User Registration Using Your Face and Mouth

Jie Yang, Fei Huang, William Kunz
Interactive System Lab
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{yang+, fhuang, kunz+}@cs.cmu.edu

Abstract

Registration is a common event in our daily life. Patients have to register before they can receive any treatment in a hospital. Participants have to register before they can attend a meeting. The objective of this research is to minimize a user's effort in the registration process. We propose a multimodal approach to user registration by combining face recognition, speech recognition and speech synthesis technologies together through an efficient dialogue manager. The multimodal user registration system consists of a face recognition module, a speech recognition module, a dialogue management module, and a speech synthesis module. When a person comes within range of the camera, the face recognition module tries to identify the user, and the speech synthesis module and speech recognition module interact with the user. The dialogue module integrates and manages the whole user registration process.

1 Introduction

Registration is the process of identifying people and associating certain personal information such as phone numbers, addresses, etc. with that person. User registration plays an important role in an intelligent meeting room environment. Participants' identities are essential for a multimedia meeting recorder to know who said what and so forth. Furthermore, knowing who is in the meeting in advance is helpful for enhancing speech recognition by taking advantage of more accurate speaker-dependent speech recognition algorithms. Moreover, some personal information (e.g., phone number, email address, etc.) is necessary to provide a method for contacting a participant after the meeting, which is often necessary in order to convey further information, follow-up on a participant's commitments, allow further discussion, and confirm the decisions of the meeting's participants. User registration, however, is a tedious and time-consuming task, especially in environments when large numbers of people must be registered rapidly and frequently, such as school or work registration. It is also time consuming to make a person register and re-register in order to attend different meetings. Therefore, it would be desirable if a user could register automatically or interactively with minimal efforts.

User registration is more than user identification and different from user authentication. In a user registration task, we assume that the user is cooperative. The task is to keep an updated copy of a user's information in a database. The database can be pre-constructed using information retrieval techniques. For example, the information of a user can be obtained from a pre-existing database, a personal home page, and a department directory. The database, however, can be incomplete and/or out-of-date. We need to complete, verify and update user's information in real-time. The task requires that we identify a user, and check if information in the database is complete and up-to-date. In this paper, we present a multimodal approach to interactive user registration by combining face recognition, speech recognition and speech synthesis technologies together through an efficient dialogue manager.

At Interactive Systems Lab in Carnegie Mellon University we have completed a working prototype of the multimodal user registration system. The system includes a face recognition module, a speech recognition module, a dialogue management module, and a speech synthesis module. Once a user comes within range of the system, the face recognition module tries to identify the user, and the speech synthesis module and speech recognition module interact with the user. The dialogue module controls the whole user registration process. During a registration process, a human face is used as the initial cue to help the system to identify the user. If the system can identify the user by his/her face and the user's personal information is up-to-date, the system will finish the registration process automatically, and the user need do nothing more. If the system can identify the user by his/her face but the user's information has not been updated recently, the system will retrieve the information from the database and verify it by confirming the information with the user through a human-computer dialogue process. If the system failed to identify a user's face, the system will perform registration interactively through a dialogue process. The dialogue management module plays a key role in minimizing the user's effort. We have developed a new dialogue management model based on a finite state automaton (FSA) [1], which uses a Bayesian network to fuse the user's information from multiple cues to reliably estimate the confidence about user identity. The FSA adjusts its weights dynamically based on available information from multiple sources. An optimal action at each succeeding state is determined by maximizing an objective function associated to the current confidence and information cues. Thus the transition between states can be done along the shortest path from the initial state to the goal state. In short, the system makes decisions about the user based on the best available information.

The organization of this paper is as follows: Section 2 presents the multimodal approach to user registration. Section 3 describes the architecture of the multimodal user registration system. Section 4 illustrates the working process of the system through an example. Section 5 details the conclusion of this method.

2 Problem Description

User registration is a tedious and time-consuming task, especially when it has to be done again and again. It is desirable that a user can register automatically. Automatic user registration requires a perfect user identification system and constantly updated database. The system could fail to identify a user because of incorrect information, and it is infeasible to maintain a database that contains every possible user. The database must be constantly updated or it will quickly become out of date, for instance, someone moves offices. Therefore, completely automatic user registration is infeasible in practice. In fact, there are five possible outcomes for an automatic user registration process:

- The system successfully identities a user and the user's information is updated (e.g., the user has registered recently).
- The system successfully identifies a user but the user's information may be out of the date.
- The system incorrectly identifies a user.
- The system fails to identify a user but the user's information is in the database.
- The user is unknown to the system.

