Computer vision, eyetracking, spoken dialog systems, and evaluation: Challenges and opportunities

Gregory Aist[*,1,2]

[1]Computer Science Department, University of Rochester, USA
[2]Institute for Human and Machine Cognition, Pensacola, Florida, USA
*Contact address: gsa@gregoryaist.com

Abstract

Spoken dialog systems have matured to the point where the underlying technologies of speech recognition, natural language understanding, dialog management, natural language generation, and speech synthesis are available for many languages and are portable to various domains. However, when people communicate, spoken language is accompanied by a host of other inputs and outputs. The most immediately salient of these processes is input and output perceived visually: gestures, eye gaze, posture, and the like. We focus on three kinds of visual input: the first can be roughly called visual awareness, and includes factors such as what objects (including people) are in the visual scene, how far away they are from the computer, and so forth. Certain types of visual input such as posture, head pose, blink rate, and pupil dilation are more fine-grained and have to do with analyzing an individual's cognitive and affective states; we'll call this cluster personal perception. The third is eyetracking: where the user is looking. Here we spell out a number of challenges posed by the introduction of the visual channel into spoken dialog systems, and enumerate some opportunities that vissual input provides. We focus especially on how these challenges relate to evaluation – since evaluation of a spoken dialog system's behavior is a central aspect of dialog system research and is especially critical for important applications such as intelligent tutoring systems.

1. Architectural Requirements for Adding Visual Input to a Dialog System

We can represent the standard model of dialog systems as:
1. Listen to the user talk.
2. When the user is done speaking,
 decide what to say (or do) in response.
3. Say it.
4. When the system is done speaking, go to step 1.

Advances such as incremental understanding and learning/adaptation require various modifications to this model, resulting in a more event-driven structure; adding vision as well yields the following:
1. Watch and listen to the user talk
2. While watching and listening,
 decide what to say or do in response.
3. If there's a response, say it
 while continuing to watch and listen.

In a sense, the use of visual input requires a shift to an incremental understanding framework (with an event-driven model) as a technological precondition: the system has to take an active role in thinking about what to do next on a continuous basis, rather than just waiting for the end of the user's turn. Fortunately, a number of dialog systems including task-oriented systems such as TRIPS (Allen et al. 2005) and tutoring-oriented systems such as the Reading Tutor have made this transition (Aist 1998).

Using visual input presents a number of additional challenges beyoind the requirement for an event-driven model with incremental understanding

1.1. Ensuring High-Quality Sensory Input

Spoken dialog systems, as with any interactive system, perform better with high-quality input. Various efforts have been made to pre-process noisy audio to either yield higher quality audio, or to modify the basic understanding process to be robust to the precense of noise. In an interactive setting, other techniques are available as well: direct feedback to the user ("Too loud! Please don't shout."), playing back a sample for the user to self-analyze (as the Reading Tutor does at times.) With video input, system designers face similar challenges; as anyone who has tried to take home video knows, the intensity and direction of available lighting plays a huge role in the quality of the resulting video. Beyond issues of lighting, for computer perception, having a somewhat predictable background would be likely to have a substantial effect on the usefulness of the resulting video. For example, a physical setup for a video-enabled tutoring system where the entire classroom was behind the student would present additional challenges compared to a setup where a static screen of some kind was in the background instead.

1.2. Turning Sensations into Perceptions

For example, unlike speech input which can be either present or absent (i.e. the system hears silence), visual input is always available unless the system is in total darkness. In a real-world setting, some sort of structure must be imposed or inferred upon the sequence of visual input in order to determine such basic facts such as (a) is anyone there? (b) how many people are present? (c) who are they? (d) are they engaged with the computer system, and so forth. This is somewhat analagous to the situation that speech-in, speech-out systems encounter when deployed in real-world settings where people may talk to one another as well as to the dialog system (e.g. Hieronymus, Aist, and Dowding ICASSP 2006) – but with the use of visual input this requirement is encountered much sooner in the research and development process. This engineering requirement of converting a stream of video into a sequence of meaningful events is analagous at an abstract level to the human cognitive process of converting sensations (e.g. excitations of certain cells in the retina) to perceptions (e.g. a blue sky).

