct decision trees for intention prediction based on the semantic features of users' multimodal input.
- Support Vector Machines (SVM)
  The SVM [8] is built by mapping instances to a high dimensional space and finding a hyperplane with the largest margin that separates the training instances into two classes in the mapped space. In prediction, an instance is classified depending the side of the hyperplane it lies in. A kernel function κ is used in SVM to achieve linear classification in the high dimensional space. Based on the semantic features of users' multimodal input, we used a polynomial kernel for user intention prediction.

  Since SVM can only handle binary (2-class) classification, a "one-against-one" method is applied to use SVM for multi-class classification [14]. For a classification task of c classes, $c(c-1)/2$ SVMs are built for all pairs of classes and each SVM is trained on the data from the pair of two classes. In the testing phase, a test instance $\mathbf{x}$ is classified through a majority voting strategy. For each of the $c(c-1)/2$ binary classifiers built for class pair (c_i, c_j), if the classifier decides $\mathbf{x}$ belongs to the class c_i, the vote for class c_i increases by one. Otherwise, the vote for class c_j increases by one. After all binary classifiers have been used to vote for the classes, the one that wins the most votes is picked as the prediction of $\mathbf{x}$.

Instance-Based Intention Prediction

We also applied k-nearest neighbor (KNN), an instance-based method, to predict users' intention. Given a set of training instances with known intention, the KNN method (k=1) predicts the intention of a testing instance by finding the testing instance's closest match in the training instances and using the match's intention as the prediction.

In instance-based intention prediction, besides semantic features, we also use phoneme features of users' spoken utterances for intention prediction. For each speech recognition hypothesis of an utterance, we can get a phoneme sequence. Each phoneme sequence is treated as a phoneme feature.

We give an example to show the potential of using phoneme features to help user intention prediction. As shown in Figure 4, the user's utterance is not correctly recognized and as a result, the semantic feature extracted from the recognized speech does not give any useful information about the user's intention of *ACT-INFO_REQUEST*. Therefore, using semantic features alone will fail to predict the user's intention. However, if we compare the two phoneme sequences of the true utterance and the speech recognition result, we can find that the phoneme sequences of the mis-recognized speech, [ax n d] [**f er**] [**m** ih **sh ax n**], is close to the true phoneme sequence [ih n **f er m** ey **sh ax n**]. This means that using phoneme sequence similarity can help recover the word "information", which is the key to identifying the user's intention in this utterance, and therefore can help predict intention of the user.

User utterance: "*information on this*"
Phonemes: [ih n f er m ey sh ax n] [ao n] [dh ih s]

Speech recognition: "*and for mission on this*"
Phonemes: [ax n d] [f er] [m ih sh ax n] [ao n] [dh ih s]

Figure 4. Phonemes of an utterance

We applied KNN to predict user intention based on semantic features and phoneme features. The similarity between a testing instance $\mathbf{x}^t$ and a training instance $\mathbf{x}^r$ is defined as

$$d_{sp}(\mathbf{x}^t, \mathbf{x}^r) = d_s(\mathbf{x}^t, \mathbf{x}^r) + d_p(\mathbf{x}^t, \mathbf{x}^r) \quad (3)$$

where $d_s(\mathbf{x}^t, \mathbf{x}^r)$ is the Hamming distance between the nominal semantic features and $d_p(\mathbf{x}^t, \mathbf{x}^r)$ is the distance between the phoneme features.

Hamming distance $d_s(\mathbf{x}^t, \mathbf{x}^r)$ is defined as:

$$d_s(\mathbf{x}^t, \mathbf{x}^r) = \sum_{k=1}^{m} \delta(x_k^t, x_k^r) \quad (4)$$

where $x_k^t(x_k^r)$ is the k-th attribute in the semantic feature, and

$$\delta(x_k^t, x_k^r) = \begin{cases} 0 & x_k^t = x_k^r \\ 1 & x_k^t \neq x_k^r \end{cases}$$

Phonemes distance $d_p(\mathbf{x}^t, \mathbf{x}^r)$ is defined as follows based on different configurations:

- when n-best speech recognition is used, and no gestural information is used:

$$d_p(\mathbf{x}^t, \mathbf{x}^r) = \min_k MED(P_k^t, P^r) \quad (5)$$

- when n-best recognized speech hypotheses [1] are used, and gestural information (i.e., objects indicated by deictic gestures) is used

$$d_p(\mathbf{x}^t, \mathbf{x}^r) = \min_k \Big(MED(P_k^t, P^r) \Big) + w_e(o_t, o_r) \quad (6)$$

where

MED	–	minimum edit distance
P_k^t	–	phonemes of the k-th speech recognition hypothesis of testing instance $\mathbf{x}^t$
P^r	–	phonemes of the speech transcript of training instance $\mathbf{x}^r$
$w_e(o_t, o_r)$	–	distance between the object o_t selected by gesture accompanying testing instance $\mathbf{x}^t$ and the object o_r selected by gesture accompanying training instance $\mathbf{x}^r$ (0 if o_t and o_r are of the same semantic type, otherwise a non-zero constant)

[1] When 1-best hypothesis is used, no *min* operation is necessary.

Utterance: "what is the power of this lamp?"

Standard speech recognition:
n-best list:
is the artist lamp
is the artist left

Gesture-tailored speech recognition:
gesture selection:
p(lamp_mr) = 0.8094
p(door_1) = 0.1363
p(table_pc) = 0.052
p(bedroom) = 0.0023
n-best list:
is the power this lamp
is the power this lamp's
is the power this left

Figure 5. N-best lists of speech recognition of an utterance

Using Gesture in Speech Recognition for Intention Recognition

As mentioned earlier, speech recognition is important for intention recognition. The better the speech recognition, the more accurate recognition hypotheses can be derived, which potentially leads to better intention recognition. In our previous work [4, 25], we have shown that incorporating gesture in speech recognition can achieve better recognition. We can use gestural information first to help recognize users' speech, then based on the gesture tailored speech recognition, to identify intention by classification-based language understanding as described in previous section.

More specifically, gesture is incorporated in speech recognition through gesture-based salience driven language modeling [25]:

$$p_s(w_i|w_{i-1}) = \frac{p(w_i|w_{i-1}) + \lambda \sum_e p(w_i|w_{i-1}, e)p(e)}{1+\lambda} \quad (7)$$

where $p(w_i|w_{i-1})$ is the standard bigram probability, $p(e)$ is the salience distribution over objects on the graphical display, which is influenced by the deictic gestures. The priming weight λ decides how much the original bigram probability will be tailored by the salient objects selected by the gestures. This equation gives the new bigram probability $p_s(w_i|w_{i-1})$ that is tailored by the deictic gestures. This new bigram model can be used in the speech decoding process (i.e., Viterbi search) to generate the gesture tailored speech recognition.

Figure 5 shows an example where using gesture improves speech recognition and thus helps intention recognition. In the example, the user asks "*what is the power of this lamp?*" while pointing to a lamp object on the screen. The standard speech recognition results are shown in the figure. As we can see, none of the n-best hypotheses preserve the important information about the utterance. However, the gesture can be incorporated to improve recognition. In this case, the pointing gesture results in a salience distribution of en-

tities in the graphic display as shown in the figure. When this salience distribution is integrated with speech recognition, the tailored n-best hypotheses preserves the important term "power". The correct recognition of word "power" can greatly help intention recognition because word "power" indicates the key semantic concept *WATTAGE*, which is critical to identify the user's intention.

EVALUATION

We empirically evaluated the role of gestural information in intention recognition. We applied both model-based and instance-based methods, and investigated utilization of gesture for intention recognition in language understanding and speech recognition stages.

Experiment Settings

We used the CMU Sphinx-4 speech recognizer [31] for speech recognition. An open acoustic model and a domain dictionary were used in recognizing users' spoken utterances.

