Multimodal Interaction System for a Household Assistive Robot

Zhi Li and Ray Jarvis
Intelligent Robotics Research Center
Monash University
Clayton, Victoria 3800, AUSTRALIA
{Zhi.Li, Ray.Jarvis}@monash.edu

Abstract

It is essential for a household assistive robot to understand multiple communication modalities during the interaction with humans. In this paper, we describe our multimodal interaction system which integrates the input information from speech, gesture and gaze. A probabilistic method fuses the results from each individual recognizer at the decision level and then the system decides whether the accumulated information is complete according to a task-orientated command table. Under a two-way dialogue paradigm, the robot can also request more information from the user or confirm its interpretation. The experiments proved that the participants preferred to communicate multimodally, and the multimodal interaction system significantly outperformed unimodal speech.

1 Introduction

For a humanoid robot to participate in our daily life and to interact with humans in a natural style which is similar to our interpersonal communication, it is essential that the robot is able to understand humans' natural communication modes, including speech, gestures, gaze, facial expression and so on. A multimodal interaction system can provide the flexibility that allows users to select and mix among various input modes. The flexibility of a multimodal interface can accommodate a wide range of users and environments, for example, handicapped users, noisy environments and other cases which cannot be recognized by unimodal input.

It is natural, effective and thus overwhelmingly preferred for humans to convey their intentions with multiple means. For example, it is visually vivid and commonly used that the verbal utterance "I like that one" is accompanied with a pointing gesturing indicating the desired object. It is impossible for the robot to fully understand the user's intention if it does not recognize the speech and the gesture at the same time. Linguistic analysis revealed that the spoken and gesture modes consistently provide complementary information, rather than redundant [Oviatt et al., 1997].

Many multimodal interface architectures [Harte and Jarvis, 2009] [Johnston et al., 1997] [Kaur et al., 2003] are designed to cope with ambiguity. Disambiguation of error-prone modalities is an important motivation for the use of multiple modalities. Spoken utterances can be ambiguous -- even a correctly recognized sentence may lead to several hypotheses [Heidemann et al., 2004]. Likewise, gestures are also uncertain. A gesture can have multiple interpretations (e.g. a pointing line may have intersections with more than one object).

Multimodal interfaces have already been proven to significantly prevent errors and improve the effectiveness of the communication. Task-critical errors and disfluent language are reported to drop by 36-50% during multimodal interaction [Oviatt, 1997]. Using multimodal pen/voice interaction, a temporal speed up of 10% was also noticed compared to an unimodal speech input [Oviatt, 1997].

Human-robot interaction often operates in a two-way dialogue paradigm [Holzapfel, 2008]. The robot needs to request more information from the user if the recognized input information is ambiguous and no action should be taken based on the given input. In other situations, for example, when the robot thinks the task to be implemented is high cost or risky, it may need to confirm its understanding with the user before the execution.

The remainder of this paper is organized as follows: we begin by a brief review of the related work in section 2. Then in section 3, we introduce our probabilistic multimodal interaction architecture in detail. Section 4 presents some experiments and the results. Finally, we draw our conclusion in section 5.

2 Related Work

Since Bolt demonstrated his seminal "Put-That-There" system [Bolt, 1980], which processed language commands together with deictic hand gestures, a number of multimodal systems have emerged, such as Virtual World [Codella et al., 1992], CUBRICON[Neal and Shapiro, 1991] and QuickSet [Johnston, 1998]. Early investigations integrated speech with mouse or pen pointing or drawing [Johnston, 1998]; while more sophisticated systems bind language commands with natural gestures [Harte and Jarvis, 2009] [Stiefelhagen and Fugen, 2004], facial expressions [Gunes and Piccardi, 2006] and/or eye gaze [Zhang et al., 2004] [Kaur et al., 2003].

Johnston et al. [Johnston et al., 1997] utilized a unification method [Moshier, 1988] over typed feature

structures [Carpenter, 1992]. A unification method determines the consistency of several pieces of information. They can be combined into a single result if they are consistent. The multimodal integrator agent determines and ranks all possible unifications of speech and gesture and issues complete commands as the output. In this way, the ambiguity of the gesture was resolved by speech and similarly gestures also compensated for the errors in speech. The method was implemented in "QuickSet", which is a distributed interactive simulation system with a multimodal interface (i.e. pen and voice).

A constraint-based multimodal fusion system for speech and pointing gestures was proposed in [Holzapfel et al., 2004]. They also employed a typed feature structure method on the semantic level. It extended Johnston's work and provided a fusion that considers false detection. The pointing gesture recognition subsystem returned a sorted list of objects based on their relative distance to the pointing direction. The speech recognition subsystem mainly specified the action command and the type of the desired object. The multimodal fusion component synchronized and combined the output of the subsystems and sent the result to a dialogue manager. It took the speech as the main modality and used the gesture for disambiguating purposes. They also considered the difference between the arrival time from the recognizer and the actual occurrence time of the event. Events remained in a pool if they were not compatible with fusion rules, and were faded out after a predefined time window.

