The Digital Office: Overview

Michael J. Black[1], **François Bérard**[3], **Allan Jepson**[4], **William Newman**[2], **Eric Saund**[1], **Gudrun Socher**[1], **Michael J. Taylor**[2]

[1]Xerox PARC, 3333 Coyote Hill Rd., Palo Alto, CA 94304, USA
[2]Xerox Research Centre Europe, 61 Regent St., Cambridge CB21AB, UK
[3]CLIPS-IMAG, BP 53, 38041 Grenoble cedex, France
[4]Department of Computer Science, University of Toronto, Toronto, Canada, M5S 1A4

e-mail: {black,saund,socher}@parc.xerox.com, {wnewman,mtaylor}@xrce.xerox.com

Abstract

This paper describes our efforts to develop a "Digital Office" in which we augment a physical office setting with cameras and other electronic devices. Our goal is to bring the worlds of electronic and physical documents closer together and to facilitate the interaction of humans with all kinds of documents. In the "Digital Office" we extend the traditional notion of "scanning" documents to include the capture of whiteboards, books, desktops, and the human office workers themselves. In particular, we give an overview of three systems in which video cameras unobtrusively observe and capture whiteboards and human gestures, papers on the desktop, and the motion of a user's face which is used to control the display of electronic documents in a browser. Each of these systems combines the features and affordances of both physical and electronic documents, and together they begin to illuminate the intelligent office environment of the future.

Introduction

In a typical office we are likely to find computers, printers, fax machines, whiteboards, desktops, people, and, quite often, piles and piles of paper. Work often centers around *documents* which may exist in electronic form (e.g. Microsoft Word documents, video, voice mail) or in physical form (e.g. printed documents, notes, whiteboard markings). While printers and digital scanners allow us to get documents from printed form to electronic and back again, this sort of interaction covers only a small fraction of the types of documents people actually use in their work. The goal of our research is to help bridge the physical and electronic worlds of documents.

One step in this direction is to expand the traditional notion of copying and scanning to allow us to "scan the world." We envision a "Digital Office" in which cameras unobtrusively "scan" documents in the office including whiteboards and papers on the desktop. Unlike the traditional model of document scanning, in this Digital Office, the users of the documents can be scanned as well. This may be a problem in cases where the user obscures the camera's view of the document. On the other hand, the presence of the user in the scanning process presents tremendous opportunities for developing new tools to help people find, scan, print, interact with, and manage documents.

The official version of this paper has been published by the American Association for Artificial Intelligence (http://www.aaai.org). AAAI Spring Symposium on Intelligent Environments, 1998.

This paper reviews some of the ongoing work at Xerox that explores the Digital Office concept. In particular, we briefly describe three different systems that bring together the physical and electronic document worlds.

The first is the "ZombieBoard" which is a video-based high resolution whiteboard scanner. ZombieBoard provides a testbed for exploring new forms of human computer interaction based on hand-drawn diagram understanding and gesture recognition. The second application is desktop scanning using a video camera with, and without, projection onto the desktop. The final application we discuss, is a "Perceptual Browser" that allows users to interact with electronic documents in a way that is somewhat analogous to their interactions with paper documents. Once cameras are common in the officeplace, we will want to be able to locate and track the motion of people and their document use. The Perceptual Browser uses motion estimation techniques to track a user's head and to control the display of an electronic document in a browser. Each of these applications is described below in greater detail.

The Digital Office is not only a place for the seamless interaction of physical and electronic documents but also an environment for experimenting with new forms of human-computer interaction. Computer vision research has reached a level of maturity that now permits its application to many problems in human-computer interaction. Numerous research groups are exploring perceptual user interfaces (PUI's) (Turk, 1997) that exploit vision techniques to track people and understand their actions thereby creating computer interfaces that are embedded in the physical world (Bobick et al., 1996; Pentland, 1996; Abowd et al., 1997; Waters et al., 1996).

