Video user interfaces

Peter Robinson

University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
pr@cl.cam.ac.uk

Abstract. The increasing power and falling cost of computers, combined with improvements in digital projectors and cameras, are making the use of video interaction in human-computer interfaces more popular. This paper presents a review of video interface projects in the Computer Laboratory at the University of Cambridge over the past 15 years. These encompass early work on augmented environments, applications in publishing, personal projected displays, and emotionally aware interfaces.

1 Introduction

The increasing power and falling cost of computers, combined with improvements in digital projectors and cameras, are making the use of video interaction in human-computer interfaces more popular. This paper reviews work on video interfaces at the University of Cambridge over the past 15 years, and presents two recent projects in more detail.

With support from the Rank Xerox Research Centre in Cambridge, we laid the foundations for a new model of interaction based on video interfaces in the early 1990s. We built a user interface based on video projection and digital cameras (the *DigitalDesk*), extended this for remote collaboration (the *DoubleDigitalDesk*), and investigated the use of a camera for input alone (*BrightBoard*). The result is an augmented environment in which everyday objects acquire computational properties, rather than virtual environments where the user is obliged to inhabit a synthetic world.

The research continued with support from the EPSRC in the later 1990s to investigate combinations of electronic and conventional publishing, with applications in education. The *Origami* project combined electronic and printed documents to give a richer presentation than that afforded by either separate medium.

People manage large amounts of information on a physical desk, using the space to arrange different documents to facilitate their work. The ‘desk top’ on a computer screen only offers a poor approximation. Thales Research & Technology have supported work on the *Escritoire*, a desk-based interface for a personal workstation that uses two overlapping projectors to create a foveal display: a large display surface

with a central, high resolution region to allow detailed work. Multiple pen input devices are calibrated to the display to allow input with both hands. A server holds the documents and programs while multiple clients connect to collaborate on them.

Facial displays are an important channel for the expression of emotions, and are often thought of as projections of a person's mental state. Computer systems generally ignore this information. *Mind-reading* interfaces infer users' mental states from facial expressions, giving the computer a degree of emotional intelligence. Video processing is used to track two dozen features on the user's face. These are then interpreted as basic action units, which are interpreted using statistical techniques as complex mental states.

2 Video augmented environments

The availability of digital video projection and digital video capture in the early 1990s led us to conceive the *DigitalDesk* – an ordinary desk augmented with a computer display using projection television and a video camera to monitor inputs [22][23]. Figure 1 shows the desk with a projector (made from an overhead projector and an early liquid crystal display) and two cameras.

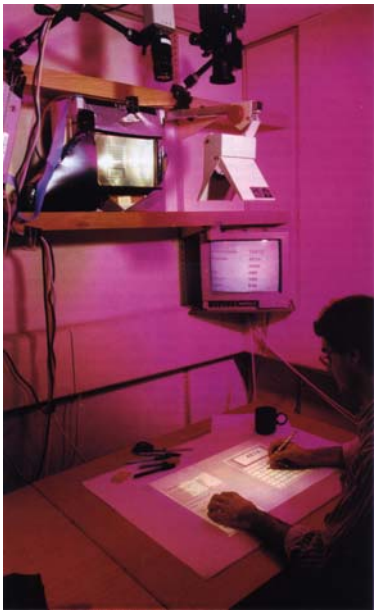


Fig. 1. The DigitalDesk

A number of prototype systems were implemented to demonstrate its feasibility. Figure 2 shows a sketching application called *PaperPaint*. The darker lines have been drawn with a pen. Some of these have then been copied electronically, and appear as grey lines in the projected image. Figure 3 shows the *DoubleDigitalDesk* where two

DigitalDesks are being used to support collaborative work [8]. The inset image at the top right shows the other participant.