In [Eisenstein and Christoudias, 2004] the salience of candidate gesture-speech bindings was calculated using a hybrid of data-driven and knowledge-based methods. A penalty function was built, which considered the factors including the time gap, the order of each gestural and spoken symbol, and affinities between different words and gestures. A naïve gradient decent algorithm was used to find the binding with the minimal penalty score.

A semantic network was established in [Russ et al., 2005], in which the nodes were activated depending on their relationship to the given commands. The fading mechanism decreased the activation values of the nodes in the network according to a fading function. Also the interpretation history was considered in their semantic network.

Harte and Jarvis [Harte and Jarvis, 2009] presented a robotic system that fused speech, vision and laser-depth data to perform tasks in a domestic environment. Contextual information, which included the history of recently uttered phrases and the objects' attributes, was taken into consideration. A probabilistic method was used to cope with the possible incorrect recognition.

[Kaiser et al., 2003] proposed a multimodal interaction system in augmented and virtual reality. It fused symbolic and statistical information from 3D gestures, spoken language and referential agents. They used four 6-DOF magnetic sensors attached to the user's hands, arms and head to achieve accurate detection of the hand gestures, pointing direction and the gaze direction. The Unification method [Johnston, 1998] with typed feature structure was then employed to fuse the complementary logical variables.

Some researchers have viewed speech as a primary input mode being self-sufficient and considered gestures, gaze direction and other input as secondary modes just providing redundant accompaniments that carry non-significant information [Oviatt, 1999]. [Holzapfel et al., 2004] took speech as the main modality and used the gestures to disambiguate speech input.

However, the speech signal can be degraded (e.g. in a noisy environment) and the speech recognizer is not guaranteed to provide correct result even if the audio signal is of good quality. More importantly, other modes can convey important information. Mutual disambiguation and error prevention are desirable features in multimodal interaction systems. Interaction systems that ignore some sources of input information will systematically fail to recognize many cases of spontaneous multimodal construction [Oviatt, 1999]. The unification-based integration of spoken and gestural input by Johnston et al. [Johnston, 1998] [Johnston et al., 1997] allowed the modalities to mutually compensate for each other's errors.

3 Probabilistic Multimodal Interaction Architecture

A household assistive robot mainly serves in domestic places such as the living room and the kitchen. For the moment, we are generally interested in two types of tasks which include navigating the robot (e.g. turn left/right, go there and go to the kitchen etc.) and requesting the robot to bring an object to the user (e.g. "give me the red cup on the table", "bring me my cup over there").

The system integrates three types of information, i.e. speech, gesture and gaze. Three channels of input data are first captured and recognized independently, generating semantic results, which are then fed to the multimodal fusion agent. A probabilistic method finds the most likely instance of each class in a task-orientated command table. The system then determines whether the recognized information is complete to trigger a physical action of the robot. A dialogue is activated if the robot needs to confirm with the user before its execution or more information is required to start any task. The flow chart of the system is shown in Figure 1.

The recognition of gestures and estimation of gaze direction have been described in our previous work [Li and Jarvis, 2010]. Our speech recognition here is simple, taking advantage of an existing speech recognition engine and followed by an intuitive words mapping method.

3.1 Speech Recognition and Words Mapping

Microsoft Speech SDK[1], which includes an advanced Automatic Speech Recognition (ASR) engine, is employed for our speech recognition component. The ASR analyzes the captured audio signal and transforms speech to text. A confidence value for each single word is generated which reflects the confidence of the ASR result.

A training stage, which involves reading aloud a list of provided articles, is normally important to improve the ASR's ability to understand the user's voice. Building a speech profile for each individual user is highly recommended rather than using the default profile. In our system, the speech profiles can be automatically switched to the current user, after the user's identity is recognized by a face recognition component.

