Vision in HCI: Embodiment, Multimodality and Information Capacity

David Powers

Artificial Intelligence Lab, School of Informatics and Engineering, Flinders University
Adelaide, Australia
Email: david.powers@flinders.edu.au

Abstract

Almost all Human Computer Interfaces involve vision and Pedagogical research encourages the use of multiple modalities including vision. The combination of visual and other modalities, as well as the many submodalities of vision, has both advantages and pitfalls. The work presented here connects psychological research into human cognitive and perceptual processes and limitations, to evaluation and optimization of multimodal HCI.

Keywords: Multimodal interfaces, interface optimization, information capacity.

1 Introduction

Why are educators encouraged to employ multimodal teaching technologies? Why do we like to use Graphical User Interfaces? Why do we need Vision in a Human Computer Interface? And how should we best utilize the various modalities and submodalities?

Most HCI interfaces do involve vision - textual interfaces involve vision both in terms of overt reading but also in terms of orientation within a document or screen. Speech recognition and speech synthesis have their technical issues, but speech has fundamental disadvantages as a sole HCI mechanism versus text, and for programming it is arguably worse - English is not a good programming or representation language, which is why we have designed mathematical and musical notations as well as programming languages.

Similarly speech lacks the persistence and position that text has in relation to other visual elements - that is we can saccade back and forward within a sentence (Huey 1908) or the text, or the program, either consciously or unconsciously, and we retain a 2D or 3D eidetic impression of where we have seen items. The formatting of a text or program, including both left and right indentation, also has a huge impact on how efficiently we can orient in a text and how fast we can read it. Standard typesetting guidelines have been developed over the centuries with a view to optimizing reading speed and orientation.

Graphic User Interfaces add another dimension but there has been a lack of Human Factors analysis in making design decisions, and there is no reason to think the designs we currently have are anything like optimal. Nonetheless, good performance can be achieved with suboptimal interfaces with sufficient training, and a better interface which is demonstrably more efficient or faster will not necessarily win widespread acceptance given the familiarity and training lock-in phenomenon of which the Qwerty vs Dvorak keyboard is a case in point.

2 Case Studies - Vision Input

2.1 Speech Reading

The classic case of a visual user interface is the use of lip-reading and its fusion with auditory information. Lip reading is good for distinguishing some phonemes that are hard to distinguish aurally, particularly in the face of noise. We will discuss developments relating to finding and tracking facial features as well as fusion of the visual and auditory information in such a way as to guarantee no significant degradation over either alone - viz. no catastrophic fusion (Lewis & Powers 2002, Lewis & Powers 2004). In addition, our research program deals with noise of different kinds, of which lighting and reverb are special cases.

2.2 Situation Awareness

One of the factors that limits the utility of natural language/speech interaction systems is the lack of shared experience/embodiment. Vision is a major part of this and can give the computer a broad view of the world, as well as detail as appropriate.

Further extensions and enhancements in relation to speech reading include the affective interpretation of facial, eye and hand gestures and movements, and the incorporate of muscular (surface EMG/sEMG), ocular (EOG) and brain (EEG) signals. sEMG alone can be used for reasonable lip reading of certain sounds, and the other signals provide correlations with a broader range of linguistic and non-linguistic communication modes and mental states, and represent an integration of AVSR and BCI (Brain Computer Interface). Again careful fusion is necessary to incorporate this information.

3 Case Studies - Vision Output

3.1 Thinking Heads

The above input modalities are complemented by speech synthesis, expression synthesis, dialogue generation and a shared interactive environment, being part of a broad Thinking Head project funded under the ARC/NHMRC Thinking Systems Special Research Initiative.

The full picture is to be exemplified and evaluated in two scenarios - a bill enquiry/complaint scenario and a Second Language (L2) teaching/learning situation. These situations afford opportunity to evaluate appropriateness of computer response and to characterize user response to different emotional/gestural expressions, including eye gaze and attention tracking. Both scenarios have the opportunity to be enhanced to take into account environmental/situational circumstances/context. In the L2 situation common reference is an essential aspect of the learning situation.

In this case we are not only talking about understanding real auditory and visual input for a Human Head (HH), but modelling and mirroring/simulating the same kind of output with a Thinking Head (TH).

3.2 Simulated Robots

The Robot World (RW) learning situation being imported into the Thinking Head L2 scenario was originally designed for studying Machine Learning of Natural Language and Ontology (L0) and First Language (L1) learning. This system avoids the problems of dealing with real robots and real vision and audition by simulating scripted scenarios and learning or teaching using these scripts.