Only for the first case, can the system finish the registration automatically. A completely automatic user registration system will fail in the rest cases. The major problem is uncertainty in both user identification and information retrieval. In order to handle uncertainties, we can take advantage of human intervention. The system can perform the registration interactively rather than automatically. Indeed, a user registration task is different from a user authentication task. In a user registration task, we can safely assume that a user is cooperative. The system can interact with a user if there is any uncertainty during a registration process as shown in Figure 1. In order to achieve such an interactive user registration scheme, it requires that a system be able to identify a user, to understand what a user is saying, to make queries and to control the interactive process. This can be achieved by combining face recognition, speech recognition, speech synthesis, and dialogue management technologies.

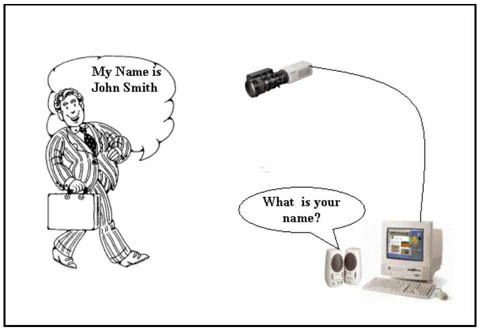

Figure 1. Interactive multimodal user registration

We hope to make a user's registration process as short as possible so as to minimize a user's effort. An optimal dialogue management strategy can lead to achieving such a goal. Much work has been directed towards the development of efficient dialogue managers for human computer interaction. Denecke and Waibel proposed the generation of clarification questions with domain modeling and underspecified representations to achieve a dialogue goal along an optimal sequences of questions [2]. Heeman et al. advocated "factoring out the grounding behavior" from the structured dialogue model [3]. Papineni et al. used a free-flow dialogue management model based on a form, which correspond to a specific task in the domain. The dialogue manager is mainly responsible for choosing the appropriate "form" which matches the user's goal best [4]. Ehrlich structured complex dialogs into sub- dialogs and thus reduced the dialogue's complexity at each state without losing its flexibility [5]. Johnston et al. applied dialogue management to multimodal integration based on unification over typed feature structure, which determines the consistency of two pieces of partial information, and combines them into a single result if they are consistent [6]. However, for multimodal user registration, the consistency between multiple pieces of partial information is already known (they are all from the same user). We are more concerned with the confidence of the user's identity given the information.

We would like to employ an optimization method. It has been demonstrated by other researchers that a dialogue management process can be formulated as an optimization problem ([7, 8, 9]). Under some assumptions about the state transition probabilities and cost assignment, a dialogue system can be considered as a Markov Decision Process (MDP). With such a framework, the supervised and reinforcement learning algorithms are applied to learn the optimal strategy. An optimal strategy is a mapping from a state to an action, i.e., a policy determines which action should be taken in every possible state. Once it is learned, it is fixed and deterministic. This is not necessarily a good approach for multimodal integration, because information cues from multiple modalities are dynamically available, and switches between different modalities are quite frequent. When and which modality should be used will depend upon the availability of the various information sources.

In order to handle such uncertainty, we designed a dialogue manager based on a finite state automaton. Similar to the MDP model, we define the states, transitions and action set. Unlike a traditional MDP model, weights of the FSA model are not fixed. We use a Bayesian network to determine the confidence with respect to a user's identity by fusing current information cues from multiple channels (e.g., face image, spoken language input and database). Multimodal information cues are integrated incrementally. The weights are adapted based on an evaluation function, which indicates the confidence score, completeness of available information and human-computer interaction cost at current state. By maximizing this function during each dialogue turn, the optimal strategy is determined runtime rather than learned in advance, and the shortest path from initial state to goal state can be dynamically determined. This will achieve the goal of minimizing a user's effort.

3 Multimodal User Registration

We have implemented a multimodal user registration system. The system architecture is shown in Figure 2. The system consists of a face recognition module, a speaker-independent large vocabulary speech recognition module, a text-to-speech synthesis module, and a dialogue management module. The system works as follows. During a registration process, a human face is used as the initial cue to identify a user. Once a user appears within the scene of the camera, the face recognition module tries to identify the user. If the system can reliably identify the user by the face and is certain about user's information, the system will finish the registration process automatically. If the system has any uncertainty in either the identification or the information retrieved, the system will start an interactive registration process. The speech synthesis module and speech recognition module interact with the user. The dialogue module optimally controls the whole user registration process. The remainder of this section discusses each module in more detail.