[1]http://www.microsoft.com/download/en/details.aspx?id=10121

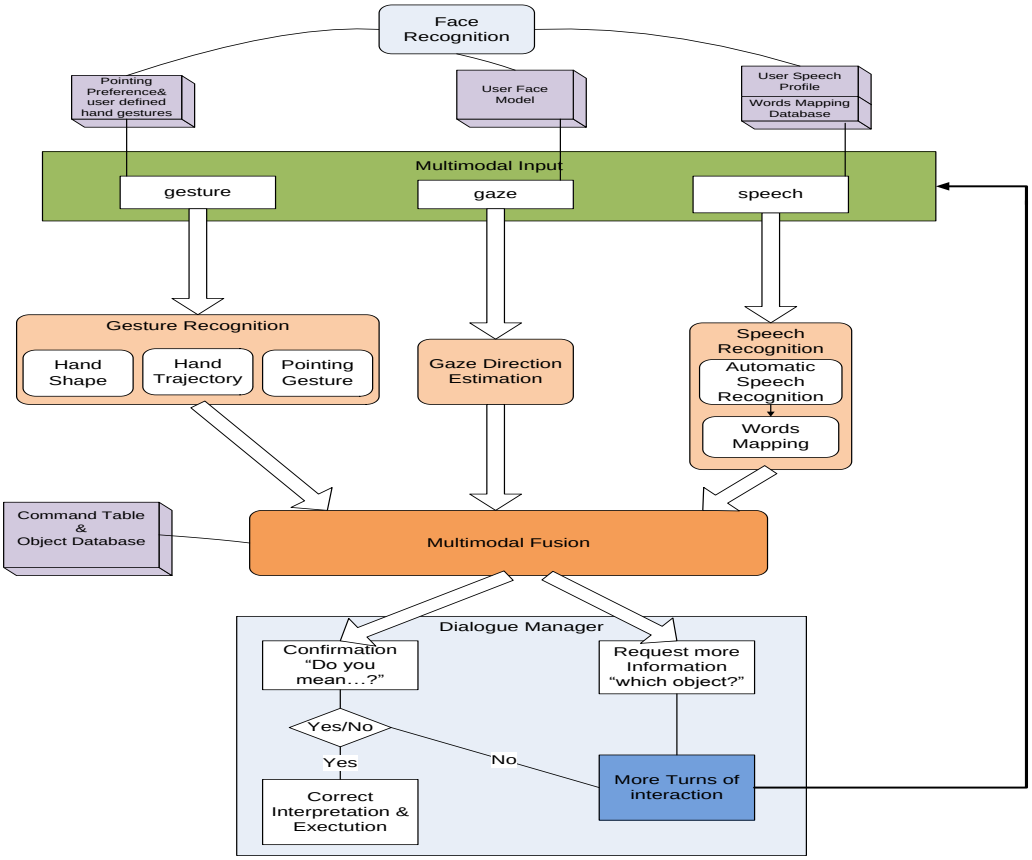

Figure 1: Flow chart of our multimodal interaction architecture

However, the effort needed in the training stage may be non-trivial. In fact, the recognition performance may be still far away from satisfactory even after hours of training, especially for the non-native speakers.

As discussed in [Li and Jarvis, 2010], we have considered the preference of pointing methods according to each individual's habits. Users can also define their own sets of hand gestures. Similarly, to compensate for insufficient training effort, we could refine the speech recognition result for each individual by a words mapping method.

Microsoft Speech SDK provides a context-free recognition result. It is then enhanced by the scenario defining the commands about which the users are likely to talk. For our household assistive robot, the words include the common objects in the living room and the kitchen, their properties and simple manipulation and navigating the robot to a specific place.

Therefore, it is reasonable to map some words which are unlikely to be spoken in these scenarios to the robot, like "cop" or "copy", to certain words which may be frequently referred in these places, like "cup". This is similar to how humans interpret sentences by considering the context of the conversation.

When some words are wrongly but consistently recognized as certain words by the ASR, the method projects them back to their original meanings. But if the errors of the ASR are not consistent, the method will fail. The disadvantage is that it may narrow down the range of topics which can be understood by the robot, because several words are mapped to one, and their meanings are changed. However, this tradeoff is acceptable and showed advantages in our experiments. The mapping database can be reduced or removed along with the development of the speech recognition engine in the future, but is currently important to compensate for the errors due to the accents of the speakers, insufficient training effort and low quality of the speech signal in a noisy environment.

Other factors can also be considered when applying the words mapping method. For example, some people may have difficulty in distinguishing certain colors, e.g. green, blue and cyan. Some people may not make a distinction between cups and mugs. This consideration would be desirable to endow the robot with humanization and sociality. As mentioned in [Shneiderman, 2002] [Jaimes and Dimitrova, 2006], for even more sophisticated robot, one should also consider about the relationship of a person's behavior with his/her personality, cultural, mood and the context in which the observed behavioral cues are encountered.

3.2 Gesture-Speech Alignment

In this subsection, we discuss the alignment problem between the recognized outputs from different modes when the command is issued multimodally.

Gesture-speech alignment involves choosing the appropriate gesture to ground key verbal utterance. This is important if there is more than one gesture occurring during the speech. For example, there may be two pointing gestures accompanying the sentence "put that

cup over there". One pointing is coupled with the word "cup" and the other one explicitly points out the spatial position referred to by the word "there". Some researchers have investigated this problem and [Eisenstein and Christoudias, 2004] summarized notable findings about gesture-speech bindings. In our system, the following points are considered:

- The gesture is usually close in time to the relevant keyword and normally precedes the keyword [Oviatt, 1999].
- Some word-gesture combinations are particularly likely. For example, a pointing gesture and the word "this", "that", "here" or "there" [Eisenstein and Christoudias, 2004].

However, the above is just a general phenomenon; actually there also can be individual differences in binding patterns. As discussed by Oviatt [Oviatt, 1999], in their experiments with a multimodal pen/voice system, users adopted either a simultaneous or sequential alignment pattern when combining speech and pen input. The integration pattern for each user was established early and remained consistent through all the trails.

