A Robust Selection System using Real-time Multi-modal User-agent Interactions

Katsumi Tanaka
Toshiba Cooperation, Kansai Research Laboratories
8-6-26 Motoyama-minami-cho
Higashinada-ku, Kobe 658-0015, Japan
+81 78 435 3105
tanaka@krl.toshiba.co.jp

ABSTRACT

This paper presents a real-time object selection system which can deal with gaze and speech inputs that include uncertainty. Although much research has focused on integration of multi-modal information, most of it assumes that each input is accurately symbolized in advance. In addition, real-time interaction with the user is an important and desirable feature which most systems have overlooked. Unlike those systems, our system is intended to satisfy these two requirements. In our system, target objects are modeled by agents which react to user's action in real-time. The agent's reactions are based on integration of multi-modal inputs. We use gaze input which enables real-time detection of focus-of-attention but has low accuracy, whereas speech input has high accuracy but non-real-time feature. Highly accurate selection with robustness is achieved by complementary effect through probabilistic integration of these two modalities. Our first experiment shows that it is possible to select target object successfully in most cases, even if either of the modalities includes great uncertainty.

Keywords

Multi-modal interface, uncertainty, real-time interaction, gaze, speech, agent model

INTRODUCTION

Recent multi-modal interface systems have a wide variety of inputs such as natural language, speech, gaze and gesture. Their main goal is to obtain the user's intention by using integration techniques for various kind of inputs[1]. They have achieved successful results in integrating symbolized information as in resolving references in natural language sentences[2]. However, their approaches have a problem when dealing with not symbolized inputs like speech and vision. In real situations, recognition systems based on speech and vision include great uncertainties that mainly result from environmental changes (noise, light, etc). A lack of real-time interaction is also considered to be a problem. Usually systems accept simultaneous inputs from user but most of them have batch input-integration processes after all inputs are finished. That causes delay in response to user, and increases the complexity of the integration process.

We aim to satisfy those two desirable requirements. Our current task is a simple object selection system in which a user tries to select his/her intended object by using multi-modal inputs, and is given a real-time feedback by the system. To do it, we employ an architecture in which each agent corresponds to each object. Every agent acts autonomously by predicting the user's intention (selected or not) based on its perception and reacts to the user so as to inform its situation to the user. Ultimately, repetition of such interactions leads to a successful selection. The usefulness of applications in this area has long been recognized [3,4]. In the cited works, gaze was the primary modality and it works as a cue for controlling the system's behavior. Notably, [4] reported the construction of object selection systems based on gaze input using a highly accurate eye-tracker. They successfully conceptualized the gaze-based software through practical applications. However, their systems are subject to the same problem as other multi-modal systems in uncertain situations derived from inaccurate and unstable input devices.

We are convinced that an agent-based model can lead to a solution. Agents are able to perceive, act autonomously, and communicate with users or other agents. These features are suitable for real-time interaction and problem solving in uncertain situations.

In this paper, first we present a basic selection system using agent-architecture in which each object is modeled on autonomous agents which predict the user's intentions based on inaccurate inputs and give appropriate feedback

IUI 99 Redondo Beach CA USA
Copyright ACM 1999 1-58113-098-8/99/01...$5.00

to users. Input devices which we used are a gaze-detection system and a speech recognition system which may include inaccurate results. Secondly, we report experimental results that confirm the system's usability (the rate of successful selection and time for selection). These results confirm that the overall system works well even in uncertain situations that arise from inaccurate input devises. In addition, the results clearly demonstrate the advantage of using multi-modal inputs: integration of gaze and speech information gives far better results than using only one modality.

AGENT MODEL FOR SELECTION

Figure 1 illustrates the basic process of a single agent. Each agent executes three steps (observation, detection, and action) repeatedly. In the observation step, an agent interprets the information from input device as a degree of focus-of-attention respecting itself. For example, an input from the gaze detection system is interpreted as the degree of being looked at by a user. A speech recognition result is dealt with in the same way, that is, speech input is interpreted as the degree of being named. In the detection step, an agent estimates the user's intention (select or not) based on its belief. Agents obtain their beliefs in advance in the learning step which is mentioned in a later section. In our system, this step outputs a probability of the user's intention to select the agent. In the action step, agents show their reactions to the user according to the user's intention. This is perceived by the user as activation of agents. He/she feels as if they activate his/her intended agent by looking and calling. Repeating these steps in a speedy cycle (currently 0.5 sec), we realize an almost real-time interface between agents and a user.