ZombieBoard Whiteboard Scanner

Computer Vision can be useful in instrumenting ordinary whiteboards in offices and conference rooms to better support individual and collaborative work. *Image mosaicing* enables high-resolution whiteboard image capture (Szeliski, 1994); *image motion and gesture analysis* are useful for detecting human activity in front of whiteboards; and *line drawing analysis* underlies interpretation of diagrammatic information drawn on whiteboards. Our prototype, the “ZombieBoard” Whiteboard Scanner (it brings to electronic “life” the ink marks on a whiteboard), is currently in routine use in about ten offices, open areas and conference rooms at Xerox PARC.

Camera-based Whiteboard Capture

The existence of several commercial devices for online or offline whiteboard image capture suggests that the material drawn on whiteboards is indeed valuable and worth preserving, editing, and sharing after the meeting or individual whiteboard use session. However, available devices all severely limit the size of the region that can be captured.

ZombieBoard performs whiteboard scanning through image mosaicing using a camera on a pan/tilt head, mounted to the ceiling. Depending upon the number and arrangement of image tiles pieced together, any size whiteboard can be scanned at any desired resolution. For a typical 8’ x 4’ conference room whiteboard, fourteen closeup snapshots are pieced together to achieve a scanning resolution of 30 dots/inch. The mosaicing algorithm is feature-based; batch, to permit large two-dimensional assemblies of tiles; requires very little overlap between tiles; and works with sparse image data as is frequently encountered in whiteboard scenes.

Diagrammatic User Interface

An intelligent room does not force the user to attend to a computer explicitly in order to exploit computationally-provided capabilities. To do so when working on ideas at a whiteboard would disrupt the pace and demeanor of office and conference room work. Therefore, ZombieBoard provides not only a Graphical User Interface via a web browser, but also a Diagrammatic User Interface (DUI), permitting the user to issue commands by drawing on the whiteboard itself.

In our current DUI protocol design, users draw a “button” accompanied by annotations, e.g. number of copies desired. When the button is “pressed” by drawing a check mark or X, the command is carried out. Figure 1 is a whiteboard scan taken in one of our conference rooms illustrating aspects of the diagrammatic protocol. The computer vision required to implement the Diagrammatic User Interface includes real-time activity detection, and line drawing interpretation (Saund, 1996).

The activity detection stage attempts to maintain a model of the static material drawn on the whiteboard, as distinguished from people and other mobile objects. Using frame differencing, filtering, histogramming, and related image processing techniques, the board model is incrementally updated and refined as changes to markings on the whiteboard become visible. As they are detected, changed regions of the board are passed to the line drawing interpretation module.

The line drawing analysis component of the system must account for tremendous variability in the appearance of diagrammatic commands due to different drawing styles, poor imaging conditions, and variations in camera imaging geometry. Our approach is based on perceptual organization through token grouping, in the style of Marr’s Primal Sketch (Marr, 1976). Primitive curve fragments are grouped into larger curvilinear strokes, corners, parallels, etc., which serve as perceptually salient objects with which to perform model-based recognition of buttons, check marks, and other diagrammatic components.

In order to support more open-ended diagrammatic analysis operations, an architecture is utilized following Ullman’s notion of Visual Routines (Ullman, 1983). For example, as part of the diagrammatic protocol users may delineate a complex polygonal region of the board which they wish to scan (cf. Stafford-Fraser, 1996), motivating the use of recognition via a curve tracing routine.

Gesture Recognition

We are also using the ZombieBoard as a testbed for human gesture recognition. In our scenario, when the user wants to perform a command, they pick up a gesture “phicon” (or physical icon) (Ishii & Ullmer, 1997) that has a distinctive color that makes it easy to locate and track. The motion of the phicon is tracked using a color histogram tracker in real time (Coutaz et al., 1996). The horizontal and vertical velocities of the phicon are used for gesture recognition. Currently the system recognizes the following gestures (see Figure 2a):

- **Start**: Tells the system to and start interpreting gestures.
- **Cut Region:** Indicate a region of the whiteboard to be scanned (possibly at higher resolution). This gesture consists of three primitive gestures
  - **Cut-On**: an upside-down “check mark” marks the upper left corner of the scanning region.
  - **Cut**: The user then moves the phicon to the lower right corner of the region.
  - **Cut-Off**: a “check mark” ends the gesture and cuts the image region.
- **Print**: To send a cut region to the printer.