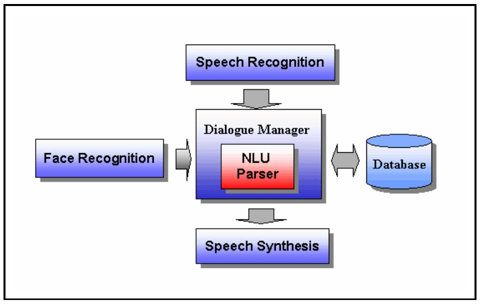

Figure 2. System architecture: communications between modules are through sockets

3.1 Face Recognition Module

The face recognition module consists of two parts: a real-time face tracker and a face recognition system. These two parts are connected through a socket. The reason behind the decision to separate face tracking from face recognition is to allow the two units to work at different rates. It is essential that the tracking unit continually updates the location of a person face for coherency reasons. This requires the tracking unit to work at a speed of many frames per second. Face recognition, on the other hand, trades off accuracy for speed. At many frames per second, accuracy of recognition is too degraded to be useful. However, it is not necessary that recognition occurs near as frequently as tracking; therefore, we have separated the units to allow the tracking unit and the recognition unit to work at speeds optimal for their purpose. This has the added advantage of allowing one machine to perform the tracking and image acquisition, while another computer performs recognition on the face, which means registration can be performed remotely while recognition can be centralized.

Locating and tracking human faces is a prerequisite for face recognition. By combining the adaptive skin color model with the motion model and the camera model, we have developed a real-time face tracker [10]. The system has achieved a rate of 30+ frames/second on both Unix and

PC platforms. The system can track a person's face while the person walks, jumps, sits and rises. Figure 3 shows an example of the result by the real-time face tracker.

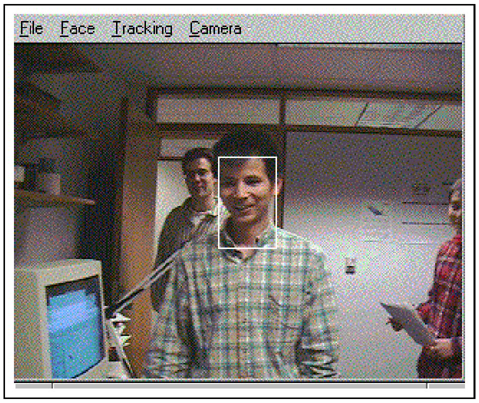

Figure 3. An example of the real-time face tracker

The face recognition part is a modular system, which can easily plug into various face recognition algorithms. We have currently implemented an "eigenface" algorithm, a dynamic space warping (DSW) algorithm [11], and an LDA algorithm [12]. The techniques based on Principal Components Analysis (PCA), namely "eigenfaces" [13], have demonstrated excellent performance. In the eigenface approach, a face image defines a point in a high dimensional space. Different face images share a number of similarities with each other, so that the points representing these images are not randomly distributed in the image space. They all fall into a lower dimensional subspace. The key idea of the recognition process is to map the face images into an appropriately chosen subspace and perform classification by distance computation. Instead of transforming a face image into one point in the eigenspace, the DSW algorithm breaks down a face image into sub-images using a moving window. When the square window covers the whole image by moving half of the window size each time, we get a sequence of sub-images. Each sub-image can be transformed to a point in the eigen-space. We then get a set of eigen-points for each face image. During the recognition process, the template set of points is compared to the unknown set of points. The DSW algorithm has better performance than the eigenface algorithm but it considerably is slower. Unlike the PCA which encodes information in an orthogonal linear space, the LDA encodes discriminatory information in a linear separable space of which bases are not necessarily orthogonal. Researchers have demonstrated that the LDA based algorithms outperform the PCA algorithm for many different tasks. However, the standard LDA algorithm has difficulty processing high dimensional image data. The PCA is often used for projecting an image into a lower dimensional space or so-called face space, and the LDA is then performed to maximize the discriminatory power. We have developed an efficient LDA algorithm to maximize the LDA criterion directly without a separate PCA step [12] .

3.2 Speech Recognition Module

The speech recognition module utilizes XCalibur, a spoken language R&D ToolKit developed at Interactive Systems Inc. [14], as its core engine. It is designed to be compatible with the Java Speech API. The XCalibur supports large vocabulary continuous speech recognition with a very high accuracy. In order to further increase recognition rate, we have written grammars for user registration. Like many task specific speech recognition applications, user registration benefits from highly predictable conversation patterns. In fact, the topics of conversation can quickly be classified into 4 different classes: greeting, determining the need to update information, exchanging information, and confirming new information. The process of exchanging information in our user registration system was further categorized into: name presentation, phone number presentation and email presentation. We write a grammar for each category. In our initial test, the speech recognition accuracy is 95% using such grammars.