For model-based intention prediction, we evaluated the intention prediction accuracies with the following classifiers based on semantic features:

- *NBayes* – naive bayes
- *DTree* – decision tree (C4.5)
- *SVM* – support vector machine (polynomial kernel)

For instance-based intention prediction, we evaluated the intention prediction accuracies with KNN classifiers based on different instance similarity functions:

- *S-KNN* – instance distance defined on semantic features (Equation (4))
- *P-KNN* – instance distance defined on phoneme features (Equations (5) and (6) depending on whether gestural information is incorporated)
- *SP-KNN* – instance distance defined on combinational features of semantics and phonemes (Equation (3))

For each method, we compared the performances of using only the 1-best speech recognition hypothesis and using all n-best speech recognition hypotheses for intention prediction. Also, to compare the influences of gestural information on intention prediction, we evaluated intention prediction under three gesture configurations:

- *noGest* – no gestural information is used
- *recoGest* – with gesture recognition results, i.e., the most likely objects selected by the user's gestures as recognized by the system.
- *trueGest* – with ground truth gesture recognition results, i.e., the objects truly selected by the user's gestures

For each method, we further evaluated intention prediction based on standard speech recognition and gesture-tailored speech recognition. When intention prediction is based on standard speech recognition, gestural information is incorporated only in language understanding for intention prediction. When intention prediction is based on gesture-tailored speech recognition, gestural information is already used in speech recognition and can also be used in language understanding stage for intention prediction.

The evaluations were done by a 10-fold cross validation on the 649 utterances with accompanying gestures that were collected in our user studies.

Results Based on Traditional Speech Recognition

Table 2 shows the intention prediction accuracies based on the standard speech recognition results that did not use gestural information. The intention prediction accuracies based on transcripts of users' spoken utterances are also given in the table to show the upper-bound performance when speech is perfectly recognized.

For all model-based methods (i.e., NBayes, DTree, SVM), the results show that using gestural information together with recognized speech (1-best or n-best) in intention prediction achieves statistically significant improvement on prediction accuracy compared to not using gestural information. Among instance-based methods (i.e., S-KNN, P-KNN, SP-KNN), only for the S-KNN that uses semantic features, intention prediction accuracies are improved significantly when gestural information is used together with recognized speech (1-best or n-best hypotheses). For the P-KNN, where only phoneme features are used, there is no significant change between the intention prediction using gesture and not using gesture, no matter if gestural information is used together with 1-best speech recognition or n-best speech recognition. For the SP-KNN that uses both semantic and phoneme features, intention prediction is significantly improved only when gestural information is used together with 1-best speech recognition.

It is found that, used together with recognized speech hypotheses in model-based methods, ground truth gesture selection achieves more accurate intention prediction than recognized gesture selection in most configurations. This indicates that improving gesture recognition and understanding can further enhance intention prediction when speech recognition is not perfect. When SVM is applied on semantic features extracted from all n-best speech recognition hypotheses, using the true gesture selection achieves slightly worse performance than using the recognized gesture selection. However, t-test shows that this difference is not significant. In instance-based methods, using true gesture selection makes no significant difference compared to using recognized gesture selection for user intention prediction.

Results Based on Gesture-Tailored Speech Recognition

Table 3 shows the intention prediction accuracies based on the gesture-tailored speech recognition hypotheses. Note that in Table 3, gestural information (all possible gesture selections recognized by the system) has been utilized in speech recognition [25], the configurations *noGest*, *reco-*

	transcript			n-best hypotheses			1-best hypothesis		
	noGest	recoGest	trueGest	noGest	recoGest	trueGest	noGest	recoGest	trueGest
NBayes	0.860	0.878	0.874	0.709	0.741	0.755	0.721	0.747	0.763
DTree	0.881	0.888	0.889	0.718	0.729	0.738	0.727	0.755	0.769
SVM	0.878	0.884	0.884	0.713	0.749	0.744	0.730	0.747	0.760
S-KNN	0.881	0.888	0.884	0.700	0.740	0.737	0.730	0.757	0.758
P-KNN	0.918	0.921	0.921	0.790	0.797	0.806	0.798	0.801	0.804
SP-KNN	0.937	0.934	0.934	0.824	0.826	0.832	0.820	0.834	0.844

