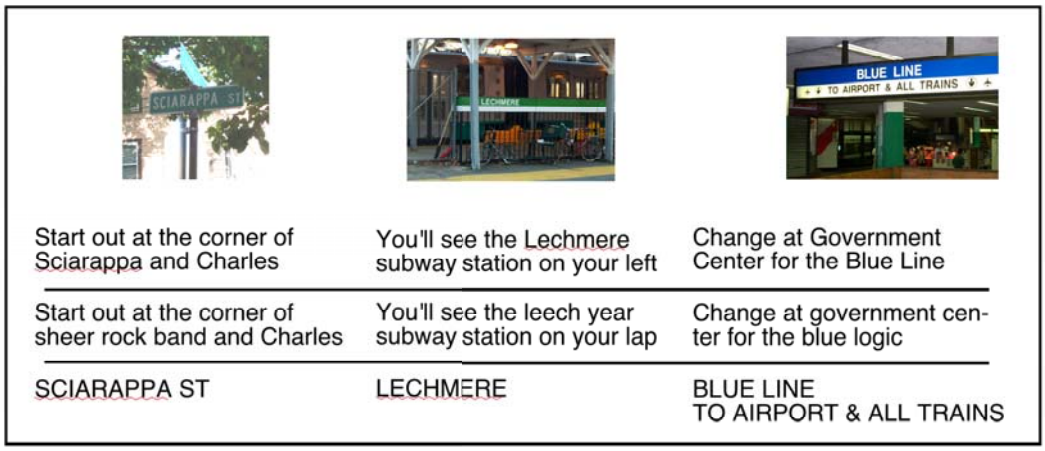
Out of Many, One:
Reliable Results from Unreliable Recognition

Henry Lieberman
Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
+1 617 253 0315
lieber@media.mit.edu



ABSTRACT

Recognition technologies such as speech recognition and optical recognition are still, by themselves. not reliable enough for many practical uses in user interfaces However, by combining input from several sources, each of which may be unreliable by itself, and with knowledge of a specific task and context that the user is engaged in, we might achieve enough recognition to provide useful results. We describe a preliminary experiment to assist the user in giving directions for urban navigation by combining partial results from unreliable speech recognition and unreliable visual recognition.

Keywords

Speech recognition, optical character recognition, navigation, heuristics, intelligent agents.

INTRODUCTION

User interface designers often shy away from designing interfaces that try to solve the "hard problems" such as voice recognition, natural language understanding, and visual recognition.

While it is true that each of these technologies has not yet, individually, reached the stage of full practical reliability, we may have gotten *too* accustomed to avoiding them. Each of these technologies has made substantial progress in the last few years, both from better algorithmic advances and from faster processing speeds. It may now be time to reassess their practicality in user interfaces.

Further, combining the use of several of these recognition procedures might be better than using each independently, since many situations contain redundant information. In addition to extracting information from the raw perceptual data, we can also take account of the user's *context*, both of the task context and the individual situation and preferences of the user [1]. Context information can greatly restrict the plausible interpretations of the data, and provide the means to correct possible errors of interpretation.

AN ASSISTANT FOR URBAN NAVIGATION

We are exploring these issues in a scenario of an assistant for giving (and, eventually, taking) directions for urban navigation. It is typical, when issuing an invitation to an event, to include directions that will enable the invitee to find their way to the event's location. Services such as MapBlast can automatically generate street maps given a starting point and destination, but these have their limitations. Based only on street connectivity, they can't include visible landmarks (Turn right at the big oak tree), or deal with multi-modal routes (partly walking, partly public transit) or exceptional circumstances.