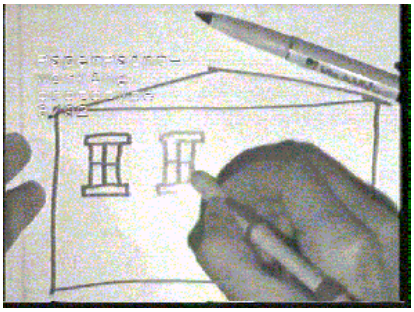


Fig. 2. PaperPaint on the DigitalDesk

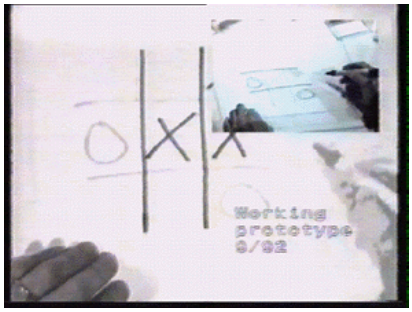


Fig. 3. The DoubleDigitalDesk

BrightBoard dispensed with the projector, and just used a camera to enable any part of the user's environment to be used to control a computer [20][21]. Figure 4 shows an ordinary whiteboard being monitored by a camera. The user could write commands on the board, for example to print a hard copy of its contents.

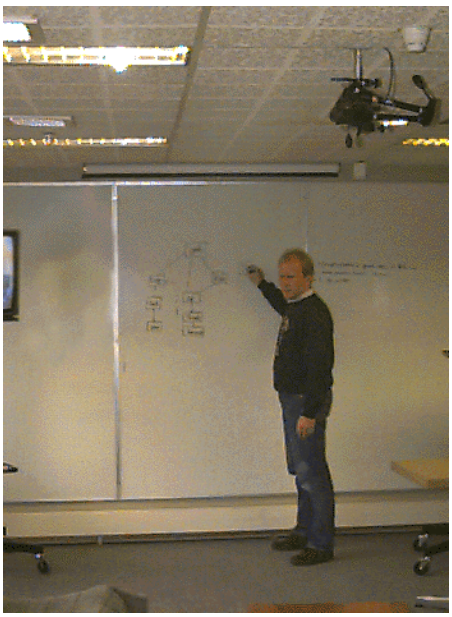


Fig. 4. BrightBoard

These early experiments established the value of *augmented environments* in which everyday objects such as paper and whiteboards acquired computational properties. This contrasts with virtual environments, where the user is obliged to inhabit a synthetic world.

3 Animated paper documents

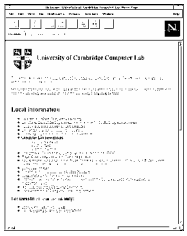
Electronic, multi-media publishing is becoming established as an alternative to conventional publishing on paper. CD-ROM and on-line versions of reference books and fiction can augment their conventional counterparts in a number of ways:

- They offer elaborate indexing, glossaries and cross-referencing.
- They allow non-linear progression through the text.
- Sound and moving images can be added.
- Sections can be copied into new documents.

However, screen-based documents have a number of disadvantages:

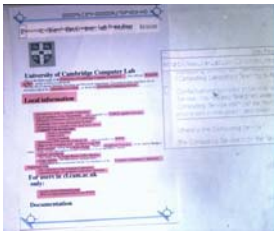
- People find screens harder to read than paper.
- Electronic bookmarks are less convenient than bits of paper or flicking through a book.
- Adding personal notes to electronic documents is more complicated than jotting in the margin of a book.
- Writing, editing and proof-reading a non-linear, multi-media document is still a specialised and difficult task.

Our solution is to publish material as an ordinary, printed document that can be read in the normal way, enjoying the usual benefits of readability, accessibility and portability. However, when observed by a camera connected to a computer, the material acquires the properties of an electronic document, blurring the distinction between the two modes of operation [16][17][19].

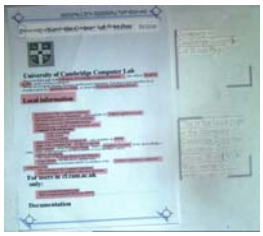


(a) Original Web page.

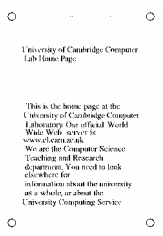
(b) Printed version.



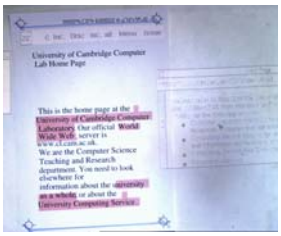
(c) Animated on the DigitalDesk.



(d) Deriving a new document.



(e) Printed version.



(f) Animating the new document.

Fig. 5. Paper access to the World-Wide Web

A simple demonstration of this principle is a system enabling interaction with printed versions of Web pages [18]. Figure 5 shows a conventional WWW page

at (a). This is imported into the system and reprinted with additional coding to assist recognition (b). When this is placed on a DigitalDesk it is recognised and active areas of the document illuminated by projected highlights. When these are selected, links are followed or programs executed and the results projected into a further window on the work surface (c).