3.3 Speech Synthesis Module

We use Festival as the speech synthesis module. Festival is a general multi-lingual speech synthesis system developed at Center for Speech Technology Research, University of Edinburgh. The Festival is a full text-to-speech (TTS) system with various APIs, and an environment for development and research of speech synthesis techniques. For more detailed information, see [15].

However, synthesizing the pronunciations of many foreign names, which have different pronunciation rules, is a rather tough task. An alternative solution is to record the user's pronunciation when he/she answers the question "What is your first/last name?". This is a part of our future work.

3.4 Dialogue Manager Module

The dialogue manager module controls the whole registration process. The dialogue manager module works as follows: it first obtains the hypothesis on a user identity from the face recognition module; then searches the related information in a pre-constructed database; determines the user identity based on the confidences from different information cues (e.g., face, name, etc.); further acquires and/or confirms

	First name	Last name	Phone number	Email address
State 1	Empty	Empty	Empty	Empty
State 2	Filled	Empty	Empty	Empty
State 3	Filled	Filled	Empty	Empty
State 4	Filled	Filled	Filled	Empty
State 5	Filled	Filled	Empty	Filled
State 6	Filled	Filled	Filled	Filled
State 7	Filled& Verified	Filled& Verified	Filled& Verified	Filled& Verified

Table 1. Definitions of different states in the FSA.

user's personal information via speech recognition module and text-to-speech synthesis module; and finally updates the database if information is changed.

The dialogue manager module employs a structure of finite state automaton (FSA) as shown in Figure 4, which contains 7 states, each corresponding to the frame filled with different information. The definitions of different states are listed in Table definition. We define a frame containing 4 slots (First_name: Last_name: Phone_number: Email_account:) as the format of the required information. The registration can be considered as a slot-filling process. It transits from the initial state to the goal state. To minimize user's effort, the dialogue manager should take the optimal action at different states so that the transitions occur along the shortest path.

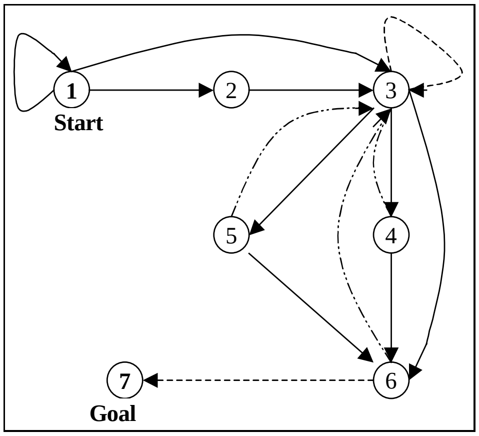

Figure 4. Structure of the FSA for the dialogue manager

The dialogue manager combines all the available information through a Bayesian Network, to reliably estimate the user's identity. By continuously maximizing an objective function at each state, i.e., dynamically adjusting the weights of the FSA, the optimal action can be selected from an action set (e.g., performing face recognition, querying the specified information, retrieving information from a database, etc.).

Since the speech recognition module has very high accuracy and the language- understanding task in a user registration process is not difficult, we have embedded the natural language understanding part into the dialogue manager for keyword spotting and parsing. The recognized speech is parsed so that user's information can be extracted with a correct format.referenced and extracted quickly.

4 An Application Example

We have tested the multimodal user registration system under our meeting room scenario. The user database is automatically constructed using information retrieval technique through searching web pages, particularly the directory of the School of Computer Science in the Carnegie Mellon University, which contains 1744 users' names, phone numbers and email addresses (however phone numbers and e-mail addresses are not always provided). To help readers to better understand a registration process using our system, we have listed a sample dialogue below:

Agent: Welcome to the Multimodal User Registration System!

(System takes Face Recognition, current confidence score is not high enough for verification, so system asks for more information)

Agent: Could you tell me your first name?

User: My first name is F-e-i.

(System compute the confidence based on face recognition and recognized first name, check the database, and have higher confidence on user's identity, so ask for verification.)

Agent: Hi, you are Fei Huang, right?

User: Yes, I am.

(System retrieves database for personal information.)