Table 2. Accuracies of intention prediction based on standard speech recognition

	transcript			n-best hypotheses			1-best hypothesis		
	noGest	recoGest	trueGest	noGest	recoGest	trueGest	noGest	recoGest	trueGest
NBayes	0.860	0.878	0.874	0.727	0.753	0.766	0.735	0.764	0.783
DTree	0.881	0.888	0.889	0.749	0.766	0.781	0.743	0.772	0.795
SVM	0.878	0.884	0.884	0.750	0.780	0.786	0.752	0.764	0.777
S-KNN	0.881	0.888	0.884	0.753	0.770	0.781	0.758	0.778	0.795
P-KNN	0.918	0.921	0.921	0.826	0.829	0.827	0.812	0.815	0.817
SP-KNN	0.937	0.934	0.934	0.858	0.857	0.860	0.843	0.855	0.860

Table 3. Accuracies of intention prediction based on gesture-tailored speech recognition

Gest, and *trueGest* only apply to how gestural information is used in the language understanding stage for intention prediction. Therefore, in Table 3, the results under configurations *n-best hypotheses + noGest* and *1-best hypothesis + noGest* are actually the intention prediction performance when gestural information is used in only speech recognition stage.

Compared to using gestural information only in speech recognition, the accuracies of intention prediction are significantly improved in all model-based methods when gestural information is used in both speech recognition and language understanding, no matter if it is used together with 1-best or n-best speech recognition. Among instance-based methods, only in S-KNN, that using gestural information in both speech recognition and language understanding (with 1-best or n-best recognition hypotheses) significantly improves intention prediction compared to using gestural information only in speech recognition. For P-KNN, whether or not gestural information id used in language understanding does not make significant change on intention prediction. For SP-KNN, it is only when gestural information is used together with 1-best speech recognition hypothesis in language understanding that intention prediction is significantly improved compared to using gestural information only in speech recognition.

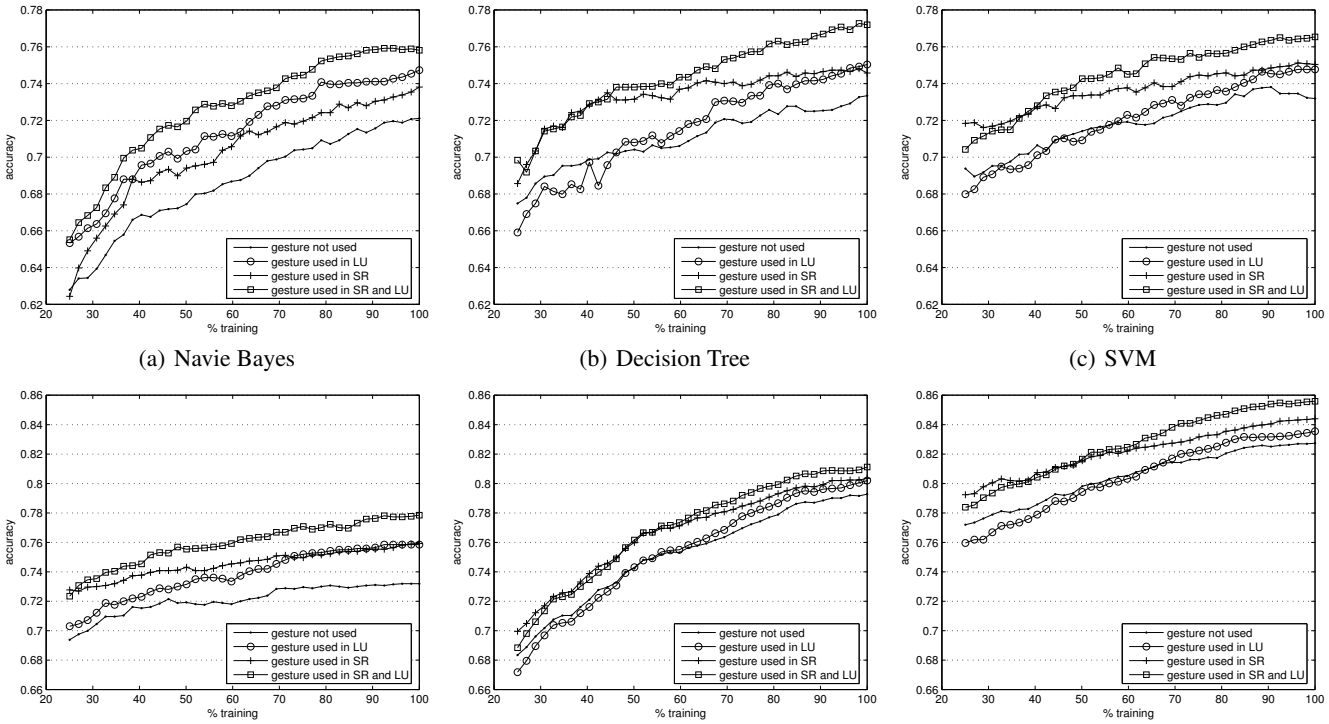
In all model-based methods, together with recognized speech, using ground truth gesture selection in language understanding is found to improve intention prediction more than the recognized gesture selection. Again, this indicates that improving gesture recognition and understanding is helpful for intention prediction. In instance-based methods, using true or recognized gesture selection in language understanding stage for intention prediction does not make significant differences when phoneme features are used.

