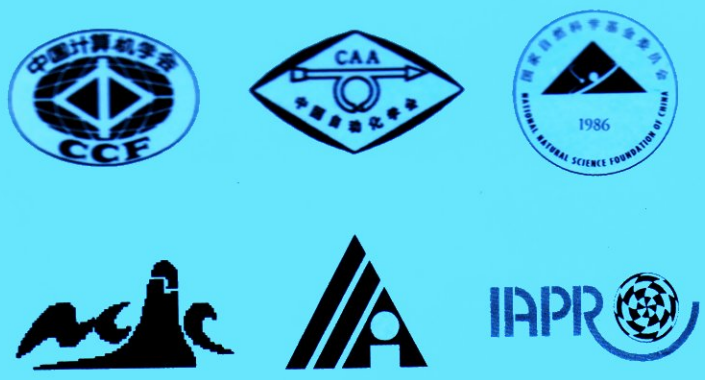
Proceedings of the First International Conference on Multimodal Interface

(ICMI'96)

October 15-17,1996 Beijing, P.R.China

Chief Editor: Wen Gao



Tsinghua University Press
Beijing

CONTENTS

Build Perception Model To Multimodal Interface

Datong Chen, Wen Gao & Xilin Chen
Department of Computer Science and Technology
Harbin Institutes of Technology
Harbin 150001 , P. R . China
e-mail : cdt@vilab.hit.edu.cn

ABSTRACT

The problem about how to build perception model for multimodal interface of an application system will be managed to discuss in this paper. To serve a multimodal interface, we propose a perception model with three levels. In the first level, it translates physical signal into some basic elements and we provoke to build standard set of elements for each physical channel. These elements are built into certain models that the application needs in the second level. The third level is to prepare abstract notions and basic combining or reasoning functions to the users. An application instance through out the while paper convinces us that all these steps are reasonable.

Keywords : multimodal interface, perception model, channel elements

INTRODUCTION

Human beings usually interact with the outer world through multiple channel upon its body functions. They favorite on sight, listening, smelling, touching, tasting and make a normal daily life on the multiple channels' information. As well as human being, an application system on computer may also ask for better running with the interdependent data from different channel independently. Unfortunately, most computers can only receive the information from simple channels such as keyboard and mouse. Each signal of these simple channels is a certain abstract information and has been used as varied meanings in different applications. This abstract information does little help for a computer application to agree each other about the same outer scene. To offer sufficient embodied information to a system, multimodal interface has become a popular research field these years. With the growing of computer power and the progress of artificial intelligence technology, there is a possibility to provide a more intelligent interface for an application system. With such interface, computers are able to receive the information from not only keyboard and mouse, but visual, audio, touching, and so on. It provides more comfortably to interact between computer and the outer world. Recently, people assume that such kind of interactions can conquer the difficulties we encountered in former interface system we used. For example, visual model information may help a Radar navigated missile to distinguish the real target from the mental forgers. An auto control car may use laser distance measuring device to make up of the visual difficulty about distinguishing where is the edge of the road and where is shadings of the trees. Further more, multimodal interface may be employed into many fields and will keep the our life stepping to intelligence. Without multimodal interface, a system seems like a handicap person who is difficult to get a powerful mind or a strong body. On the other hand, multimodal interface will improve human computer interface greatly. It commonly tends to introduce knowledge about human being into HCI (human-computer interface) which aims at building human centered environment and helps computer to understand the human behavior. The idea of introducing multimodal interface into human-computer interaction was wildly developed in the mid 1980's by Alan Kay. Until now, more efforts have been made by other researchers in the world. Takebayashi has developed a spoken dialogue system (SDS) which interacts with the user by natural language and verify a user by some special switches [1]. Nagao et al also developed a SDS to join human-human conversations. It offers an ability to identify users by their voice [2]. Osamn et al made a great progress that they developed a multimodal interface system involved both visual and audio functions to communicate with users [3]. In their system, a human face recognition method was used to name the user. In our work, we hope to build multimodal interface for out system which tend to understand human behaviors by both natural language and body language upon a visual and audio base. The remainder of this paper is organized as following, a perception model of multimodal interface with three levels – translator, constructor, and abstracter are proposed in second section, the detail of each level is discussed differently in third, fourth and fifth section, and the conclusion and future work are presented in the final section.