Moreover, fragments can be copied from the paper document into new electronic documents also projected onto the desk (d). The new document can be printed to give a new paper document (e) which can be animated on the desk in just the same way (f).

Two further applications explored the use of this technology for educational material. The first is a course book for teaching mathematics [9]. The software which accompanies the course book is automatically launched when the book is first placed on the desk. Figure 6 shows a section on curve-sketching for polynomials. The generic equation of a quadratic polynomial is given with spaces for the values of the coefficients and an empty box underneath for plotting the graph. The software projects default values and draws the graph into the box. However, it also projects controls alongside the coefficients to allow the reader to change these values while observing the corresponding change in the graph.

Further down the page of the course book there is an assessment exercise. This time the polynomials are fixed and the student must draw the curve into the box (the active pen also has a real nib for writing). Clicking a projected button asks the computer to assess the sketch. The image is captured and analysed for features such as maxima, minima and axis crossings, and marked accordingly.

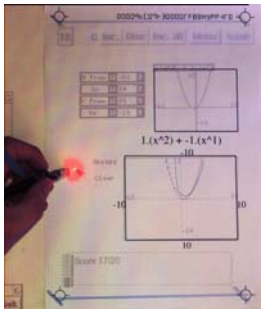


Fig. 6. A maths tutor

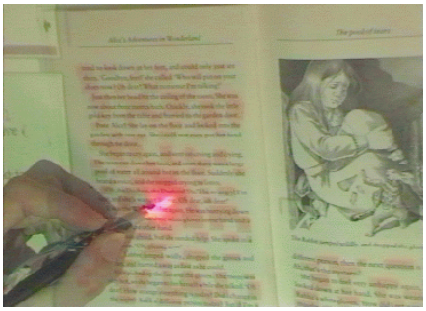


Fig. 7. A grammar tutor

Figure 7 shows a second educational application that teaches elementary grammar by animating a standard printed book [6][7]. This uses additional information from an SGML edition of the book distributed as part of the Text Encoding Initiative.

4 Personal projected displays

Since the inception of the personal computer, the interface presented to users has been defined by the monitor screen, keyboard, and mouse, and by the framework of the desktop metaphor. It is very different from a physical desktop which has a large

horizontal surface, allows paper documents to be arranged, browsed, and annotated, and is controlled via continuous movements with both hands. The desktop metaphor will not scale to such a large display; the continuing profusion of paper, which is used as much as ever, attests to its unsurpassed affordances as a medium for manipulating documents; and despite its proven benefits, two-handed input is still not used in computer interfaces [14][15].

The *Escritoire* [1] uses a novel configuration of overlapping projectors to create a large desk display that fills the area of a conventional desk and also has a high resolution region in front of the user for precise work. The projectors need not be positioned exactly—the projected imagery is warped using standard 3D video hardware to compensate for rough projector positioning and oblique projection. Calibration involves computing planar homographies between the 2D co-ordinate spaces of the warped textures, projector framebuffers, desk, and input devices. The video hardware can easily perform the necessary warping and achieves 30 frames per second for the dual-projector display. Oblique projection has proved to be a solution to the problem of occlusion common to front-projection systems. The combination of an electromagnetic digitizer and an ultrasonic pen allows simultaneous input with two hands. The pen for the non-dominant hand is simpler and coarser than that for the dominant hand, reflecting the differing roles of the hands in bimanual manipulation. We use a new algorithm for calibrating a pen, that uses piecewise linear interpolation between control points. We can also calibrate a wall display at distance using a device whose position and orientation are tracked in three dimensions.

The Escritoire software is divided into a client that exploits the video hardware and handles the input devices, and a server that processes events and stores all of the system state. Multiple clients can connect to a single server to support collaboration. Sheets of virtual paper on the Escritoire can be put in piles which can be browsed and reordered. As with physical paper this allows items to be arranged quickly and informally, avoiding the premature work required to add an item to a hierarchical file system. Another interface feature is pen traces, which allow remote users to gesture to each other. We report the results of tests with individuals and with pairs collaborating remotely. Collaborating participants found an audio channel and the shared desk surface much more useful than a video channel showing their faces.