1.3. Relating Visual Perception to Cognitive and Affective State

Once a stream of visual sensory input has been processed and has yielded a stream of percepts, those perceived events mush be analyzed for their correspondence to the cognitive and/or affective state of the user. Some visual events may have relatively straightforward mappings to cognitive or affective states; for example, blink rate is generally held to increase with arousal. Others may require temporal alignment; for example, people tend to naturally look at objects not while they are speaking about them, but at some previous time – thus, using eyetracking to help resolve spoken references requires knowledge of the relative timeing of speech and gaze. Yet other events may bear only partial witness to underlying cognitive or affective state: while at times the user's face may dynamically reflect emotions, at other times (especially during concentration and learning) the face may be a "blank" mask as the user focuses on the task at hand. Despite these challenges the step of relating visual percepts to the user's cognitive and/or affective state is necessary for integrating visual input into spoken dialog systems.

1.4. Using Visually-Enhanced Models to make Dialog Decisions

There are two fundamental paradigms for using visual information in order to make dialog decisions (whether task-based, or tutorial.) The first can be broadly considered gaze-contingent (or vision-contingent). For example, if the system detects that a person has walked into the field of view, it could immediately trigger a response such as "Hello there!" Or, if the system detected that the user has fixated on a word, it could speak the work aloud a la Sibert et al.'s Reading Assistant (2000). or translate the word into another language as in Hyrskykari eta al's iDict (2003). The second paradigm is to integrate the visual information into the user modeling or dialog management process, and use the reuslting richer model to drive dialog system responses. For example, Merten and Conati (2006) describe how they used eye-tracking data to improve user modeling accuracy in the context of user exploration and user self-explanation (see also Conati et al. 2006); such improved models could then be used to result in more effective responses.

1.5. Closing the Loop: Monitoring the Effectiveness of Dialog Behaviors using Visual Input

But how can the dialog systme be sure that the responses are in fact more effective? In addition to the usual tools such as monitoring the content and speech act(s) of the user's next utterance, a dialog system that incorporates visual input can make use of that visual input (or of the user model(s) that such visual input enriches) in order to monitor the effectiveness of its own behavior. For example, if the system intends to call attention to a particular part of the screen (such as a Help text box, a part of a diagram, or a glossary explanation of a word) it would be able to monitor the user's gaze to see if in fact the action (such as highlighting the text) had the desired effect (such as getting the user to read the text in question.)

2. Opportunities Afforded by Adding Visual Input to a Dialog System

Visual input may be a Pandora's box of technological challenges – yet there is great hope for how visual input will afford new capabilities for interactive systems. For example:

VISUAL AWARENESS

1. Knowing whether there is a user present may help determine when to try to engage the user in the software; for example, a museum exhibit could actively try to recruit children to learn about its particular dinosaur ("Hey kids! Ever hear about Stegosuarus?")

2. Knowing who the user is has obvious computer security implications, but also has implications for universal usability as well; for example, upon recognition of the user, the system might adjust various factors such as screen brightness, font size, and volume level to match a particular user's preferences or accommodate a particular user's access needs.

PERSON PERCEPTION

1. Visual input added to a user model may account for more of the variance in the user's behavior than a model based on GUI and speech input alone.

2. Visual input may help in detecting short-term lapses in attention, as well as longer-term distractions. For example, a hand-analyzed sample of video from Reading Tutor use found that approximately one longish distraction occurred every session (145/149 sessions), with shorter glances away from the screen (one or two seconds) coming more frequently at four times per session (602/149) – or about once every thirty seconds (Kominek, Aist, and Mostow 1998).

EYETRACKING

1. Looking at eye-gaze patterns during reading can shed light on the reading process and on comprehension of the text. Integrating eye-tracking into dialog systems, especially into tutorial systems, offers the prospect of knowing which parts of the text are being well understood and which are more challenging.

2. Looking at patterns of eye movement may also help in determining "good" (i.e. useful) patterns of interaction, and distinguishing those patterns from less useful patterns. For example, if a user is presented with information to verify, such as a digram than summarizes the results of an interaction, a successful verification pattern might include a quick scan of the entire diagram, followed by fixations on important components. A nonuseful pattern might indicate shallow understanding of the diagram and less careful verification, such as just glancing at the title text of the diagram and then returning gaze to the main screen. (We might expect to find different behaviors for experts and novices here.)