Grammars, morphologies, ontologies and semantics can all be learned in this L0RW context. The Robot World has its limitations, and a system that is totally simulated based on existing models fails to convince after a point - after all, we are only learning the models we built in. In fact, currently we are working with the CHILDES corpus and building our scenarios around actual sentences and constructs used in child-directed speech. Nonetheless, eventually the robot learners need to see the real world.

3.3 Real Robots

Real robots are able to sense and interact with the real world, and dealing with real robots introduces considerable complexity that takes us away from the human learning and human interface.

Our robotics research has included building a doll that crawls and orients towards a voice, the original version being blind, with a new and rather too heavy head being designed with verging USB cameras and head turning/panning capability. For the new head we also developed an 8-microphone USB array that could be oriented tetrahedrally on the head (ears, mouth and crown) in noise-cancelling 180° pairs for a TH-centred soundfield. The same array can also be worn as a headset for an HH-centred soundfield.

We also use a garbage can on wheels style robot to navigate our building and develop an ontology. Using Wizard of Oz techniques using 802.11 WLAN technology, we have also used it as a building guide. This has a variety of sensors including sonar, an omnidirectional camera. We also use several USB webcams, one of which is used to track our position very precisely. We are also developing a system to read the room numbers (and eventually occupant names and other information). At this stage that is being trained with photos taken from 10 known positions and orientations for each room number, but eventually the image will be taken from the robot's cameras.

3.4 Graphical User Interfaces

The flip side of vision in HCI is the GUI or Graphical User Interface. GUI design has largely neglected human perceptual and cognitive limitations, cognitive load and situation awareness. There has been an implicit assumption that natural is better, and as a corollary, that 3D is better. But this has not been borne out empirically - the converse can be true. Better performance can result from 2D displays in an information retrieval/search context.

We have developed techniques to allow us to display up to 26 simultaneous dimensions in an IR GUI. We are also experimenting with clustering and hyperspace navigation models. But because you can do it doesn't mean you should do it or it is useful to do it.

We therefore have a research focus on understanding the interplay of the linguistic/search dimensions and the visual/graphical dimensions. There are same basic questions about how many dimensions and how many bits of information per dimension people choose to deal with or are capable of dealing with. The work in this area that Miller cited in his Magical Number Seven paper (Miller 1956a), as well as a variety of follow on studies (Miller 1956b), demonstrates that chunking and combination of dimensions can increase the amount of information that can be conveyed to at least 150 distinctions (7 to 8 bits).

In our work we are particularly interested in distinguishing between and controlling for the working memory/cognitive load aspects versus the perceptual aspects, as well as in specifying the optimum matching of application attribute/information dimensions and graphics/display dimensions (Pfitzner, Hobbs & Powers 2003).

We are also evaluating the effectiveness of animation, both as an iconic display dimension and in relation to continuity and situation awareness versus change blindness.

References

Huey, E.B. (1908), *The Psychology and Pedagogy of Reading*, MIT Press (reprinted 1968).

Lewis, T.W. & Powers, D.M.W. (2002). 'Audio-Visual Speech Recognition using Red Exclusion and Neural Networks', *in Proc. 25th Australasian Computer Science Conference (ACSC2002)*, Oudshoorn, M.J., ed., Conferences in Research and Practice in Information Technology, Vol. 4, Melbourne, Australia, Australian Computer Society, pp. 149–156.

Lewis, T.W. & Powers, D.M.W. (2004). 'Sensor Fusion Weighting Measures in Audio-Visual Speech Recognition', *in Proc. 26th Australasian Computer Science Conference (ACSC2004)*, Estivill-Castro, V., ed., Conferences in Research and Practice in Information Technology, Vol. 26, Dunedin, New Zealand, Australian Computer Society, pp. 305–314.

Miller, G.A. (1956), 'The magical number seven, plus or minus two: Some limits on our capacity for processing information', *Psychology Review*, Vol. 63, pp. 91–97.

Miller, G.A. (1956), 'Human Memory and the Storage of Information', *IRE Transactions on Information Theory*, Vol. IT-2, No. 3, pp. 129–137.

Pfitzner, D., Hobbs, V. & Powers, D.M.W. (2003), 'A unified taxonomic framework for information visualization', *in Proc. Australian Symposium on Information Visualization*, Pattison, T. & Thomas, B., eds., Conferences in Research and Practice in Information Technology, Vol. 24, Adelaide, Australia, Australian Computer Society, pp. 57–66.