The Escritoire is constructed from commodity components, and unlike multi-projector display walls its cost is feasible for an individual user and it fits into a normal office setting. It demonstrates a hardware configuration, calibration algorithm, graphics warping process, set of interface features, and distributed architecture that can make personal projected displays a reality.

4.1 Foveal display

To create a display that fills an entire desk but also allows life-sized documents to be displayed and manipulated we have created what we call a *foveal* display. One projector fills the desk with a low-resolution display, while a second overlapping projector displays a high-resolution area in front of the user. The optical path of the first projector is folded using a mirror above the desk to enable it to generate a display of

the desired size without being mounted at an inconveniently high position above the desk surface. Figure 8 shows the general arrangement. Baudisch et al. have combined an LCD monitor and a projector to get a dual-resolution display [2], although they do not address calibration, have used only a conventional keyboard and mouse for input, and get a display with different affordances because of its vertical rather than horizontal placement.

The user can arrange items on the desk, identify them at a glance, reach out and grab them, and quickly move them to the high-resolution region where the text becomes legible and they can be worked on in detail. Figure 9 shows a document being moved from the periphery into the fovea.

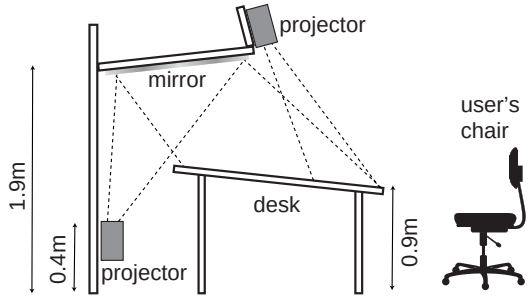


Fig. 8. The two-projector configuration of the Escritoire

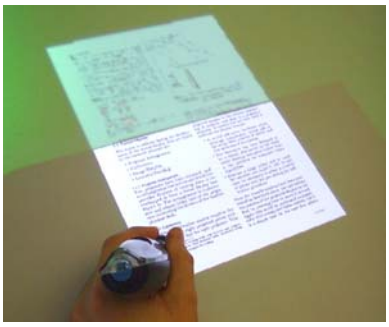


Fig. 9. Moving a sheet of virtual paper to the high-resolution region

4.2 Two-handed interaction

Bimanual input—using two hands—has manual benefits from increased time-motion efficiency due to twice as many degrees of freedom being simultaneously available to the user, and also cognitive benefits which arise as a result of reducing the load of mentally composing and visualizing a task at an unnaturally low level imposed by traditional single-handed techniques.

We have combined a desk-sized digitizer and stylus that provide accurate input for the user's dominant hand, with an ultrasonic whiteboard pen that provides simple and less accurate tracking for the user's non-dominant hand. The non-dominant hand is used to move items around on the desk, setting up a frame of reference for the dominant hand to do its more detailed work such as writing and drawing.

4.3 Collaboration

We have implemented the Escritoire in two parts: a server written in Java that stores the details of the items on the desk , and a client written in C++ that handles the input

and output devices. This allows multiple desks to connect to the same server over the Internet allowing geographically separated users to share the desk contents.

We have conducted tests in which pairs of participants converse over a standard videoconference while using Escritoire desks whose contents are shared in a What You See Is What I See fashion. Figure 10 shows a videoconference being conducted on an ordinary computer, but where both participants are also using a pair of Escritoires driven from the same server. As they talk they can work together to read and annotate documents, gesturing in the shared graphical space as they do so. Systems for remote collaboration often concentrate on optimizing the talking heads model of a standard videoconference but we have found that a shared task space is often more useful. The shared space provided by the Escritoire is much larger than a monitor screen and supports fast and natural interaction over the whole area, so users share a large visual context while being able to easily refer to and collaborate on specific items.

Fig. 10. Augmenting a videoconference with a desk surface that is shared between collaborators

5 Mind-reading interfaces

People routinely express their emotions and mental states through their facial expressions. Other people are used to this, and read their minds accordingly. This non-verbal communication is a vital part of human society, and those who lack the ability to read facial expressions are at a disadvantage. All computers suffer this disadvantage by failing to read their users' minds. In effect, computers are autistic. We have developed an automated system to remedy this problem [11][12][13].

In order to support intelligent man-machine interaction the system is designed to meet three important criteria. These are full automation so that it requires no human intervention, the ability to execute in real-time, and the categorization of mental states early enough after their onset to ensure that the resulting knowledge is current and actionable. Other aspects include being user-independent and dealing with substantial

rigid head motion. The experimental evaluation shows promising results for 24 classes of complex mental states (sampled from 6 groups) in different interaction scenarios.