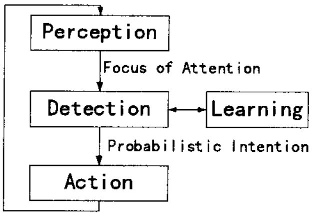

Figure 1: Basic process in an agent

Learning beliefs of agents

Learning belief of a user's intention from inputs to an agent should be based on factual data. To do so, we employ a Bayesian approach. According to Bayes' theorem, a posteriori probability that the user selects the agent in the case of a particular input *Input* is given by

$$P(Selected|Input) = P(Selected)P(Input|Selected)/P(Input)$$

If a prior $P(Selected)$ and a likelihood $P(Input|Selected)$ are given in advance, we can compute a posteriori probability. We need a learning step which obtains them together with teaching signals which give the information that each agent is selected or not by the user.

When dealing with a multi-modal system, we should consider causal relationships between input modalities. Recent progresses in Bayesian networks [5] helps us model complicated causal relationships of this kind in probabilistic framework. A Bayesian network we used is illustrated in Figure 2, where each node denotes a proposition and each directed arc denotes a causality between propositions. A "selection" node means user's intention to select. "Gaze-n" and "Speech-n" nodes mean that there exist gaze and speech inputs at n clocks before, respectively. We have the following intuitions of causal relationships.

1) Being currently looked at is a cause of intention for selection.

2) Being called and looked at simultaneously in the past time is a cause of intention for selection.

We use speech inputs in the past because it is difficult to do real-time processing with current speech recognition systems. As speech recognition results reach agents after a little delay, we need to handle speech inputs in such a way that they can match gaze inputs at the same clock. Uncertainty in Bayesian networks is represented by probabilities (priors or likelihoods) of propositions based on causal relationships. Obtaining these probabilities corresponds to our learning process. In order to obtain these probabilities, we employ a popular approach, a conditional probability table (CPT). An example of a CPT, which represents "speech-1" node, is shown in Figure 3. In the figure, columns denote the degree of focus of attention obtained by speech inputs, where the domain is divided discretely into 10 regions. Rows denote all the combinations of causal sources that correspond to the source of the directed arc in Figure 2. In the learning step, we give all the agents' intentions (*Selected* or *NotSelected*) with gaze and speech inputs normalized from 0 to 1. Agents add the counts to corresponding cells in the CPT.

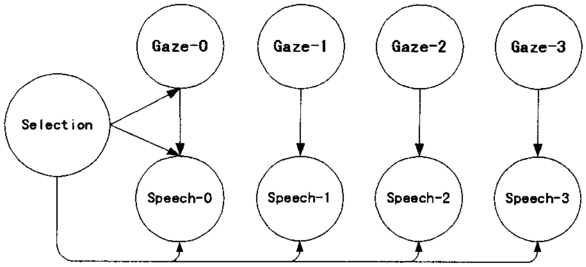

Figure 2:Bayesian network for intention detection

Speech-1\ Gaze-1 & Selection	0-0.1	0.1-0.2	….	0.8-0.9	0.9-1
0-0.1 & Selected					
0-0.1 & NotSelected					
…					
0.9-1 & Selected					

Figure 3: An Example of conditional probability table

In the detection step, posteriori probabilities of agents are computed using the values of the CPT and Bayes' Theorem.

Mechanism of making action

We limit an agent's action to the most basic cases: activating itself or not. An agent's action should reflect the user's intention and importance of that intention when it is selected. It can be modeled by utility theory. Using utility theory, an agent's expected utility for activation *E[u(Activate)]* is formulated as follows.

$$E[u(Activate)]=P(Selected|Input)u(ActivateWhenSelected)$$
$$+\ P(NotSelected|Input)u(ActivateWhenNotSelected)$$

In the above formula, *u(ActivateWhenSelected)* (*u(ActivateWhenNotSelected)*) denote the utility value when the agent activates when it is selected(not selected), and so they mean the importance level of each case. We use *E[u(Activate)]* for determining each agent's activation level. To make the interface robust, *E[u(Activate)]* are accumulated over time and used as the final activation level.

