CONNECTIONIST MODELS IN MULTIMODAL HUMAN-COMPUTER INTERACTION

Alex Waibel  
Paul Duchnowski
Carnegie Mellon University, Pittsburgh PA, USA
University of Karlsruhe, Karlsruhe, Germany

Abstract

We present an overview of our laboratories’ research on multimodal Human-Computer Interfaces. By exploiting all available channels of human communication we aim to increase flexibility, robustness, and naturalness of human-computer interaction. The information sources we process include Speech-, Character-, and Gesture Recognition, Face- and Eye Tracking, Lipreading, and Sound Source Localization. Connectionist and hybrid techniques are used throughout.

Introduction

Recent developments in the computer and communication industries are rapidly increasing the amount and variety of information available to a wide and diverse audience. The multimedia nature of this data explosion, heralded by the advent of the “Information Superhighway”, offers images, text, etc. as output presented to the information consumer. This is in stark contrast to the information available only to a small subset of society. Attempts at the use of alternate communication methods have focused on single alternatives and limited acceptance.

To improve this situation, we have begun to process a multiplicity of signals that are carry meaning in human communication. Understanding written character-, lipreading, face-tracking, eye-localization. In combination, information is known to provide redundant information for greater efficiency and freedom to choose the signal best suited for human interaction. Such signals form a rich language in human-computer interaction, such as English, German, French, etc.
recognition and is preferred to a static, bitmapped representation of gesture's shape. The coordinates are normalized and resampled at regular intervals to eliminate differences in size and drawing speed; from these resampled coordinates we extract local geometric information at each point, such as the direction of pen movement and the curvature of the ectory.

Each coordinate is represented in the classifying NN by high-level features. Their temporal sequence contains the input layer. Ten units in the first hidden layer patterns form the input, eight units in the second layer spot patterns typical of a given gesture. Out (one per gesture) integrate over time the evidence corresponding unit in the second hidden layer. The highest activation level determines the class of one gesture. The network is trained on a set of manually labeled examples using a modified backpropagation algo-

Also incorporates a method for active “on the fly”, i.e., while the systems still operates, the system updates the network's output and creates new templates that project onto the output units. If the input is similar to the template used for training then the network projects it onto the active units. This technique is called an “exhausted network”. The system's performance is, of course, the accuracy of the output to try to supplement the system's own a teach

Also, it operates on a teach

Additional options

The system can also be trained using backpropagation,
There is a system of networked cameras that track the three-dimensional shape of the objects in a scene and enhance depth and localization of the virtual camera, indicating the individual's spatial position along the face. Appropriate commands are given to the virtual camera by the user if the face moves out of the field of view. A model of the face is then used to calibrate the physical camera. Finally, the camera captures the user's face and the area classified as the face is cropped. This approach to providing clean sources of facial features is efficient in front of the camera. The edge is often used for signal acquisition and extracting features.

![Image 2](camera.png)

**Figure 2.** Camera image and extracted largest skin-colored object.

Two neural networks are used for centering and size estimation, respectively. They were trained by backpropagation on 5000 artificially scaled and shifted example images generated with a database containing 72 images of 24 faces of different sex, age, hair style, skin color, etc. Performance was evaluated on test sequences of 20000 images of 7 persons (with different skin types) performing arbitrary movements in front of different backgrounds. Depending on the sequence, the face was located in 99% to 100% of all images in the sequence. The average difference of the actual position of the face and the output of the system were less than 10% of the size of the head.

**Eye Tracking**

The goal of eye tracking is to determine where a person is looking from the appearance of his eye. Two potential uses of a gaze tracker are as an alternative to the mouse as an input modality and as an analysis tool for human-computer interaction studies. The direction of eye fixation can be used to determine the user's focus of attention in a visual interface; for instance, knowing whether the user is looking at the screen or somewhere else while talking is important in deciding whether automated speech should be activated.

We have developed a neural-network gaze tracker based on camera input. The system incorporates a gaze tracker, a gaze-based gaze tracking, the user is looking at the screen, not to keep the system from falling behind the user. An infrared light is used on the eye. The gaze relative positions of the system extracts on the eye. The gray-scale input is to a neural network that performs an analysis.
tic communication situations including competing speakers.

Considering visual aspects to locate the speaker's position overcomes these limitations. Specifically, the face-tracker supplies the coordinates of a moving speaker to the microphone array which then form a beam to that location. Our experiments have confirmed this synergy, demonstrating improved signal-to-noise ratio even for speakers moving in an environment with another loud sound source.

Another application of fast tracking and beamforming is to enhance the lip-reading/speech recognition system. The face-tracker allows for a noninvasive acquisition of the visual data, which enhances the quality of the received audio signal. Work on complete systems is already in progress.

Figure 3. Basic recognition network architecture (integration at a single, flexible, and robust human-computer interface).

Acknowledgments

The authors gratefully acknowledge the support of the Defense Advanced Research Projects Agency, Microtechnology Office who made this research possible.

Dr. Wu Bih, Martin Bum, Stefan Bum, Wezel Brine, Ivica Brine, and their work on each of the modules provided assistance in preparation of the manuscript.

References


We have tested the recognizer on data sets of 200 letter sequences from single speakers. On average, LDA preprocessed visual input produces best results, reducing the audio-alone error rate by 33.7%.

Speech and Gesture Recognition

We have developed a speech- and gesture-based text editor as another step towards modality integration. The word spotter (see above) was trained to spot 11 keywords representing editing commands such as move, delete,... and 14 textual units such as character, word,... The effect is to let the user speak naturally without having to worry about grammar and vocabulary, as long as the utterance contains the relevant keywords. For example, an utterance such as "Please delete this word for me" is equivalent to "Delete word".

We based the interpretation of multimodal inputs on frames consisting of slots representing parts of an interpretation. The speech and gesture recognizers produce partial hypotheses in the form of partially filled frames. The output of the interpreter is obtained by unifying the information contained in the partial frames. For example, a user draws a circle and says "Please delete this word". The gesture-processing subsystem recognizes the circle and fills in the command scope (what to operate on) specified by the circle in the gesture frame. The word spotter produces "Delete word", from which the parser fills in the action and textual unit slot in the speech frame. The frame merger then outputs a unified frame indicating that the operation delete is to be carried out on the word specified by the scope of the circle.

One important advantage of this frame-based approach is its flexibility, which will facilitate the integration of more than two modalities. All we have to do is define a general plan for interpretation and specify the ways in which slots are filled by each input modality. In a general implementation, it is possible that the slots may be filled in different and performing a search to find the best merge would be necessary.

Modeling and Transforming

As described earlier, it is the goal of the system to track a moving talker in reall...