5.1 Multi-level representation

A person's mental state is not directly available to an observer (the machine in this case) and as a result has to be inferred from observable behaviour such as facial signals. The process of reading a person's mental state in the face is inherently uncertain. Different people with the same mental state may exhibit very different facial expressions, with varying intensities and durations. In addition, the recognition of head and facial displays is a noisy process.

To account for this uncertainty, we use a multi-level representation of the video input, combined in a Bayesian inference framework. Our system abstracts raw video input into three levels, each conveying face-based events at different granularities of spatial and temporal abstraction. Each level captures a different degree of temporal detail depicted by the physical property of the events at that level. As shown in Figure 11, the observation (input) at any one level is a temporal sequence of the output of lower layers. At the bottom level, 24 facial feature points are tracked in each new frame every 33ms. Figure 12 shows hierarchy of the spatial analysis consisting of:

- *actions* which are explicitly coded being detected every 166ms,
- *displays* recognised by Hidden Markov Models (HMMs) every second,
- *mental states* assigned probabilities by Dynamic Bayesian Networks (DBNs) every two seconds.

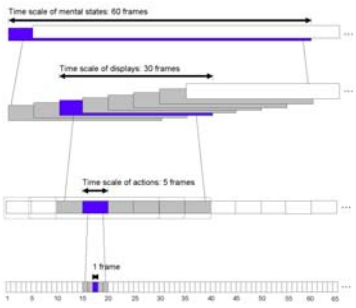


Fig. 11. Temporal abstraction in the mind-reading machine

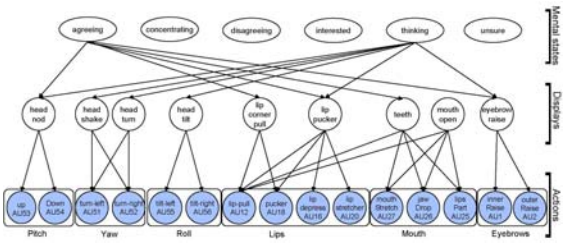


Fig. 12. Classification hierarchy

This approach has a number of advantages. First, higher-level classifiers are less sensitive to variations in the environment because their observations are the outputs of the middle classifiers. Second, with each of the layers being trained independently, the system is easier to interpret and improve at different levels. Third, the Bayesian framework provides a principled approach to combine multiple sources of information. Finally, by combining dynamic modelling with multi-level temporal abstraction, the model fully accounts for the dynamics inherent in facial behaviour. In terms of

implementation, the system is user-independent, unobtrusive, and accounts for rigid head motion while recognizing meaningful head gestures.

5.2 Training

A great deal of data was necessary to determine the window sizes in the temporal abstraction and to train the statistical classifiers in the inference system. We have used the Mind Reading DVD [5], a computer-based guide to emotions, developed by a team of psychologists led by Professor Simon Baron-Cohen at the Autism Research Centre in the University of Cambridge. The DVD was designed to help individuals diagnosed along the autism spectrum recognize facial expressions of emotions.

The DVD is based on a taxonomy of emotion by Baron-Cohen *et al*. [4] that covers a wide range of affective and cognitive mental states. The taxonomy lists 412 mental state concepts, each assigned to one (and only one) of 24 mental state classes. The 24 classes were chosen such that the semantic distinctiveness of the emotion concepts within one class is preserved. The number of concepts within a mental state class that one is able to identify reflect one's empathizing ability [3].

Out of the 24 classes, we focus on the automated recognition of 6 classes that are particularly relevant in a human-computer interaction context, and that are not in the basic emotion set. The 6 classes are: *agreeing*, *concentrating*, *disagreeing*, *interested*, *thinking* and *unsure*. The classes include affective states such as *interested*, and cognitive ones such as *thinking*, and encompass 29 mental state concepts, or fine shades, of the 6 mental states. For instance, *brooding*, *calculating*, and *fantasizing* are different shades of the *thinking* class; likewise, *baffled*, *confused* and *puzzled* are concepts within the *unsure* class.