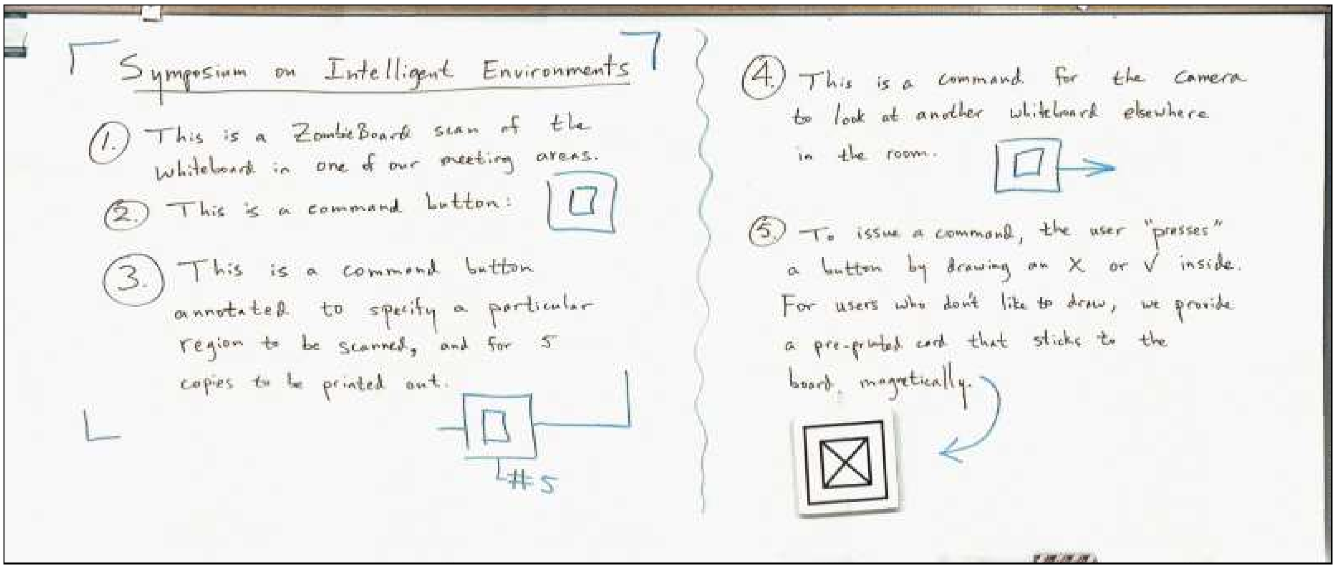
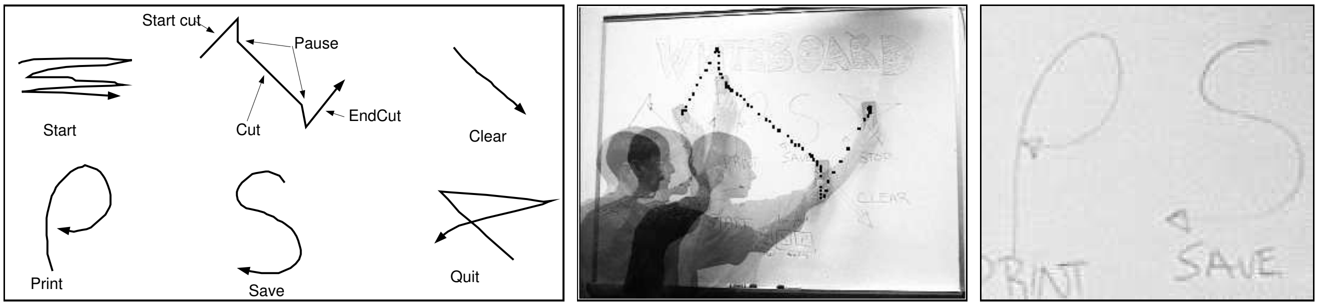


Figure 1: ZombieBoard scan that illustrates aspects of the diagrammatic protocol.



a) b) c)

Figure 2: Gesture Recognition: a) Example gestures understood by the system. b) Example of a "Cut" gesture. The user makes a gesture with the phicon. c) The detailed region that is cut out of the larger image.

