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.

e speech recognition are un-
h several connectionist
Large Vocabulary Con-
eous speech recog-
ceptrons (MLP),
g Vector

Introduction

Recent developments in the computer and communication
industries are rapidly increasing the amount and variety of
nformation available to a wide and diverse audience. The
ti-media nature of this data explosion, heralded by the
pt of the "Information Superhighway", offers images,
ext, etc. as the output presented to the informa-
sumer. This is in stark contrast to the impover-
of input options which are still largely limited to
ard and mouse. Attempts at the use of alternate
ave mostly focused on single alternatives and
ited 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-
ocalization. In combination,
rmation are known to pro-
l information for effec-
hey allow for greater
lundant information
ty and freedom to
n channel. Such
ul in human-
g, speech
such as
ery

h-
n

recognition[7] 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
jectory.

Each coordinate is represented in the classifying TDNN by
such low-level features. Their temporal sequence con-
s the input layer. Ten units in the first hidden layer
patterns from the input, eight units in the second
ayer spot patterns typical of a given gesture. Out-
(one per gesture) integrate over time the evidence
rresponding unit in the second hidden layer. The
th the highest activation level determines the
The network is trained on a set of manually
es using a modified backpropagation algo-
ng data of 80 samples/gesture, we have
ependent recognition rate of 98.8% on

also incorporates a method for ac-
s "on the fly", i.e., while the system is
nition error occurs, the system queries
ect output and creates new template-
at project onto the output units. If
n is similar to the template used
is turned on and can influence
t. This technique is called an

r handwriting
ementariness to the acous-
able for distinguishing the
conveys most robustly man-
systems' performance is,
egradation of the acous-
ral to try to supplement
nformation
ly scientific, but also
for notepad comput-
ative input options
ng of on-line
temporal infor-
ented above.

on a touch
and grabbing the
80 frames/sec,
, speakers are
lips appear
busly shown
markers,
ordinarise-
e with a
sure data
blem
cted
ks-
or

e a system of networks. The network considers shape of the objects in pro-
and enhance speech ... the ... of the *virtual camera*, indicating the
s individual regions ... by controlling the face. Appropriate commands
ng to carry ... such ... are issued if the face moves out of
tc. a pre-defined area in the center of the physical camera. Fig-
ure 2 shows an example of an image and the area classified
as face by the tracking system.
pproach to providing clean
source.
onal microphone array
plane in front of the
rds a given spot the
the microphones
y from this lo-
signals, one
Competing
m other
pro-

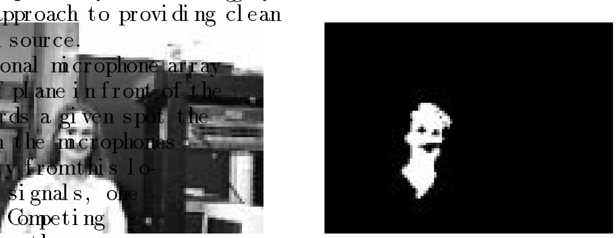

Figure 2. Camera image and extracted largest skin-colored object.

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 over 2000 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 96% 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 gaze tracking is to determine where a person is looking from the appearance of his eye. Two potential es of a gaze tracker are as an alternative to the mouse an input modality and as an analysis tool for human-er interaction studies. The direction of eye fixation be used to determine the user's focus of attention in odal interface; for instance, knowing whether the king at the screen or somewhere else while talking rtant in deciding whether automated speech d be activated.

l on we have developed a neural-network-gaze tracker based on camera input lvanced gaze tracking, the user is ny special equipment, nor to keep he system comes from a camera monitor. An infrared light on on the eye. The gaze relative positions of The system extracts on. The gray-scale put to a neural ut units for rformed by

y we

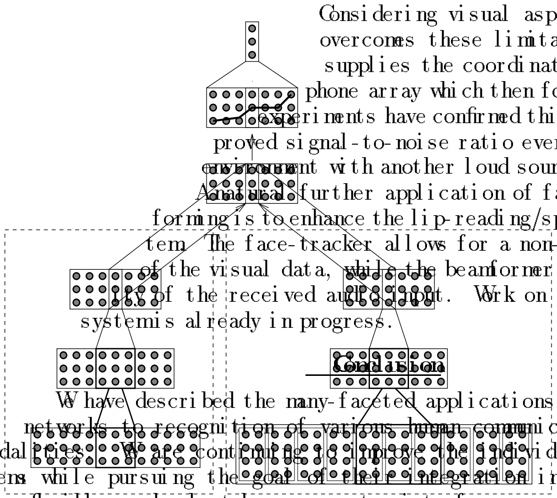

Figure 3. Basic recognition network architecture (integration at the phoneme/viseme level).

We have tested the recognizer on data sets of 200 letter sequences from single speakers. On the 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 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 ts flexibility, which will facilitate the integration of more than two modalities. All we have to do is define a general ame for interpretation and specify the ways in which slots be filled by each input modality. In a general implemen-, it is possible that the slots may be filled in different and performing a search to find the best merge would r.

ing and Beamforming

cribed earlier picks its target as the s vicinity. It, therefore, encounters probting to track a moving talker in realistic 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 forms 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.

An additional further application of face-tracking and beamforming is to enhance the lip-reading/speech recognition system. The face-tracker allows for a non-invasive acquisition of the visual data, while the beamformer improves the quality of the received audio input. Work on such a complete system is already in progress.

Conclusion

We have described the many-faceted applications of neural networks to recognition of various human communication modalities. We are continuing to improve the individual systems while pursuing the goal of their integration in a single, flexible, and robust human-computer interface.

Acknowledgements

The authors gratefully acknowledge the support of the Ad-nced Research Projects Agency, Microtechnology Office le this research possible.

s to Udo Bub, Martin Hunke, Stefan Manke, Uwe rleau, Ivica Rogina, Minh Tue Vo, and r their work on each of the modules arious assistance in preparation of

eferences

au (1993). Non-Intrusive Gaze al Networks, in *Neural In-* (NIPS-6), Morgan Kauf-

ibel (1994). See ech Recogni- *CP 94*.

. Inte- gh