Each of the 29 mental states is captured through six video clips. The resulting 174 videos were recorded at 30 frames per second, and last between 5 to 8 seconds at a resolution of 320×240. The videos were acted by 30 actors of varying age ranges and ethnic origins. All the videos were frontal with a uniform white background. The process of labelling the videos involved a panel of 10 judges who were asked "could this be *the emotion name*?" When 8 out of 10 judges agreed, a statistically significant majority, the video was included. To the best of our knowledge, the Mind Reading DVD is the only available, labelled resource with such a rich collection of mental states, even if they are posed.

5.3 Operation

Figure 13 shows the system in operation. Seven frames from a six second performance of the *undecided* emotion are shown. These are followed by the outputs from the HMMs during the video for five displays – *head nod*, *head shake*, *head tilt*, *head turn*, and *lip pull*. Finally, the outputs from the DBNs are shown giving the probabilities of the six mental state classes during the clip.

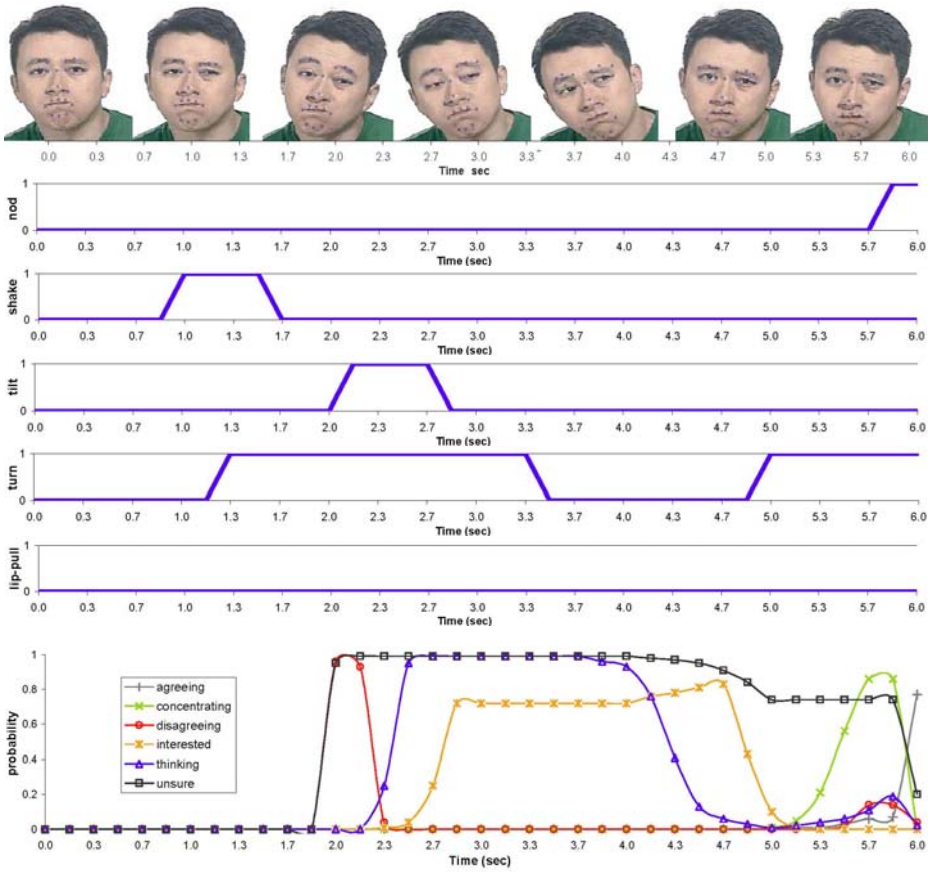


Fig. 13. Selected frames, traces of display recognition, and probabilities of mental state inference in a video labeled as *undecided*

The probabilities of the different mental state classes vary during the course of the video, and there are several plausible interpretations. This reflects the position with recognition of emotions by humans. A principal state can be inferred by measuring the area under the six graphs, and selecting the largest. In this case, *unsure* is correctly selected as the class within which *undecided* falls.

The overall accuracy of the system was evaluated by testing the inference results of 164 videos representing the six mental state classes. The videos span 25645 frames, or approximately 855 seconds. Using a leave-one-out methodology, 164 runs were carried out, where for each run the system was trained on all but one video, and then tested with that video. The classification rule that is used to deem whether a classification result is correct is defined as follows: compare the overall probability of each of the mental states over the course of a video. If the video's label matches that of the most likely mental state or the overall probability of the mental state exceeds 0.6, then it is a correct classification.