- **Save**: To save the region to a file.
- **Clear**: A sharp diagonal motion "clears" the current stored whiteboard region.
- **Quit**: An "X" gesture stops the gesture recognition function.

Figures 2b and 2c illustrate the performance of a "cut" gesture. We use a bright red block as our gesture phicon. The black dots in the figure represent tracked locations of the phicon. Figure 2c is the region that is cut out by our algorithm. The gesture recognition method uses the "Condensation algorithm" to incrementally match learned gesture models to the tracked phicon data (see Black & Jepson (1998) for details).

Scanning over the desk: LightWorks and CamWorks

Paper still plays a key part in many document related tasks. Principally, paper is cheap, portable, easy to annotate and easy to read (O'Hara & Sellen, 1997; Sellen & Harper, 1997). These properties are proving rather difficult to replicate electronically, and in the meantime, paper is likely to maintain its role in the document life-cycle for some time to come. As a consequence of these properties, much information continues to arrive at our desks in a paper form. Familiar examples are annotations on a draft report, articles in newspapers and magazines bought on the way to work and passages from books.

Existing paper-to-paper scanners typically come in one of three main forms: namely the flat-bed scanner, the sheet-feed "keyboard" scanner and the hand-held scanner. All three are rather cumbersome to use. The flat-bed and sheet-

a) b) c)

Figure 3: Video scanning: a) LightWorks: a projected user- interface based on the DigitalDesk. The normal computer display is projected onto the desk where the paper is being read. The user copies directly from the page as he reads. b) CamWorks, a screen- based video scanning system: the user-interface is on a live video image on the computer screen. c) The CamWorks user interface. A live image is displayed on the computer screen. Parts of the image can be copied very conveniently without moving the document or executing a pre-scan.

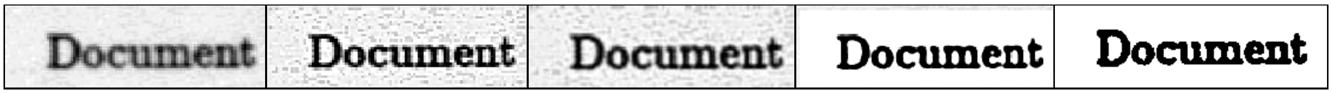
feed scanners require that the document be moved from its place of reading for scanning. The flat bed is worst in this respect as it takes up rather a lot of desk space and is therefore often situated away from the reader's desk. The sheet-feed scanner usually resides on the reader's desk but only takes single sheets of a limited range of paper sizes and therefore cannot handle newspapers, magazines, books or bound reports. Those documents it can scan typically need to have the staples removed and require a time-consuming manual feed. All three devices use a contact scanning process and as such are limited to scanning flat objects. This is how high resolution images of documents can be achieved for relatively little cost. However, ideally it would be possible to capture images of objects as well as documents with the same input device. We would like to combine the functionality of the traditional paper scanning devices with that of the digital camera.

Two possible Interfaces for Video Scanning

Over-the-desk video scanning systems avoid these inconvenient aspects of the user-interface. The LightWorks system, a development of the original DigitalDesk concept (Wellner, 1993), is shown in Figure 3a. A computer display is projected onto the desk surface where the document lies. In this manner, the user is able to select a region to scan with the mouse without taking his eyes off the page. Figure 3b shows CamWorks, an alternative arrangement where the user-interface is on the screen instead of on the page. The user selects regions to scan from a live image of the desk displayed on the computer screen (see Figure 3c). The user can then copy or cut the region and send it to a nearby printer or integrate it into a different document.

The projected interface has the advantage that the interface is all on the paper and therefore supports a potentially more natural mode of interaction during the reading task. However, as we have already noted, scanning tasks involve copying paper-based information into an electronic form. This implies that there is usually a destination application for the copied image, such as a word processor, an optical character recognition (OCR) program, an image editor, a spreadsheet, a fax server or a video-conferencing system. The screen-based interface is better suited to these destination applications, and so the shift in user focus from the page to the screen is not necessarily a disadvantage.