In addition, a falsely detected gestures can also be filtered out by the gesture-speech alignment. Take the pointing gesture as an example. In our previous work [Li and Jarvis, 2010], a threshold for the duration of the pointing pose is used to tell whether a pointing arm is stable enough to be considered as a valid pointing gesture. Lowering this threshold would reduce the risk of miss detection, but with the side effect of more false positives. The false positives are filtered out based on two factors: one is the occurrence time of the speech and the detected gestures. If the pointing vector is far away from spoken keywords then it is filtered out (shown in Figure 2); the second one is the quality of the pointing vector itself (the quality is calculated based on the duration and within-group variance of the pointing vectors, see [Li and Jarvis, 2010] for details). Therefore, if there are more than one pointing gesture in the temporal neighborhood of the keyword, we do not simply keep the closest one (Figure 3). A Probability Related Score (PRS) of the gesture to be coupled with the spoken word is found by Equation (1), where Q_i is the quality of the i^{th} gesture, $\triangle T_i$ is the temporal distance between the occurrence of the i^{th} gesture and the spoken word, T_w is the predefined time window.

$$p_i = \begin{cases} 0 & , if \ \triangle T_i \ is \ out \ of \ T_w \\ w_1 \cdot Q_i + w_2 \cdot e^{-\frac{\triangle T_i^2}{T_w^2}} & , otherwise \end{cases} \quad (1)$$

Let $j = \arg\max_i p_i$, if p_j is over a specified threshold then the j^{th} gesture is coupled with this spoken word.

Similarly, only the captured eye gaze data within a time window specified by the spoken keyword is taken into consideration. The usage of the pointing line and gaze direction for selecting the target is described in section 3.3.2.

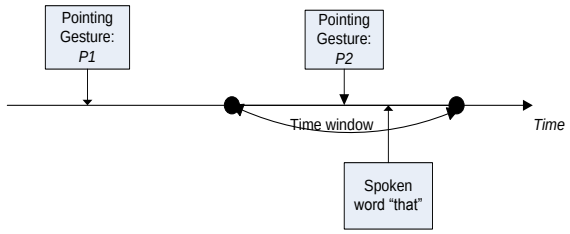

Figure 2: The pointing gesture *P1* is filter out by the recognized spoken word "that", because they are temporally far away from each other.

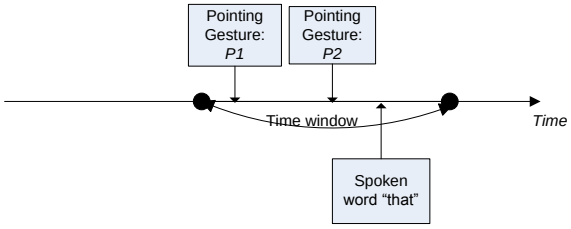

Figure 3: When two pointing gestures are detected within the time window of the spoken word "that", we do not simply choose the closest. Instead, the valid one is chosen using equation (1).

It is worthwhile to mention that in the implementation, the actual occurrence time of the event and the output time from the recognizer could be different. For example, the ASR only gives the recognized text after the voice signal has an obvious pause which indicates the sentence is complete. [Holzapfel et al., 2004] also took it into consideration. However, they did not explicitly find the occurrence time, but simply extended the constraints so that the fusion agent waited for a certain amount of time for a gesture event if the speech is ambiguous. In our system, the occurrence time of each event is found explicitly. The occurrence time of each spoken word is calculated by the start time of the phrase and the offset time of each word from the start time, both of which are generated by the ASR. The start time of the pointing gesture is found by the output time from the recognizer subtracting the duration threshold of the stability evaluation. The time of the gaze direction is calculated by the output time subtracting the estimated delay time.

3.3 Fusion at decision level

Information fusion methods are categorized to three levels (i.e. data, feature and decision) in [Sharma et al., 1998]. Fusion at the low level of raw data is the integration of information from the sensors of the same type; feature fusion refers to the closely coupled modalities such as speech and lips movement; decision fusion fuses the interpretation results of each individual mode [Russ et al., 2005]. In our task-orientated interaction system, the fusion is implemented at the decision level. Three recognizers for speech, hand gestures and gaze direction first process the captured data from different sources independently and generate recognition results at the semantic level which are then fed to the fusion agent.

In multimodal systems, it is ideal that each mode provides complementary information without errors and the fusion agent just assembles them as the final output. [Koons et al., 1993] presented a multimodal interface that

accepted speech, gestures and gaze input from a user. The command was given verbally while the location was provided from the user's pointing gestures or gaze. It was assumed that the recognition from each mode was accurate in advance.

However, the recognition result from each mode can be redundant, complementary or even contradictory. More importantly, in our realistic experiment, no recognizer can be assumed to be error-free: not all of the spoken words are correctly recognized; also the pointing line may be far away from the target; the gesture may be miss-detected, or in contrary several pointing vectors are falsely generated by the recognizer.

3.3.1 Task-orientated command table

According to the possible tasks involved in our household robot applications, we defined a Command Table which decides what information is needed to perform a task. Some examples are shown in Table 1.

Table 1: examples in the Command Table

Action	Object	Place	Direction
Pick up	√	×	×
Turn	×	×	√
Go to	×	√	×
Put down	×	×	×
Move	√	√	×
…			

Different constituents of a sentence are categorized into several classes such as action, object, direction and place. Only keywords are considered while some words like "could", "you", "please", are ignored. Each class has several instances. For example, the action could be "*pick up"*, "*put down" and "turn"* and so on.