Results Based on Different Sizes of Training Data

The empirical results have shown that using gestural information improves user intention recognition. To examine whether this improvement by using gestural information is dependent on the size of training data, we compare the accuracies of intention prediction with different sizes of training sets. The results from each method are shown in Figure 6. The semantic features and phoneme features are extracted from the 1-best speech recognition and the recognized gesture selection are used in intention prediction.

The intention prediction accuracy curves are generated in the following way. The whole data set is first separated into 5 folds in a stratified way such that the class distributions in each fold are the same. In each round of evaluation, two different folds are picked as the testing set and initial training set, instances in the other 3 folds are added to the training set incrementally by random picking to get intention prediction accuracies based on different sizes of training sets. After each fold of data has been used as testing set and initial training set, the intention prediction accuracy curves of the 20 round evaluations are averaged to get the curves in Figure 6.

We can see that, for all model-based and instance-based methods, using gestural information in both speech recognition stage and language understanding stage always outperforms using gestural information in only language understanding stage or not using gestural information at all for intention prediction. Using gestural information only in speech recognition stage is found to always outperform not using gestural information for intention prediction in all model-based and instance-based methods despite the training size. When gestural information is used only in language understanding stage, Naive Bayes and S-KNN always improve intention prediction despite the training size. For the other methods (Decision Tree, SVM, P-KNN, and SP-KNN), sufficient training data is needed to make gestural information helpful for intention prediction.



(a) Navie Bayes (b) Decision Tree (c) SVM

(d) S-KNN (e) P-KNN (f) SP-KNN

Figure 6. Intention prediction performance based on different training size

DISCUSSION

The empirical results lead to several findings about the role of deictic gestures in intention recognition.

First, *deictic gesture helps intention recognition given the current speech recognition technology. The earlier deictic gesture is used in the processing stage, the more effect it brings to intention recognition.* Speech recognition technology still has significant limitations. Consistent with previous findings [24], gestures can help language processing through mutual disambiguation, however, in a much earlier stage. Figure 7 shows the role of gesture in intention recognition by different methods at different stages: used in speech recognition, used in language understanding, used in both speech recognition and language understanding. Across all methods, we can easily see that using gestural information in speech recognition stage or language understanding stage improves intention prediction. Using gestural information in both speech recognition stage and language understanding stage further improves intention prediction. Therefore, it is desirable to incorporate gesture earlier in the pipeline (i.e., for speech recognition).