Over-the-desk scanning with video cameras has many advantages over traditional scanning techniques. Perhaps the most obvious is that of convenience: documents need not be moved from their usual place of reading, so the user is able to scan without the usual overhead of moving the document. This promotes the act of "casual" scanning: the user is able to scan small amounts of interesting paper information as they are encountered whilst reading, rather than having to make a note of their position in a document so that they can be scanned or copied at a later stage. A second advantage is that the non-contact nature of the scanning process means that 3D objects (such as a pointing finger) can be captured by the same scanning process as the document. The third is the ability to display video sequences as visual feedback on the computer screen. This shows the user exactly what is being scanned, rather like a very efficient "pre-scan" phase that is often performed with traditional flat-bed scanners. This real-time video scanning also has the potential to support a very rich set of user interactions such as tracking gestures, hand-writing, or annotations for either local or collaborative user interfaces.



a) b) c) d) e)

Figure 4: The four stages of the image restoration algorithm: a) Original. b) After high frequency boost. c) Resolution enhancement via linear interpolation. d) After contrast enhancement. e) Binary image after thresholding.

Image Restoration

The main problem in video scanning is how to convert the grey-scale images from the camera to a form as close to the original paper image as possible. We assume that the original document image is bimodal in color space (typically black and white) and that the lighting conditions on the desk are within a specified tolerance. We require a computationally efficient conversion of low resolution grey-scale images to higher resolution binary images of documents. The method we use for binary image restoration has four stages (see Fig. 4): (1) High frequency boost, (2) spatial resolution enhancement, (3) adaptive contrast enhancement, and (4) thresholding. The four stages are described in detail by Taylor & Zappala (1997). The restored images are good enough for fax and reasonable OCR accuracy.

Perceptual Browser

CamWorks and ZombieBoard provide a bridge between physical and electronic documents and this leads to new genres of documents that exist simultaneously in multiple forms. Many issues arise regarding the interaction of a user with these multiple forms of the document. In particular, the ways we interact with electronic documents are typically very different from interactions with paper documents. Not only do we want to enhance physical documents with many of the properties of electronic documents, but we also want to take our experiences with physical artifacts and use them for developing new ways of interacting with electronic documents that mirror our interactions in the physical world. The common way of reading electronic documents is to open a window or browser on a computer and to use keyboard and mouse to navigate through the document. This differs from our actions in the physical world where we have a paper document and we move our head or eyes to conveniently read or look at interesting parts of the document.

"Scanning" the human user while reading an electronic document enables us to create an interface for reading electronic documents which appears more like reading documents in the real world. The "Perceptual Browser" behaves mostly like a standard web browser, except that the scrolling of the content is controlled by user's head movements. As the user moves his or her head downward from a "neutral" position, the content in the browser window starts to scroll up. The more the head is tilted downwards, the faster the scrolling. The symmetrical behavior occurs when looking upwards (Figure 5).

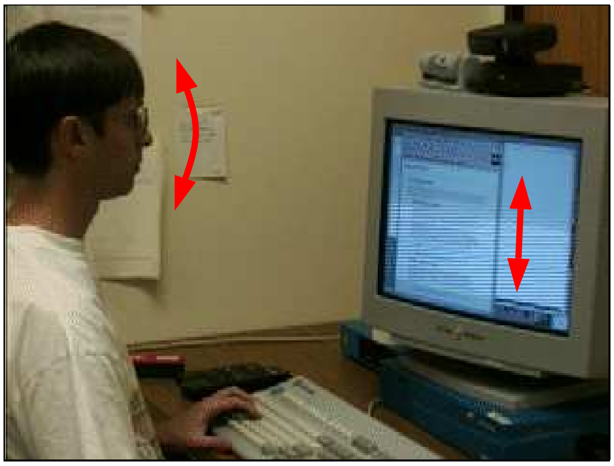


Figure 5: The Perceptual Browser: As the human user moves his head the content in the browser is scrolled according to the observed head movements.