3. Candidate Technologies for Visual Input to a Dialog System

At the present time there are a wide variety of hardware options for adding visual input to a computational system that operate well in the laboratory – and a scarcity of devices that are inexpensive, durable, and robust enough to varying input in order to operate well in long-term field studies. The ideal visual input hardware would be:

1. Precise and accurate enough to be useful
2. Non-intrusive – that is, not affixed to the user
3. Durable enough to be used in field work

4. Low-cost to enable use of many systems for large-scale studies.
5. Fast enough to provide high-frame-rate data to the system
6. Sophisticated enough to require minimal computational commitment from the main computer running the systemn
7. Able to perform calibration automatically
8. Robust to variations in the input such as ambient lighting, platform vibration, changes in lighting or in camera position, and so forth.

We now present a list of some currently available technologies, evaluated along a subset of the dimensions enumerated above: functions, speed, resolution, computation & networking requirements, and cost.

3.1. Visual Awareness and Person Perception: Hardware

There has been substantial improvement in consumer-level webcams over the last several years; at this point, the best on the market come in at around $100 for a fixed camera and as low as $150 for a camera with pan and tilt. For example, one current model from Logitech is as follows:
Model: Logitech QuickCam ® Orbit MP
Functions: Automatic Face-Tracking, Pan and Tilt
Resolution: 1.3 Megapixels
Cost: $130 as of April 30,2006.
A similar model, the Logitech QuickCam ® Fusion (TM) lacks the pan and tilt feature and is priced at $100. Creative (makers of SoundBlaster cards) also market a similar webcam, the Creative Live Motion (with pan-and-title) for $150 and a model without pan-and-tilt for $100.
Various networked cameras are available (e.g. from StarDot and Linksys) but the consensus review opinion at this time seems to be that the frame rate and image quality of a webcam connected directly to a computer is currently much better.

3.2. Eye-Tracking Hardware

Among the companies offering commercial hardware for eyetracking research are Tobii and SR Research. There are several current types of eyetracking hardware on the market, ranging from head-supported (SR Research's EyeLink 1000) to head-mounted systems (SR Research's EyeLink II). A high degree of precision can be obtained with head-supported or head-mounted systems, where the relative positions of the head and the camera(s) picking up images of the eyes are more or less fixed. For example, in a rather glowing testimonial on SR Research's web page, we have:

We have run our first full study with the new EyeLink 1000 and I must say that the machine is even better than I had expected. This is by far the best piece of eye tracking technology I know. The ultimate test was me talking while reading and the system stayed right on track, pretty amazing.

Dr. Ralph Radach
Department of Psychology
Florida State University

Selected technical specifications are as follows:
EyeLink 1000 – head-supported (like at an optician's)
Sample rate: 1000 Hz
Average delay: 3 ms
Gaze accuracy: 0.25 to 0.50 degrees
Pupil size resolution: 0.2% of diameter
Price: Tens of thousands of dollars

EyeLink II – head-mounted (on "helmet")
Sample rate: 250 Hz
Average delay: 10 ms
Gaze accuracy: < 0.50 degree average
Pupil size resolution: 0.1% of diameter
Price: Tens of thousands of dollars

Arrington Research EyeFrame BS007
Type: Head-mounted (on "glasses")
Sample rate: 60 Hz
Average delay: Not specified
Gaze accuracy: 0.25 to 1.00 degrees
Price: $12,000 including software

These head-supported or head-mounted systems do have the distinct disadvantage of requiring the user to either put his or her face into a device (in the head-supported case) or to wear a special helmet or glasses for the duration of the experiment (in the case of a head-mounted unit.) Eyetrackers integrated into monitors or otherwise capable of operating at a distance from the subject aim to avoid such issues. One leading example is:

Tobii 1750 Eye-tracker
Type: Integrated into monitor
Sample rate: 50 Hz
Average delay: Not specified
Gaze accuracy: 0.50 degrees
Price: Approximately $25,000 including software

4. Evaluation and Vision for Spoken Dialog Systems

There are two specific ways in which computer vision and eye-tracking present unique challenges and opportunities for experimental evaluation of spoken dialog systems. The first is that vision and eye-tracking can be used to evaluate a pre-existing dialog system. The second is that vision and eye-tracking, when integrated into a dialog system, present interesting challenges in terms of evaluating the overall system and its effects on user performance (or student learning.)