An action word is almost always necessary to define a task. In some situations, the verb is omitted because it can be inferred from the context. Other basic constituents of a sentence may include objects, ("pick up that *apple*"), directional words ("turn *left*") and places ("go to the *kitchen*"). Each object has a unique ID since the robot must localize a specified object when there are several identical ones in the scene. An object has some attributes including *type, color, location, size and movability*. A place here refers to a 2D coordinates which can be specified by a pointing gesture or a predefined location. For example, for the command "go to the kitchen", actually, the spot where the robot should stand when it enters the kitchen is defined in advance.

A number of researchers have noticed that, although users preferred to interact multimodally rather than unimodally, they nonetheless did not issue every command multimodally [Oviatt et al., 1997] [Oviatt, 1999]. In addition, the information for some classes can be supplied by either mode or several modes at the same time. For instance, when the user says "pick up that red cup" issued with a pointing gesture, the spoken words "red cup" provide the type and color properties of the intended object, while the pointing gesture designates its 3D location. In some cases, even the phrase "pick up" can be accompanied with a mimetic hand movement.

3.3.2 Probabilistic method for multimodal information fusion

A probabilistic method is proposed here for information fusion. It determines whether the accumulated information is complete at the decision level and also takes into consideration the possible errors from each recognizer.

As mentioned in section 3.1, the ASR gives a confidence value for each recognized text. Among the properties of the objects which are concerned in our system, color and type are most likely to appear in the spoken utterances. Sometimes the user speaks the approximate region of the target. We only deal with simple cases, for example, "*center of the table*" and "bring me *the cup next to the box*".

Thus, the Probability Related Score (PRS) of an object to be referred by speech is calculated in Equation (2).

$$p_{speech}(o_i) = w_c \cdot color(o_i) + w_t \cdot type(o_i) + w_r \cdot region(o_i) \quad (2)$$

where w_c, w_t and w_r are the weights for color, type and region attributes separately. The value of the function $color(o_i)$ equals to 1, if the color attribute of the object o_i is the same as the spoken color; otherwise 0. The same applies to the functions $type(o_i)$ and $region(o_i)$.

The gesture recognition modes also provide a probability associated with each result, which were described in our previous papers. In detail, for the hand shape and trajectory recognition, the matching scores of each hand gesture candidate are used to calculate the probabilities of the recognition result [Li and Jarvis, 2009].

The pointing gesture recognizer and the gaze direction estimator generate deictic vectors [Li and Jarvis, 2010]. The duration time and the within-group variance of these vectors are used as the indication of their qualities. Each object to be selected is determined by the angle between the hand-object (or eye-object) line and the deictic vector (shown in Figure 4) by equation (3), where w is the quality of the deictic gesture.

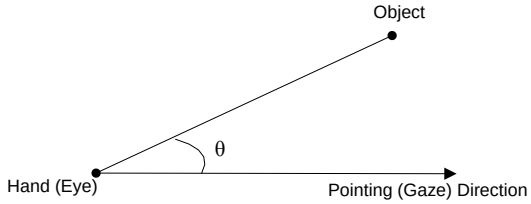

Figure 4: the angle between the hand-object (or eye-object) line and the deictic vector

$$p_{\text{pointing}}(o_i) = w \cdot e^{-\frac{\theta^2}{2\sigma^2}} \quad (3)$$

False positives of pointing gestures are filtered out by the gesture-speech alignment method when both modes have recognition output. However, if the spoken pronoun and nouns are not detected by the ASR, we still utilize the recognized pointing gestures, i.e. they are compared and chosen based on their qualities only.

The PRS from each mode is normalized before the fusion stage but the original PRS is also kept as an indication of the confidence of each recognition mode separately. For example, the confidence of the recognized spoken utterance is

$$p_{sp\_conf}(o_i) = w_c \cdot p_c + w_t \cdot p_t \quad (4)$$

where p_c and p_t are the original confidences of these

words from the ASR. For the moment, we only consider color and type, since the region property is always expressed as phrases. If $\max p_{sp\_conf}(o_i)$ is smaller than a specified threshold, it means the speech recognition result is not trustworthy.

For an autonomous and intelligent robot, it should be able to schedule the tasks according to the order of urgency or importance. However, we simplify this consideration because for the moment we do not issue a second command until the current task is finished or given up by the robot. But the "stop" command is one exception. This command is given the highest priority from the consideration of safety issues. If the stop command is recognized by any mode even with a low probability, the current operation is terminated immediately.

If none of the above two situations happens, the PRSs from each mode are normalized and then use equation (5) to calculate the final PRS of each instance. We take the object class as an example to illustrate our method. During a command requesting an object, each object has three PRSs from speech, pointing gesture and gaze separately. The final PRS of each object to be selected is calculated by weighted addition.

$$p(o_i) = \sum_j w_j \cdot \bar{p}(o_i, m_j) \quad (5)$$

where w_j is the weight for each mode, $\bar{p}(o_i, m_j)$ is the normalized PRS of object *i* selected by mode *j*. We give a high weight to speech as 0.5. Because the pointing gesture is strongly aimed, we assign its weight as 0.35. Gaze direction estimation in our system is relatively error prone, so its weight is 0.15.

The objects are sorted by their PRSs $p(o_i)$. If the top two values are different, then it indicates that the information for the object class is provided. Otherwise, it is still ambiguous about which object is intended.