Instead, we imagine that the user will have a PDA and digital camera, and *demonstrate* the route that they would like their guests to follow, actually traveling between the starting point and the destination. Along the way, the directions-giver would take pictures of each important step (usually turns or other decision points), and annotate the pictures with verbal narration of directions. For example, the user passes Sonya's Cafe, takes a picture of the cafe sign, and says into the audio recorder "You'll see Sonya's Cafe on your left". The challenge, then, is to make a system that collects the audio and image data, relates it to any other data sources that might be available (such as GPS), and compiles a set of coherent, readable directions to be transmitted to the guest. A further step would be to transmit the directions to a guest's computing device in electronic form, for use by a direction-following assistant that could provide help of the form "Is this is street where I'm supposed to make a left?".

Right now, the tasks of downloading the images, correlating the narration with the images, and presenting the information in readable form present a daunting barrier to effectively using the easily captured data.

IS PRESENT SPEECH AND VISUAL RECOGNITION UP TO THE TASK?

To assess the feasibility of such an application, we performed an experiment to collect data for some routes. The routes involved a combination of walking and public transportation (taking pictures from a car would require the passenger to perform the data entry). We did two things to interpret the resulting data.

First, we ran the narration through a commercial speech recognizer, IBM Via Voice, a large vocabulary speaker-dependent speech recognizer. It was trained with only a minimal training session before the data collection began.

Second, we ran the images through a commercial Optical Character Recognition program, Caere Omnipage 8.0., in order to recognize the written information in street signs.

It is worth noting that neither program was intended for this sort of use. ViaVoice is for dictated documents in an office setting; Omnipage is intended for the interpretation of text from scanned paper documents. We had, for example, to process the images to make them acceptable to Omnipage, simulating paper resolution, and inverting them to enable the recognition of light-on-dark sign text.

RESULTS

As might be expected, the reliability of recognition was pretty poor -- if transcription accuracy is the criterion. The speech recognizer often got proper names such as street names wrong. This could be improved by preloading the speech recognizer with a dictionary of street names or subway stop names for the city in question.

The OCR often missed information on signs, especially where the sign was tilted with respect to the plane of the camera, (as in cross streets on corners). Some of this could be alleviated by more careful image capture. We were also able to "fix in the mix" some of these problems simply by rotating the images a reasonable amount in both directions and trying again.

The table above shows, respectively, some images, the actual narration, its interpretation by the speech recognizer, and the results of the OCR. Sometimes the results were nothing short of amusing. What was remarkable, though, was how often essential information did come through.

In the numbers below, we present the raw word-level accuracy rates of the voice recognition and the *landmark accuracy* – the percentage of "landmark" names and phrases necessary for the directions that were actually recognized. The *combined landmark accuracy* is the percentage of landmarks recognized by *either* the voice *or* OCR.

Voice reco word-level accuracy:	**83%**
Voice landmark accuracy:	**62%**
Combined landmark accuracy:	**75%**

First of all, note that the landmark accuracy is worse than the word-level accuracy. This is because landmarks tend to be proper names easily missed by the voice recognition This could be considerably improved if a street directory were used to train the voice recognizer. Second, it is encouraging that the combined landmark accuracy is significantly higher than the voice landmark accuracy. This is because the landmarks most likely to be gotten wrong by the voice reco are often those that would be likely to appear on street signs. Finally, it is not necessary to achieve 100% landmark accuracy to make a useful system, because the user can be queried about possible ambiguities, and a reasonably high landmark recognition rate would still result in saving labor for the user.

RELATED WORK

This is most related to the field of multi-modal input, of which [2] is a good modern survey. The urban navigation assistant application, and use of multi-modal input in such an application by nonprofessional users, we believe to be unique.

ACKNOWLEDGMENTS

This investigation is part of a project, Information Pipette, to develop a PDA-based "informal data capture" appliance, in collaboration with Nick Kushmerick of University College, Dublin.

REFERENCES

1. Henry Lieberman and Ted Selker, Out of Context: Computer Systems that Learn About, and Adapt to, Context, IBM Systems Journal, Vol 39, Nos 3&4, pp. 617-631, 2000
2. Sharon Oviatt and Philip Cohen, Perceptual user interfaces: multimodal interfaces that process what comes naturally, CACM, 43(3), March 2000, p.45-53