4.1. Using Vision and Eye-Tracking to Evaluate an Existing Spoken Dialog System

One example of this would be automating analysis such as that needed to perform the study by Kominek et al. (1998) previously mentioned. Other possible directions include:
using Visual Awareness to see how often the fundamental social assumptions of the dialog system are met (that there is one person there, that he or she is speaking to the system, and so forth.)
using person perception to see if the system's interpretation of pauses (as either pauses to think or pauses due to distraction) matches what is going on visually
using eyetracking to see if various audiovisual system responses do in fact have the desired effecs on the user in terms of attending to the relevant items.

4.2. Evaluating the use of Vision and Eye-Tracking to Enhance an Existing Spoken Dialog System

Examples of this include:
using visual awareness to notify the dialog system of when a user (or potential user) has entered the scene;
using person perception to tell when the user is engaged, frustrated, or distracted;
or using eyetracking to tell when the user is reading with concentration and perhaps even to guage the level of reading comprehension – on text alone or text with diagrams.
For each of these capabilities, various basic evaluation questions can be asked: How accurate are such notifications? How accurate is the system's interpretation of those notifications, in light of what else is going on in the dialog? What is the effect of the system's subsequent actions on vision-detectable events such as user presence/absence, attention, and so forth?

5. Conclusion

In this position paper, we have spelled out a number of interesting interactions between computer vision, eye-tracking, spoken dialog systems, and evaluation. We first discussed some architectural requirements for adding computer vision and/or eyetracking to a dialog system. We then discussed some of the opportunities that adding visual input might provide. We reviewed some of the candidate technologies available for adding visual input to dialog systems (as a service to the interested reader.) Finally we discussed some specific issues involving visual input, dialog systems, and evaluation. As the price and availability of visual input hardware improves, we exopect to see more and more dialog systems researchers and developers becoming interested in incorporating visual input into their systems, and so we believe the time is ripe for discussion of visual input and dialog systems and the unique challenges and opportunities this combination presents in terms of evaluation.

6. References

[1] Aist, G. 1998. Expanding a time-sensitive conversational architecture for turn-taking to handle content-driven interruption. International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, Nov. 30-Dec. 4. Paper 928.

[2] Allen, J., Ferguson, G., Stent, A., Stoness, S., Swift, M., Galescu, L., Chambers, N., Campana, E., and Aist, G.S. 2005. Two diverse systems built using generic components for spoken dialogue (Recent Progress on TRIPS). Interactive Demonstration track, Association for Computational Linguistics Annual Meeting, Ann Arbor, Michigan, June.

[3] Conati, C., Merten, C., Muldner, K., and Ternes, D. 2005. Exploring eye tracking to increase bandwidth in user modeling. User Modeling.

[4] Hieronymus, J., Aist, G.S., and Dowding, J. 2006. Open microphone speech understanding: Correct discrimination of in-domain speech. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Toulouse, France, May 14-19.

[5] Aulikki Hyrskykari, Päivi Majaranta, Antti Aaltonen & Kari-Jouko Räihä, Design Issues of iDict: A Gaze-Assisted Translation Aid. In: Eye Tracking Research & Applications Symposium 2000, ACM Press, November 2000, 9–14.

[6] Kominek, J., Aist, G., and Mostow, J. 1998. When listening Is not enough: Potential uses of vision for a Reading Tutor that listens. *AAAI Spring Symposium on Intelligent Environments*, Stanford, CA, March.

[7] Merten, Christina, and Conati, Cristina. 2006. Eye-tracking to model and adapt to user meta-cognition in intelligent learning environments. Intelligent User Interfaces.

[8] Sibert JL, Gokturk, M., and Lavine, RA (2000). The Reading Assistant: Eye Gaze Triggered Auditory Prompting for Reading Remediation, CHI Letters.