The fusion agent decides whether the accumulated information from all input modes is complete based on the Command Table. If all the necessary information to accomplish a task is available, the system is then able to translate the command, according to a predefined Translation Table, to several sequential steps of executions which can be performed by the robot. For example, "bring me that cup" can be translated to three steps: localize the cup, grab it, and go to the user.

3.4 Dialogue Manager

More turns of interaction will be requested by the robot under three kinds of situations in our system.

- The perceived information is not complete for any task in the command table. This may be caused by two factors: some information is still missing or some information is ambiguous (i.e. more than one instance of a certain class has equally largest PRS). The robot needs to clarify the missing or ambiguous information from the user, for example, by saying "which object do you want to pick up?"
- None of the recognition from all modes is trustworthy. Although we can still find the action or object which has relative high PRS compared to other instances in the same class, it is prone to get wrong decision when using the non-trustworthy information sources.
- The robot's interpretation is incorrect. If the execution of a task involves risky operation to the user or the robot itself, or requires complicated operation or takes long time to finish, the robot should confirm the command before it starts the operation. This kind of confirmation is normally simple: for example, the robot asks "do you want me to go the kitchen?", and waits for a "yes or no" answer. However, currently we require the robot to confirm its interpretation with the user for every command, because it is suitable and effective when testing the cognitive ability without the real operations by the robot.

Our system utilizes the recognized information to the most degree. When possible, it tries to infer the missing information rather than request from the user, so that the number of cycles of interaction can be reduced. Unlike the work by [Harte and Jarvis, 2009], in which the whole utterance is ignored if the action word is not recognized, in our system the action word can be inferred if other classes, e.g. objects and directions, are recognized. It searches in the database for what action is associated with the recognized components. If only one is associated, then the action can be decided with a simple confirmation. For example, the robot will ask "do you want me to *turn* left?", if the word "left" is correctly recognized while the action "turn" is inferred because there is no other verb related to "left". Actually this is also affected by the state of the object (or the robot), e.g. if the robot is holding a cup, then it hears the user says "the cup", but misses the verb, it is most likely that the user wants the robot put it down rather than "pick up".

Although requesting more information is necessary and asking for confirmation reduces the risk of wrong operations, however, they influence the naturalness and effectiveness of the interaction, which may decrease the satisfaction from the user if too many turns of conversation are required.

It is also important that the robot reports its state to the user such as "waiting for command", "working on the task", "task finished" or "task failed". Synthetic audio feedback is implemented using the synthetic speech engine in the Microsoft speech SDK.

4 Experiments and Results

Our evaluation was conducted in a laboratory-like office room. The setup consists of many objects which are often used in a kitchen and a living room, including cups, plates, bowls and boxes etc. Some objects are distinct in color while some objects are identical. The objects were placed on a table, on the floor, or on a chair. Figure 5 shows a partial view of the setting. The participants gave commands to the robot, such as "go there", "turn left", and requesting the robot to bring some objects in the scene.

4.1 Experiments design

Some spoken descriptions about the objects, such as big, small, long and short, are not included in our experiments, although they are widely used in our daily language, because these words are subjective for each individual and relative to other objects in the scene too, which impose more difficulties to the system. Also, the subjects were advised not to refer to the usage of objects, e.g. "water bottle" and "coffee cup".

Figure 5: A snapshot of experiment setup. The potential targets are marked by red circles. They are placed at various locations, at different height and distance levels. Some of the objects are not shown in this picture due to the view of the camera.

The user's identity is normally recognized before the subsequent interactions, so that the user's profiles, including the self-defined hand gestures [Li and Jarvis, 2009], pointing method preference [Li and Jarvis, 2010] and trained speech profile are loaded. This helps to obtain better interpretation results from each mode, giving the multimodal fusion a good starting point. However, if the user is new to the robot, the default profiles will be used.

To test the multimodal interaction method and compare with the unimodal speech input, we designed and conducted three categories of experiments.

- Firstly, the subjects try to communicate with the robot by their speech only. This would be easy for navigating tasks, but may become difficult for requesting certain objects, because we had arranged some identical objects in the scene. Some objects are distinct in color or type, while the identical ones can only be distinguished by their spatial locations.
- Secondly, the users choose the means of interaction in their natural way. In this stage, the subjects are free to choose unimodal or multimodal method to communicate with the robot. They can still use speech alone, and they can also combine speech and gestures. In addition, the gaze direction is always captured and used.
- The users are encouraged to use both speech and gestures to provide as much information as possible to the robot. They would try to use a pointing hand when referring to an object and also speak the properties of the object, even if there is only one object of that type. The information provided can be redundant at this stage. The idea is to treat the robot like a young child. We do the same when we meet some foreigners or small kids in order to increase the chances of being understood. However, it is not restricted. Actually, the subjects did forget to do so in some trails.

The robot will try to figure out the user's intention using the available information from all modes and the pre-built database including the user's profile, the objects' properties, the command table and the context. If a task is determined, it will confirm with the user, for example by saying "Do you want me to pick up the cup whose ID is number 1". Note that it would more natural and preferred if the robot could also point to the referent to confirm its interpretation like in [Sugiyamal et al., 2006], however, it is more effective to speak out the object's ID number when we only test the cognitive part without the real operation by the robot. This number is only used for the confirmation purpose and not used to request objects. If the interpretation is correct, it is considered as a successful interaction.