PERCEPTION MODEL

In 1994, Bernsen had developed a modality theory to support multimodal interface design.[4] He emphasis the importance of modeling in the design process. Following his theory, we manage to build our perception model according to the functions of multimodal interface.

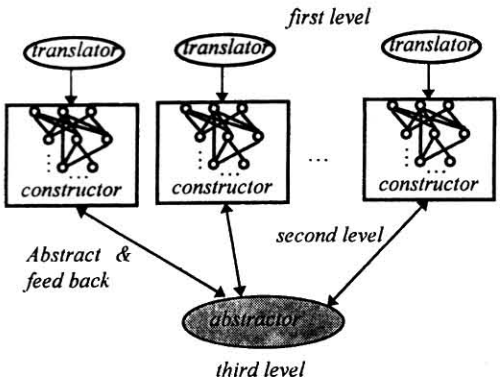


Fig.1 perception model of multimodal interface

A farther analysis about the function of multimodal interface is shown in Fig. 1. There are three levels. Each level performs its special functions. As a general interface, to build multimodal interface for a system, physical input and output devices are necessary. Through these devices, the system can get the information from the outer world and express its "feeling" to the outer world. How to produce these devices is the work of the engineer. What is we really concerned is how to deal with this original information - physical signal. Obviously, it is unwise to transmit the original information to the users directly. Original information involves large amount of redundant information and usually exists in an unfit form. The direct transmissions tend only to drown the user in the sea of the data. The experience leads the solution that, in the first level, interface should translate the original information into some fixed elements. These elements can make up all the useful information that we needed. Because one set of these elements belongs to one physical channel, we call them channel elements. The first level is called translator mostly for its function.

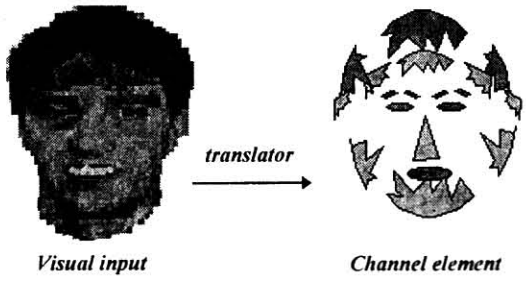


Fig. 2 channel elements of visual channel

Channel elements of the keyboard can be defined as the position of the key pressing. The visual channel elements are more complex that we assume it consisted of shape, color, and their special organization which keeps the information of position as in Fig. 2.

Channel elements are not enough to build a clear "mind" directly. The second level of the interface is called constructor which constructs channel elements into the models of the system following their own rule. The models may be notations cross many the physical channel. For example, character "A", "B", "C" is the notion cross keyboard inputting, visual and audio. The function of this level somewhat likes the function of the operation system. It organized the channel elements into notations while OS organized the instructions of the computer into basic processes and commands.

The last level of the perception model utilizes the notations from multiple models to complete the tasks of the system. Different systems may distribute this level different functions. It should provide basic rules for all the applications developed on the interface. It is more important that this level should build abstract notions and logical reason rules. It is called abstracter. Then, the advantage of multimodal interface is decided by how clever the abstracter is.

In summery, the whole process of the information flowing through the perception model can be described as followed. The physical signals were accepted by translator and translated into channel elements. And then these channel elements were built into notations by constructor. At last the notations from different constructors were complied together to support the system. Part of functions in abstracter can be off to the users. Abstracter is the workshop of the applications.

TRANSLATOR

The translator plays an important and basic roll in model. Many facts should be considered in designing a translator, for the system can get the information from nothing but the results of translator.

The first fact is that the channel elements should provide enough information to support the running of the system. From the results of the translator the system can fins whatever it wanted.

The second is that the same kind of physical channels should share the same set of channel elements. Usually, we assumed that each channel employs its own channel elements, such as the visual translator only product visual channel elements while the audio translator only product audio channel elements. This is important to separate the first and the second level of model in their functions.

Apart from offering enough information to system, the translator should insure that the form of the channel elements was suitable to be modulated and convenient to be used in the abstracter. This is the third fact that is worthy of noticing.

Other fact is that the translator must reduce the information as possible as it can. The more information it kept the more difficulties will be faced by constructor and abstracter.