THE SYSTEM

Our system configuration is shown in Figure 4. The system comprises input systems and devices (a gaze detection, a speech recognition, a mouse, and a footswitch) and agent-based selection interface software which comprises a manager and agents implemented on a PC.

Input Systems and Devices

All the inputs are gathered to the manager and are then distributed to each agent's perception step. We used a gaze-detection system based on a face-recognition technique[6]. The system registers face patterns when a user looks at specified regions in advance. After that, similarities between an input face pattern and registered ones are calculated by pattern-matching method. An advantage of the system is that it allows the user's flexible posture because simple pattern-matching does not require any geometric measurements of the user's absolute eyeball positions. But a system with high resolution does not currently exist. Using the current system implemented on SGI O_2(R10000 CPU), we can obtain pattern-matching results for the regions divided by 2x2 on PC display 8 times/sec. Then an agent calculates the focus of attention respecting itself from these similarities based on geometric positions of itself and the divided regions. Use of this method sometimes causes low accuracy, particularly when there are many selected objects (and thus, size of each object becomes smaller). The speech recognition system we used is a word-based, speaker-independent one which allows high-level noises. In a manner similar to the treatment of gaze inputs, agents calculate focus of attention from phonetic similarity output from the system, based on their names. Considering the combination of these input systems, gaze inputs work as a modality with low accuracy but speedy response, whereas speech inputs work as a modality with high accuracy but delayed response. A mouse and a footswitch are used for giving signals of starting and stopping learning, and confirmation of selection.

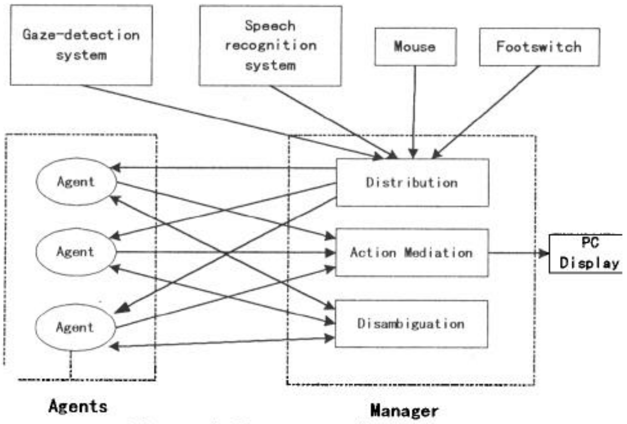

Figure 4: System configuration

Manager and Controlling Output

The manager acts as the overall interface system by mediating information between inputs and agents, and by regulating visual outputs to a PC display. The appearance of a PC display is shown in Figure 5. Each agent is represented as a squared potion of which it is conscious as its bodily region depicting its name and a level meter. A level meter whose length expresses the agent's activation level works as a feedback to a user. The user can confirm his/her final selection by switching on the footswitch when his/her intended agent's level is full. Selection is considered to be completed successfully if there is a sole agent with a full level. In the case that there are multiple agents with full level (surviving agents), the system performs a disambiguation routine. This disambiguation process is simple. Agents other than surviving ones disappear, whereas surviving agents change their position so that the distances between each other become greater. This repositioning is to complement the low accuracy of gaze detection system. After that, surviving agents execute the same perception-detection-action processes until they are disambiguated to a sole agent. Rather than have

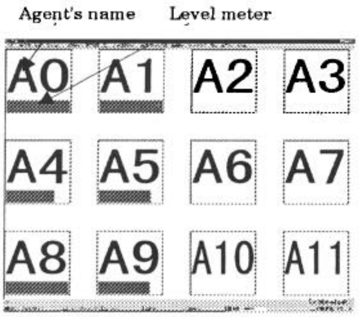

Figure 5: Appearance of display

centralized control of the system, the manager works as a mediator between agents.