4.2 Results

4.2.1 Speech only

The Microsoft SR Engine is supposed to work well under certain conditions, requiring quite environment, good quality of input signal and extensive training efforts. However, our experiments were conducted in an office environment, with other colleagues working and a copy machine and an air conditioner operating in the background. Only one of the four subjects is from English speaking country and only two persons have taken short time for training the ASR (a person spent half an hour and another spent 10 minutes).

Ten spoken utterances by each subject were used to build the words mapping database for individual. This has significantly improved the speech recognition result. In details, among the testing trials by speech only, 100 sentences have been spoken, comprising 594 keywords. Only 296 words were correctly (with any ASR confidence value) recognized by the original ASR; while 437 words were correctly recognized when we apply the words mapping database. Therefore, the subsequent processing uses the mapping results rather than the original ones from the ASR. Note that the mapping method only works when the errors of the ASR in the training dataset and in the testing dataset are consistent.

In total, 48 out of 100 spoken commands are successfully recognized in the testing stage. Besides inconsistent ASR errors, many failures are caused by the pauses and repetition in the long sentences. Unnecessary pauses separate one complete sentence into two. Repetition, hesitation and mispronunciation-correction, for example, "give me, *erm*…, that red, no blue cup (*pause*) at the right corner of the table", could not be handled by our current system. Long sentences have low chances to be understood. To interpret the spatial relationship between the intended target and other referents, like "the cup next to the box", every word must be correctly recognized. It can even be misleading if parts of the information are missing. For example, for the sentence "give me the cup next to box", if the part "the cup next to" is miss detected, what the robot hears becomes "give me the box".

Also the user's command itself may be ambiguous. In seven trials, the user did not notice the existence of other identical object when he is looking at the target. It also shows that the gaze direction is a good indication of the user's attention and has the potential to resolve the ambiguity.

4.2.2 Multimodal/Unimodal interaction in a natural way

When the subjects are free to choose unimodal or multimodal method in their natural ways, spoken sentences are much shorter in most cases, with few pauses and repetition. In total, 106 out of 151 commands were

successfully recognized. Users normally select the interaction methods that could eliminate linguistic complexities. The typical commands combine the speech like "give me that cup" and a pointing gesture indicating the location of the desired cup. Our experiments observation confirms the finding of [Oviatt, 1997] that users showed a obvious preference to interact multimodally rather than unimodally, especially in the spatial domain.

However, this preference does not guarantee that they would issue every command multimodally. In the experiments, in most cases, the actions and the objects' color and type properties were spoken verbally; the spatial information (objects' position) was indicated by pointing gestures explicitly and by gaze implicitly. In addition, if there is only one object of a type (e.g. only one box in the scene), the users may also use the unimodal speech, because this property is sufficient to distinguish it from other objects.

In the trials of requesting an object, inputs of the three modalities were used together to calculate the PRSs of the objects as in equation (5). Overall, 96 out of 140 object-requesting commands were successfully recognized. The pointing gestures helped to select the correct object among several identical ones, in which situation the unimodal speech could not succeed. The spoken words were also important when the pointing vector and the gaze direction were not able to select correctly from a crowded setup. Since multiple modes tend to provide complementary, rather than redundant, information, the loss of information by the independent recognizers is likely to lead to failures. Many failures were because some spoken keywords were not recognized and meanwhile the deictic gestures were not accurate enough to select the correct target.

4.3.3 Multimodal interaction in an encouraged way

When the users are encouraged to provide as much information as possible, despite of redundancy, the system provides the best performance among these three categories of experiments, as expected. 125 out of 148 trials are successfully recognized. It seemed that the users have input redundant information to the robot, however, actually due to the possible errors in the recognizers of each mode, the loss of information is sometimes serious, so the originally redundant information may become complementory. Even if the recognized results from different information sources are really redundant, it would not affect the final decision adversely.

It showed great advantages in the object-requesting commands. For example, when the user was free to use their natural way, shorter sentences, e.g. "give me that bowl", are often preferred; whereas if the user was encouraged to provide more information, the sentence "give me that red bowl in the center of the table" is used. Both utterances accompany with a pointing gesture. If the pointing vector is not accurate, and the noun "bowl" was miss-heard, the word "red" could still largely increase the chance of picking out the correct object, if there were only one or two objects are in red color. However, it did not show improvement for navigating commands when combining speech and hand gestures, because these commands were really short utterances, e.g. "turn left/right", and already have a high recognition rate by speech recognition alone. But the gestures would be helpful if the audio input degrades much.

It shows that, the system does generate better, at least not worse, results when the user provides redundant information. It is particularly practical when the user wants to maximize the recognition results and put the naturalness at the second priority.