Second, *deictic gesture does not help much in intention recognition for a simple/small domain if speech is perfectly recognized.* As we can see in Table 2, when gestural information is used together with the transcripts of user utterances to predict intention, the effect is not as significant as when gesture information is used with recognized speech hypotheses. This is within our expectation. Given a simple domain with the limited number of vocabularies (the vocabulary size for our current domain is 250), it is relatively easier to come up with sufficient semantic grammar to cover the variations of language. In other words, once user utterances are correctly recognized, the semantics of the input can most likely be correctly identified by the language understanding component. So the bottleneck in interpretation appears in speech recognition (due to many possible reasons such as background noise, accent, etc.) The better speech recognition is, the better the language understanding component processes the hypotheses, and the less effect the gesture is likely to bring. When speech is perfectly recognized (i.e., same as transcriptions), the addition of gesture information will not bring extra advantage. In fact, it may hurt the performance if gesture recognition is not adequate. However, we feel that when the domain becomes more complex and the variations of language become more difficult to process, the use of gesture may begin to show advantage even when speech recognition performs reasonably well. Certainly this hypothesis is yet to be validated in our future work. After all, speech recognition is far from being perfect in reality, which makes gestural information valuable in intention recognition.

Third, *deictic gesture helps more significantly when combined with semantic features than with phoneme features for intention prediction.* As shown in Figure 7, for NBaeys, DTree, SVM and S-KNN where only semantic features are used, the addition of deictic gesture in both speech recognition and language understanding can improve the performance between 4.7% and 6.6%. For P-KNN where only the phonemes features are used, the improvement is 2.1%.

Although the addition of phoneme features significantly improves the intention recognition performance, it is computationally much more expensive than the use of only semantic features. Using phoneme features may become impractical in real-time systems for complex domains. Thus the incorporation of the gestural information could be even more important.

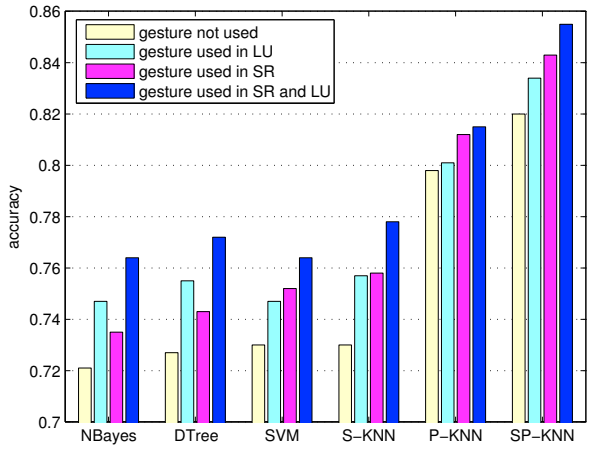


Figure 7. Results of incorporating gestural information in different stages for intention recognition

CONCLUSIONS

This paper presents an empirical investigation on the role of deictic gesture in recognizing user intention during interaction with a speech and gesture interface. Different model-based methods and instance-based methods utilizing gestural information have been applied to recognize users' intention. Our empirical results have shown that using gestural information in either speech recognition or language understanding stage is able to improve user intention recognition. Moreover, when gestural information is used in both speech recognition and language understanding, intention recognition can be further improved. These results indicates that deictic gestures, although most indicative to reflect user attention, are helpful in recognizing user intention. These results further point out when and how deictic gesture should be effectively incorporated in building practical speech-gesture systems.

ACKNOWLEDGMENTS

This work was supported by a Career Award IIS-0347548 and IIS-0535112 from the National Science Foundation. The authors would like to thank anonymous reviewers for their valuable comments and suggestions.