The results are summarized as a 3D bar chart in Figure 14. The horizontal axis represents the classification results of each mental state class. The percentage of recognition of a certain mental state is represented along the z-axis.

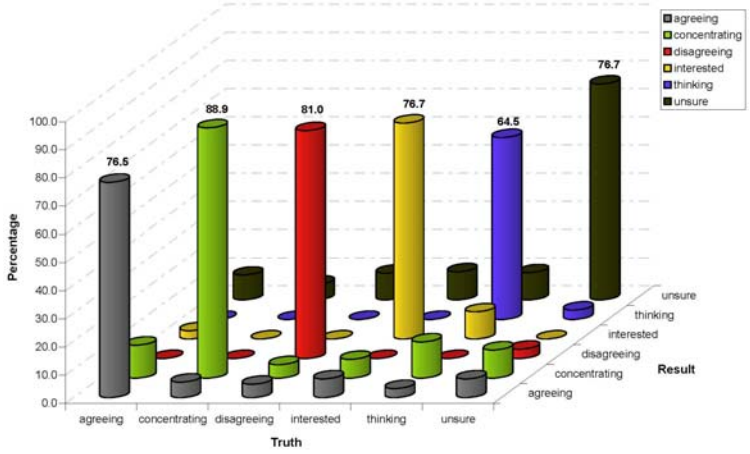


Fig. 14. Confusion matrix for the six classes of mental state used in the trials

For a mean false positive rate of 4.7%, the overall accuracy of the system is 77.4%. These results compare favourably with an earlier experiment in which the performance of a group of people in recognizing complex mental states in a similar set of videos from the Mind-reading DVD was tested []. In that experiment, human recognition rate reached an upper bound of 71.0 %. Thus, the accuracy of the automated mind-reading system in classifying complex mental states from videos of the Mind-reading DVD compares favourably to that of humans. Moreover, the system operates in real time on a standard computer workstation.

We are currently evaluating the performance of the automated mind-reading system in a more general context. The idea is to train the system on videos from the Mind-reading DVD, and test its performance on a previously unseen corpus with different recording conditions and subjects than those used in training the system. The generalization performance of a system is an important indicator to how well the system does outside of laboratory settings.

6 Conclusions

This paper has reviewed work on video user interfaces over 15 years at the University of Cambridge Computer Laboratory. The initial view that using cameras and projectors as part of the human-computer interface has proved extremely fruitful. Indeed, the steady improvements in technology over this period mean that computers are now 1000 times faster and have 1000 times the memory. Cameras have fallen in price by a similar factor. Projectors have also improved in brightness and resolution, and fallen in price, albeit by a rather smaller factor.

Many of the technical challenges remain the same. Projection systems require non-linear transformations to accommodate oblique projection and to correlate the coordinate systems of the different input and output devices. Analysing video input is expensive in terms of both processing and memory. However, the hardware of modern graphics cards can be exploited to offload much of this processing, and the systems now run comfortably on commodity hardware.

The experimental systems and applications investigated over the past 15 years in Cambridge are now entering the main stream. The Escritoire is being used to support distributed command and control systems for crisis management. Mind-reading interfaces are being used to augment teleconferencing systems and to control figures in computer animations. Video input and output are focal in the movement towards improved availability and usability of computer systems.