5 Conclusion and Future Work

We have described our multimodal interaction system for a household assistive robot. The system integrates the input information from three modes, i.e. speech, gesture and gaze. The proposed probabilistic fusion method finds the most likely instance of each class, and determines whether the accumulated information is complete to take an execution according to a predefined Command Table. The robot reports its status to the user and also could request more information from the user in a two-way interaction pattern. Experiments showed that the participants preferred to interact multimodally, and our multimodal interpretation system significantly outperformed the unimodal speech input.

A more sophisticated speech interpretation method seems to be needed to deal with the ASR errors and the variety of the spoken utterance. For example, when people say "give me the orange juice", they actually mean the container of the juice.

In our future work, we will extend the vocabulary of the speech and gesture system. We also hope that the robot could notice the user's emotion by facial expression and affective gesture recognition. It would be more interesting and challenging if multiple users are present and communicating with the robot at the same time or alternatively.

Reference

[Bolt,1980]. "“Put-that-there”: Voice and gesture at the graphics interface." SIGGRAPH Comput. Graph. 14(3): 262-270.

[Carpenter,1992]. The logic of typed feature structures: with applications to unification grammars, logic programs, and constraint resolution. Cambrige, England, Cambrige University Press.

[Codella, et al.,1992] Interactive simulation in a multi-person virtual world. Proceedings of the SIGCHI conference on Human factors in computing systems. Monterey, California, United States, ACM: 329-334.

[Eisenstein and Christoudias,2004]. A Salience-Based Approach to Gesture-Speech Alignment. Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics annual meeting.

[Gunes and Piccardi,2006]. "Bi-modal emotion recognition from expressive face and body gestures."

[Harte and Jarvis,2009]. Multimodal Human-Robot Interaction in an Assistive Technology Context. Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions (ACHI '09).

[Heidemann, et al.,2004] Multimodal interaction in an augmented reality scenario. Proceedings of the 6th international conference on Multimodal interfaces. State College, PA, USA, ACM: 53-60.

[Holzapfel,2008]. "A Dialogue Manager for Multimodal Human-Robot Interaction and Learning of a Humanoid Robot " Industrial Robots 35(6): 528-535.

[Holzapfel, et al.,2004] Implementation and evaluation of a constraint-based multimodal fusion system for speech and 3D pointing gestures. Proceedings of the 6th international conference on Multimodal interfaces. State College, PA, USA, ACM: 175-182.

[Jaimes and Dimitrova,2006]. "Human-centered multimedia: culture, deployment, and access." Multimedia, IEEE 13(1): 12-19.

[Johnston,1998]. Unification-based multimodal parsing. Proceedings of the 17th international conference on Computational linguistics.

[Johnston, et al.,1997] Unification-based multimodal integration. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Madrid, Spain, Association for Computational Linguistics: 281-288.

[Kaiser, et al.,2003]. Mutual Disambiguation of 3D Multimodal Interaction in Augmented and Virtual Reality. Proceedings of the 5th international conference on Multimodal interfaces, Vancouver, British Columbia, Canada.

[Kaur, et al.,2003]. Where is "it"? Event Synchronization in Gaze-Speech Input Systems. International Conference on Multimodal Interfaces.

[Koons, et al.,1993]. Integrating simultaneous input from speech, gaze, and hand gestures. Intelligent multimedia interfaces, American Association for Artificial Intelligence: 257-276.

[Li and Jarvis,2009]. Real time Hand Gesture Recognition using a Range Camera. Australasian Conference on Robotics and Automation (ACRA), Sydney, Australia.

[Li and Jarvis,2010]. Visual interpretation of natural pointing gestures in 3D space for human-robot interaction. Control Automation Robotics & Vision (ICARCV), 2010 11th International Conference on.

[Moshier,1988] Extensions to unification grammar for the description of programming languages. Michigan University of Michigan Ph.D.

[Neal and Shapiro,1991]. Intelligent multi-media interface technology. Intelligent user interfaces, ACM: 11-43.

[Oviatt,1997]. "Multimodal interactive maps: designing for human performance." Hum.-Comput. Interact. 12(1): 93-129.

[Oviatt,1999]. "Ten myths of multimodal interaction." Commun. ACM 42(11): 74-81.

[Oviatt, et al.,1997] Integration and synchronization of input modes during multimodal human-computer interaction. Proceedings of the SIGCHI conference on Human factors in computing systems. Atlanta, Georgia, United States, ACM: 415-422.

[Russ, et al.,2005]. Semantic based information fusion in a multimodal interface. International Conference on human-computer interaction, Las Vegas, Nevada, USA.

[Sharma, et al.,1998]. Toward Multimodal Human-Computer Interface. Proceedings of the IEEE, special issue on Multimedia Signal Processing.

[Shneiderman,2002]. Leonardo's Laptop: Human Needs and the New Computing Technologies. Cambridge, MIT Press.

[Stiefelhagen and Fugen,2004]. "Natural human-robot interaction using speech, head pose and gestures." Proceedings of 2004 IEEE/RSJ international conference on Intelligent Robots and Systems: 2422-2427.

[Sugiyamal, et al.,2006]. Three-Layer Model for Generation and Recognition of Attention-Drawing Behavior. Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, beijing, China.

[Zhang, et al.,2004]. A gaze and speech multimodal interface. 24th International Conference on Distributed Computing Systems Workshops.