The last fact we should not omit is that the translator must complete its work at a special time limit. There is no chance to take a long delay in this level.

Studying these facts carefully, it is easy to find some contradicts among them.

1. The information should be condensed as much as possible while the result of such condensing should remain enough information to satisfy the need of the system. The more accurate we pursuit, the larger mount of data we had to keep. Only the experiences come from large mount of this kind of work can tell us how to make a compromise.

2. One channel only employed one set of channel elements while the sets of channel elements must fit to every model depend upon this channel. For a visual channel, first, the visual channel elements are considered to be fitful in object recognition, human body recognition, character reading, scene describing, and so on. In each of these fields we have developed so many different methods and models. A corroding to each of these models and methods, it needs different kinds of suitable elements. It is really a hard work to find a uniform set of channel elements, but it is a significant one. If we can reach a content or make a proper standard to one channel of information, we catch an opportunity to take on a development on it. It is the way of standardization which means to build common language for one channel. Through the language, the base of the channel, we can develop a communication among application. We can rich our standard set little by little just as the same process as we rich our language by new words. The works of standardization for one channel need the effort of all the people in the related field.

3. Different channels choose different length of life for their channel elements. Keyboard may ask the last time element disappeared as soon as be picked. On the contrary, visual channel prefer a longer life for its elements. In one channel, the exist time of channel elements may be different with their kinds. In visual channel, shape elements keep a shorter time than color elements. To design a standard set of channel elements to every channel of information such as visual channel and audio channel are a hard work, but it is future reworded.

CONSTRUCTOR

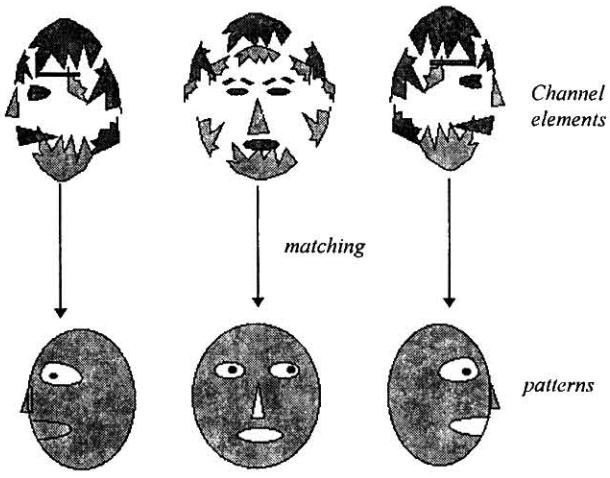


Fig. 3 the head model

Nietzsche had said that the intelligence of human being was reflected in the ability of differentiation. The second level of multimodal interface seems to follow the principle of differentiation. Because we find nothing familiar in channel elements, we need something can reduce the distance between channel elements and the understanding ability of computer. The notations produced by constructor always contained the exact meaning of our world. Constructors just act as a bridge to do this. That is to say all the channel elements were combined into the familiar items in the level. An example may help us to understand the designing of constructor. Let's assume that in our system the user may look forward to know whether the movement of the head of human being is nodding or wagging. According to this demanding, we designed middle constructors as follows:

First of all, the notation of the "head" is necessity. The channel elements gained from visual channel were used to seek the head in the head constructor. Shape visual channel elements that include the information of the position cooperated with color visual channel elements to portrait candidates of head. Fig, 3 describes the process to find the candidates. A rough matching pointed out the group of channel elements which involve 3 enough shapes with right position and right color as the candidates. More over, the candidates are transformed to match some patterns. There are some differences between these patterns and the common notation of patterns. These patterns only tend to present some gestalts of the head, instead of managing to impose the important characters of the head on the objects in the scene. Our aim is only to know how much degree the object looks like ahead and not to tell it is a head or not. In order to transform to pattern the object looks like a head and not to tell it is a head or not. In order to transform to pattern the constructor keeps special visual rule such as simple, continuous, and some other attributes. Because these matching can not offer the accurate result, a parameter named trust degree was proposed to describe about the matching. The more difficult to transform an object to match a pattern, the lower trust degree it got. Because this constructor only produces the notation of "head", its results automatically symbolize the head. In fact all we gain from this constructor is its Id number and the trust degrees.