References

1. Ashdown, M.: *Personal Projected Displays.* PhD Dissertation, University of Cambridge Computer Laboratory Technical report 585, September 2003.
2. Baudisch P., Good N., Stewart P.: Focus Plus Context Screens: Combining Display Technology with Visualization Techniques. *Proceedings of UIST 2001*, pages 31–40.
3. Baron-Cohen,S.: *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, 1995.
4. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.: *A New Taxonomy of Human Emotions*. 2004.
5. Baron-Cohen, S., Golan, O., Wheelwright, S., Hill, J.: *Mind Reading: The Interactive Guide to Emotions*. London: Jessica Kingsley Publishers, 2004.
6. Brown H., Harding R.D., Lay S.W., Robinson P., Sheppard D.P., Watts R.R.: *Active paper for active learning.* Proceedings 4th annual conference Association for Learning Technology, Telford, September 1997, reprinted in Association for Learning Technology Journal, 1998.
7. Brown H., Harding R.D., Lay S.W., Robinson P., Sheppard D.P., Watts R.R.: *Active Alice - using real paper to interact with electronic text*. Proceedings 7th International Conference on Electronic Publishing, Saint Malo, April 1998, ISBN 3 540 64298 6, pp 407-419.
8. Freeman, S.M.G. *An architecture of distributed user interfaces.* PhD Dissertation, University of Cambridge Computer Laboratory Technical Report 342, July 1994.
9. Harding R.D., Lay S.W., Robinson P., Sheppard D.P., Watts R.R.: *New technology for interactive CAL - the Origami project*. Proceedings 3rd annual conference Association for Learning Technology, September 1996, reprinted in Association for Learning Technology Journal 5(1), 1997, ISSN 0968 7769, pp 6-12.
10. el Kaliouby R., Robinson P., Keates S.: Temporal Context and the Recognition of Emotion from Facial Expression. Proceedings of the *HCI International Conference*, June 2003.
11. el Kaliouby R., Robinson P.: Mind-reading Machines: Automated Inference of Cognitive Mental States from Video. Proceedings of IEEE International Conference on Systems, Machines and Cybernetics, 2004.
12. el Kaliouby R., Robinson P.: Real-time inference of complex mental states from facial expressions and head gestures. Workshop on Real-Time Vision for Human-Computer Interaction at the IEEE CVPR Conference, 2004.
13. el Kaliouby R., Robinson P.: The emotional hearing aid: an assistive tool for children with Asperger's Syndrome. International Workshop on Universal Access and Assistive Technology, pages 244–246, 2004.

14.Leganchuk A., Zhai S,, Buxton W.: Manual and Cognitive Benefits of Two-Handed Input: An Experimental Study. *Trans. on HCI 5(4)*, pages 326–359, 1998.
15.Norman D.: *The Psychology of Everyday Things.* Basic Books, 1988.
16.Robinson P., Sheppard D.P., Watts R.R., Harding R.D., Lay S.W.: *Animated paper documents.* Proceedings HCI '97, San Fransisco, August 1997, reprinted in Design of computing systems: social and ergonomic considerations 21B, Elsevier 1997, ISBN 0 444 82183 X, pp 655-658.
17.Robinson P., Sheppard D.P., Watts R.R., Harding R.D., Lay S.W.: *A framework for interacting with paper.* Proceedings Eurographics '97, Computer Graphics Forum 16(3), September 1997, ISSN 0167 7055, pp 339-324.
18.Robinson P., Sheppard D.P., Watts R.R., Harding R.D., Lay S.W.: *Paper interfaces to the World-wide.* Procedings WebNet ’97. Toronto, November 1997, ISBN 1 880094 27 4, pp 426-431.
19.Robinson P.: *Digital manuscripts and electronic publishing.* International Congress on Production and Context, Constantijn Huygens Institute, The Hague, March 1998; reprinted in *Editio* 13, Autumn 1999, pp 337-346.
20.Stafford-Fraser, J.Q.: *Video augmented environments*. PhD Dissertation, University of Cambridge Computer Laboratory Technical Report 419, February 1996.
21.Stafford-Fraser J.Q., Robinson P.: *BrightBoard - a video augmented environment.* Proceedings ACM Conference on Computer-Human Interaction, April 1996, pp 134-141.
22.Wellner P.D.: *Interacting with paper on the DigitalDesk.* Communications of the ACM 36(7), July 1993, pp 87-96.
23.Wellner P.D.: *Interacting with paper on the DigitalDesk*. PhD Dissertation, University of Cambridge Computer Laboratory Technical Report 330, October 1993.

Acknowledgements

The systems described in this paper are the results of work by many people. Pierre Wellner, Steve Freeman and Quentin Stafford-Fraser undertook early experiments on the DigitalDesk, DoubleDigitalDesk and BrightBoard while working as research students in the Computer Laboratory with sponsorship from the Rank Xerox Research Centre in Cambridge. Steve Lay, Dan Sheppard and Richard Watts investigated animated paper documents under a grant from the UK Engineering and Physical Sciences Research Council. Robert Harding and Heather Brown also contributed to the work. Mark Ashdown designed and built the Escritoire while working as a research student in the Computer Laboratory with sponsorship from Thales Research and Technology. Rana el Kaliouby designed and built the mindreading system as a research student. Simon Baron-Cohen and his colleagues in the Autism Research Centre gave valuable advice on emotional intelligence, and supplied test data on DVD.