EXPERIMENTAL RESULT

We designed an object-selection experiment and applied our system to several cases. The experiment consists of two steps. First, a user has the system learn by giving teaching signals and inputs using gaze and speech (learning step). Next he/she tries to select his/her intended objects using gaze and speech inputs in the same way as in the learning step. The results are shown in Table 1 and 2. In every case, 12(48) agents are arranged in an orderly 4x3(8x6) array in a display. We think that the rate of successful selection (RSS) and the average selection time (AST) are essential yardsticks for evaluating selection systems. We also experimented regarding cases in which the system uses only one modality (gaze or speech input) for comparison. We attach the data which show the performance of detection/recognition of gaze/speech.

Table 1: Performance of selection (in 12 agent case)

Items\ Modalities used	RSS	AST	Gaze Performance (Recall/Precision)	Speech Peformance (Recall/Precision)
Gaze & Speech	100%	6.4sec	69.1%/38.4%	83.3%/83.3%
Gaze Only	83.3%	12.0sec	49.4%/31.8%	(-/-)
Speech Only	83.3%	8.1sec	(-/-)	84.2%/84.2%

Table 2: Performance of selection (in 48 agent case)

Items\ Modalities used	RSS	AST	Gaze Performance (Recall/Precision)	Speech Peformance (Recall/Precision)
Gaze & Speech	95.8%	9.0sec	62.1%/6.9%	84.6%/84.6%
Gaze Only	63.3%	22.0sec	59.7%/9.7%	(-/-)
Speech Only	77.6%	9.9sec	(-/-)	77.4%/77.4%

RSS: Rate of Successful Selection
AST: Average Selection Time

Implications

Both tables clearly imply the following.

1) RSSs of our system are always greater than any recalls and precisions of modality inputs(gaze/speech) in any cases. That indicates that our agent model basically works well in selection. Particularly noteworthy is the fact that the performance markedly improved when only gaze-detection was used. That indicates that the agent's autonomous behavior together with disambiguation procedure is effective even in the case of uncertain situations.

2) The performance (RSS and AST) of our system using multi-modal inputs is always higher than those systems which use uni-modality(gaze/speech). Multi-modal integration in the system clearly works well. Specifically, it is a evidence in favor of complementarily of modalities with heterogeneous features (accuracy and real-time processing).

Future Works

We don't think that the system's current performance is sufficient for practical applications. Having analyzed the logs of the experiment, we have identified the following as subjects for future work.

1) Improvement of the detection method is needed to realize more appropriate behaviors of agents. Estimated successful detection rate was 37-58% in the overall experiment. We may need more advanced Bayesian techniques to improve on it.

2) Interaction between agents is needed for more speedy selection. It can lead a reduction in the average selection time, particularly in the case that the number of selected agents increases. There are several possibilities other than the repositioning we tried this time. We think that it would be effective for changes in currently unused properties (size, colors, etc) to be handled cooperatively by multiple agents.

CONCLUSIONS

There are many potential applications for hand-free object-selection systems. To achieve robust selection, we have presented agent-based interaction mechanism using gaze and speech inputs and visual feedback. The result of a basic experiment is encouraging. We achieved 96-100% successful selections using an inaccurate (7-85% average precision rate) input system. An agent-based implementation with a good detection mechanism is essential for achieving a robust selection process through real-time interaction. We expect the incorporation of further progress in this area, especially respecting multi-agent interaction, to lead to the practical application of such systems.

ACKNOWLEDGEMENTS

The author would like to thank members of Toshiba Kansai Research Laboratories who discussed the work.

REFERENCES

1. Maybury, M. T. Research in Multimedia Parsing and Generation. *Artificial Intelligence Review*, 9(2-3),103-127, 1995
2. Cohen, P. R. et al. QuickSet: Multimodal interaction for distributed applications., *Proceedings of the Fifth Annual International Multimodal Conference*, 1997
3. Bolt, R.A. Gaze-Orchestrated Dynamic Windows. *Computer Graphics*, 15(3), 109-119, 1981
4. Jacob, R.J.K. What You Look at Is What You Get: Eye Movement-based Interaction Techniques, *CHI'90 Proceedings*, 11-18, 1990
5. Heckerman, D. A Tutorial on Learning with Bayesian Networks. *MSR Technical Report*, MSR-TR-95-06,1995
6. Yamaguchi, O. et al. Face Recognition Using Temporal Image Sequence. *IEEE Face and Gesture Recognition '98*, 318-323, 1998