The second model we need is the "face". Same processes rough matching and transformation exist in building the notation of "face" as outlined above. The notation of direction seems needed, but it is difficult to be constructed because it is an abstract notation. In this example the direction was simply defined as the transformation from a positive face pattern to the side face pattern. The definition just likes we image a face turning from the positive orientation to the side orientation in the scene. Naturally, this definition is not the real meaning of the direction. It exactly means the orientation of face. Usually, we consider the orientation of face is consist of the "face" and "direction". But it is hard to say such an idea is true or not. More complex notations must be built to perception nodding and wagging. The notations' sequence of head and orientation of face used as channel elements are put to repeat the process of rough matching and transformation again. The only difference is that the patterns are movement of nodding and wagging this time. From the example, we find the second level offer all its models and their trust degrees to the abstracter.

ABSTRACTER

Abstracter is a field for users to show their talents. The important fact we should notice that abstracter had to deal with the uncertain information. Of cause abstracter can builds notation for its own. The abstract notations, such as direction , fast, slow should be built there. The notation here, take on a changeable character. Some methods of channel may beautiful to support a abstracter, for example the methods of uncertain reasoning. The boundary between constructor and abstracter is not clear as well as the boundary between concrete and abstract is not clear, too. The roll of abstracter is very important, but there are seldom things should do in designing of multimodal

interface. Apart from build abstract notations, another important work for multimodal interface in this level is to put the same notations from deferent models together to identify that they are the same meaning.

CONCLUSION AND FUTURE WORK

we limit the work in building multimodal interface with the function of the perception model. The visual channel perception model we designed for multimodal interface have something like the pandemonium [5]. In concluding the whole model, function of each level can be defined formally as follow:

The first level is defined as (S, translator, T). here, S is the set of physical signals. T is the set of channel elements. Translator is defined as the mapping

$$translator{:}2^{S} \rightarrow T.$$

In our model, the second level is described as (T, constructor, M). T has the same meaning as above. M means the set of models which is defined as

$$M = N \times D.$$

N is the set of notations and D means the value of trust degree. Model constructor is the mapping

$$constructor{:}2^{T} \rightarrow M.$$

The basic part of abstracter can be defined as (M, B, U). M has the same meaning as before. U is the set of notations built in abstracter. B is also defined as a mapping

$$B{:}2^{M} \rightarrow U.$$

The running of the model seems to cost much time. In fact, its process is almost data drive and model feedback only takes place in case of dilemma. If there is nothing can help the abstracter to make up its mind, abstracter chooses one of these dilemma model results to feedback to the second level. Here, feedback means to change the cost of some certain kind of transformation. After the second level reported again, all these changed costs would recover to their original value. Consequently, another "channel" was resolute in abstracter.

Abstracter produced uncertain results. It's degree of this kind of uncertain hindered the character of the each abstracter. If you ask for the result of the head movement, it may answer you "I think may you are nodding" or "It seems nodding". The most interesting answer it produced may be "You say 'yes', don't you?". Then you got the a chance to change or stress your ideas. Our farther work will focus on building same notations from the same perception channel between human being and computer. Let computer understand the world which human familiar. In further future, we suggest fixed symbol to presented in certain meaning for each channel and build special language upon them. We hope to produce visual language based on movie film and audio language on something like combining of music and voice. It will not only give more abilities to the computer, but expand the abilities of human beings.

REFERENCES

[1] Y. Takebayashi, Y. Nagata and H. Kanazawa, "Noisy immunity keyword spotting with adaptive speech response cancellation" Proc. of IEEE ppII115-II119, 1993.

[2] K. Nagao and A. Takeuchi "Social Interaction: Multimodal Conversation with Social Agents" Proc. of 12th AAAI, 1992.

[3] H. Osmn, I. Katsunnobu, K. Takio, H. Satoru, T. Kazuyo, Y. Kazuhiko and O. Nobuyuki "Active Agent Oriented Multimodal Interface System" Proc. of IJCAI'95 Vol. 1 p82-p87, 1995.

[4] N. O. Bernsen, "Modality Theory in Support of Multimodal Interface Design" Proc. of AAAI, March 1994.

[5] O. G. Selfridge, "Pandemonium: A Paradigm for learning" Proc. of Symposium on the Mechanization of Thought Process: H. M. stationary office 1959.