REFERENCES

1. R. A. Bolt. Put that there: Voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262–270, 1980.
2. J. Chai, S. Pan, and M. Zhou. Mind: A context-based multimodal interpretation framework in conversational systems. In O. Bernsen, L. Dybkjaer, and J. van Kuppevelt, editors, *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems*. Kluwer Academic Publishers, 2005.
3. J. Chai, Z. Prasov, and S. Qu. Cognitive principles in robutst multimodal interpretation. *Journal of Artificial Intelligence Research*, 27:55–83, 2006.
4. J. Chai and S. Qu. A salience driven approach to robust input interpretation in multimodal conversational systems. In *Proceedings of HLT/EMNLP'05*, 2005.
5. J. Y. Chai, P. Hong, and M. X. Zhou. A probabilistic approach to reference resolution in multimodal user interfaces. In *Proceedings of IUI'04*, pages 70–77, 2004.
6. A. Chotimongkol and A. Rudnicky. N-best speech hypotheses reordering using linear regression. In *Proceedings of 7th EUROSPEECH*, pages 1829–1832, 2001.
7. P. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: multimodal interaction for distributed applications. In *Proceedings of the Fifth ACM International Conference on Multimedia*, pages 31–40, 1997.
8. C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*, 20:273C297, 1995.
9. J. Eisenstein and C. M. Christoudias. A salience-based approach to gesture-speech alignment. In *Proceedings of HLT/NAACL'04*, 2004.
10. M. Gabsdil and O. Lemon. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proceedings of ACL*, 2004.
11. B. J. Grosz and C. Sidner. Attention, intention, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.
12. A. Gruenstein, C. Wang, and S. Seneff. Context-sensitive statistical language modeling. In *Proceedings of Eurospeech'05*, 2005.
13. J. Gustafson, L. Bell, J. Beskow, B. J., R. Carlson, J. Edlund, B. Granstrom, H. D., and M. Wiren. Adapt - a multimodal conversational dialogue system in an apartment domain. In *Proceedings of 6th International Conference on Spoken Language Processing (ICSLP)*, 2000.
14. C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.
15. C. Huls, E. Bos, and W. Classen. Automatic referent resolution of deictic and anaphoric expressions. *Computational Linguistics*, 21(1):59–79, 1995.
16. S. Iba, C. Paredis, and P. Khosla. Intention aware interactive multi-modal robot programming. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003.

17. M. Johnston, S. Bangalore, G. Vasireddy, A. Stent, P. Ehlen, M. Walker, S. Whittaker, and P. Maloor. Match: An architecture for multimodal dialogue systems. In *Proceedings of the ACL'02*, pages 376–383, 2002.

18. Z. Kazi, S. Chen, M. Beitler, D. Chester, and R. Foulds. Multimodal hci for robot control: Towards an intelligent robotic assistant f or people with disablities. In *Proceedings of AAAI'96 Fall Symposium on Developing AI Applications for the Disabled*, 1996.

19. A. Kehler. Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of AAAI'00*, pages 685–689, 2000.

20. P. Kiefer and C. Schlieder. Exploring context-sensitivity in spatial intention recognition. In *Proceedings of the Workshop on Behaviour Monitoring and Interpretation (BMI'07)*, 2007.

21. O. Lemon and A. Gruenstein. Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments. *ACM Transactions on Computer-Human Interaction*, 11(3):241–267, 2004.

22. J. G. Neal, C. Y. Thielman, Z. H. Dobes, S. M., and S. C. Shapiro. Natural language with integrated deictic and graphic gestures. In M. Maybury and W. Wahlster, editors, *Intelligent User Interfaces*, pages 38–51. CA: Morgan Kaufmann Press, 1998.

23. S. Oviatt. Mulitmodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12:93–129, 1997.

24. S. Oviatt. Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of Conference on Human Factors in Computing Systems: CHI'99*, 1999.

25. S. Qu and J. Chai. Salience modeling based on non-verbal modalities for spoken language understanding. In *Proceedings of ICMI'06*, 2006.

26. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

27. P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of CHI'05*, 2005.

28. D. Roy and N. Mukherjee. Towards situated speech understanding: Visual context priming of language models. *Computer Speech and Language*, 19(2):227–248, 2005.

29. R. A. Solsona, E. Fosler-Lussier, H.-K. J. Kuo, A. Potamianos, and I. Zitouni. Adaptive language models for spoken dialogue systems. In *Proceedings of ICASSP*, 2002.

30. K. A. Tahboub. Intelligent humancmachine interaction based on dynamic bayesian networks probabilistic intention recognition. *Journal of Intelligent and Robotic Systems*, 45:31–52, 2006.

31. W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel. Sphinx-4: A flexible open source framework for speech recognition. Technical Report TR-2004-139, Sun Microsystems Laboratories, 2004.

32. S.-J. Youn and K.-W. Oh. Intention recognition using a graph representation. *International Journal of Applied Science, Engineering and Techcnology*, 4:13–18, 2007.

33. M. Zancanaro, O. Stock, and C. Strapparava. Multimodal interaction for information access: Exploiting cohesion. *Computational Intelligence*, 13(7):439–464, 1997.