Agent: Hi, Fei, your phone number is two six eight five two nine nine, your email account is fhuang@cs.cmu.edu. Do you want to update your personal information record?

User: Yes.

Agent: Which part do you want to change, phone number or email account?

User: I will update my phone number.

Agent: What is your new phone number?

User: my new phone number is two six eight one four seven eight.

Agent: You information has been updated. Do you want any other change?

User: No. Thank you.

Agent: The registration is done, Thank you for using this system. Goodbye!

5 Conclusion

We have described a multimodal approach for user registration. We believe that an interactive approach is an efficient way to handle uncertainties in both user identification and information retrieval. By optimally controlling the interactive process, it is possible to minimize a user's effort. We have employed an adaptive FSA model to represent the dialogue process. The parameters of the model are dynamically adjusted based on the confidence score, completeness of available information, and human-computer interaction cost at the current state. We use a Bayesian network to determine the confidence with respect to a user's identity by fusing current information cues from multiple channels (e.g., face image, spoken language input and database information). We have developed a multimodal user registration system by combining face recognition, speech recognition and speech synthesis technologies together through an efficient dialogue manager.

ACKNOWLEDGEMENT

The authors would like to thank Robert Malkin, Matthias Denecke and other members in the Interactive Systems Labs for their inspiring discussions and support. This research is partly supported by Defense Advanced Research Projects Agency under contract number DAAD17-99-C-0061.

References

[1] F. Huang, J. Yang, A. Waibel, "Dialogue Management for Multimodal User Registration," *Proceedings of the International Conference on Spoken Language Processing (ICSLP'2000),* Beijing, China, October, 2000.

[2] M. Denecke and A. Waibel, "Dialogue Strategies Guiding Users To Their Communicative Goals," *Proceedings of EUROSPEECH'97*, Vol. 3, pp. 1339-1342, Rhodes, Greece, September 1997.

[3] P. A. Heeman, M. Johnston, J. Denney and E. Kaiser, "Beyond Structured Dialogues: Factoring Out Grounding," *Proceedings of the International Conference on Spoken Language Processing (ICSLP-98),* pp. 933-936, Sydney, Australia, November.

[4] K. A. Papineni, S. Roukos, and R. T. Ward. "Free-flow Dialog Management Using Forms," P*Proceedings of EUROSPEECH'99,* Vol. 3, pp. 1411-1414, Budapest, Hungary, September 1999.

[5] U. Ehrlich, "Task Hierarchies Representing Sub-Dialogs in Speech Dialog Systems," *Proceedings of EUROSPEECH'99,* Vol. 3, pp. 1387-1390, Budapest, Hungary, September 1999.

[6] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith, "Unification-based multimodal integration," *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics'97,* Association for Computational Linguistics Press: March, 1997.

[7] E. Levin, R. Pieraccini, and W. Eckert, "Using Markov Decision Process for Learning Dialogue Strategies," *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP-98),* Vol. 1, pp. 201-204, Seattle, U.S., May 1998.

[8] E. Levin, R. Pieraccini, and W. Eckert, "A Stochastic Model of Human-Machine Interaction for Learning Dialog Strategies," *IEEE Trans. on Speech and Audio Processing,* Vol. 8, No. 1, pp11-23, January, 2000.

[9] S. Singh, M. S. Kearns, D. J. Litman, and M. A. Walker, "Reinforcement Learning for Spoken Dialogue Systems," *Proceedings of Neural Information Processing System (NIPS-99),* Denver, U.S., November 1999.

[10] J. Yang and A. Waibel, "A Real-time Face Tracker," *Proceedings of Third IEEE Workshop on Applications of Computer Vision (WACV-96),* pp. 142-147, Sarasota, Florida, USA, December 1996.

[11] R. Gross, J. Yang, and A. Waibel, "Face Recognition in a Meeting Room," *Proceedings of Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'2000),* Grenoble, France, March, 2000.

[12] J. Yang, Y. Yu, W. Kunz, "An Efficient LDA Algorithm for Face Recognition," The Sixth International Conference on Control, Automation, Robotics and Vision (ICARCV2000).

[13] M.A. Turk and A. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* pp. 586-591, Hawaii, USA, 1991.

[14] M. Finke, J. Fritsch, D. Koll and A. Waibel, "Modeling and Efficient Decoding of Large Vocabulary Conversational Speech," *Proceedings of the EUROSPEECH'99,* Vol. 1, pp. 467-470, Budapest, Hungary, September 1999.

[15] *http://www.cstr.ed.ac.uk/projects/festival/.*