The input to the Perceptual Browser is provided by a video camera (see Figure 5) set on top of the workstation's monitor. The information about head movements is gathered by tracking a region on the user's face over time. Currently this region is chosen manually with a graphical interface tool. When tracking is initiated, the Perceptual Browser tracks the target region in real-time by estimating its translation in the image plane. Lowering or raising the head results in translations of facial features in the captured image sequence which, in turn, controls the presentation of the document. We assume for now that the user sits in a comfortable position in front of the computer and moves only the head to read the electronic document.

The real-time tracking of the selected face region is achieved using correlation matching (Coutaz et al., 1996). Cross-correlation operates by comparing a reference template to an image neighborhood at each position within a search region. The reference template is a small neighborhood of pixels in an image. The search region in the following image frame is estimated from the expected speed of the user's movements. As the processing time is an important issue in our application, the reference template and the search window are kept small. Focusing on the recovery of translational parameters of one facial feature only, our system is able to process between 25 and 30 frames per second on a SGI Indy workstation. This gives the Perceptual Browser a response time below 50 ms, and preliminary experiments indicate that it provides a comfortable form of interaction.

The Perceptual Browser is a first step towards more natural human-computer interaction. Time constraints and processing speed limit our application at the moment to track only the translational movements of one facial region. A more natural interaction will require the recovery of more complex human movements and a complete system will need to be able to automatically locate the head (using skin color) and select a region, or regions, to be tracked. However, our simple system illustrates the promise of seamless interaction between humans and physical or electronic documents in the Digital Office.

Conclusions and Further Work

As video cameras increase in resolution and decrease in price, their presence in the officeplace will become more common. What will they enable, how will we exploit them, and how will the workplace be improved? The Digital Office project at Xerox is exploring these questions. Here we have presented three Digital-Office applications that bridge the physical and electronic worlds of documents. These applications extend the traditional notion of document "scanning" to a broader notion of "scanning the world." In doing so, they are leading us to new document types and new forms of interacting with documents.

The presented applications are promising steps towards a Digital Office where physical and electronic documents coexist without barriers and are brought together without effort. However, much remains to be done. The Digital Office must be more aware of its human inhabitants. This will require extending the recognition and understanding of gestures, actions, and human behaviors. Additionally, based on the technologies described here, we envision a Digital Desktop in which documents are automatically scanned and their locations are tracked as they are moved. "Connecting" these physical documents with their electronic counterparts would allow us to augment the paper documents to make them easier, to print, find, and edit.

Technologies like the ZombieBoard are already becoming an integral part of the daily work practice at PARC. They serve as testbeds for new forms of human computer interaction based on hand-drawn diagram understanding and gesture recognition. They also generate new types of documents (e.g. a scanned whiteboard image) and we are only beginning to understand how people exploit these in their work.

References

Abowd, G., Atkeson, C., & Essa, I. (1997). Potential Applications of Perception in Future Computing Environments. In M. Turk (Ed.), *Workshop on Perceptual User Interfaces (PUI'97)*, Banff, Canada, pp. 24–25.

Black, M. J. & Jepson, A. D. (1998). Recognizing Temporal Trajectories using the Condensation Algorithm. In *International Conference on Automatic Face and Gesture Recognition*, Nara, Japan. to appear.

Bobick, A., Intille, S., Davis, J., Baird, F., Pinhanez, C., Campbell, L., Ivanov, Y., Schütte, A., & Wilson, A. (1996). The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. Technical Report 398, M.I.T. Media Laboratory Perceptual Computing Section.

Coutaz, J., Bérard, F., & Crowley, J. (1996). Coordination of perceptual processes for computer mediated communication. In *International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, pp. 106–111.

Ishii, H. & Ullmer, B. (1997). Tangible Bits: Towards Seamless Interfaces between People, Bits and Atoms. In *Proceedings of ACM CHI'97 Conference on Human Factors in Computing Systems*, Atlanta, GA.

Marr, D. (1976). Early Processing of Visual Information. *Phil. Trans. R. Soc. Lond. B 275*, 483–519.

O'Hara, K. & Sellen, A. J. (1997). A Comparison of Reading Paper and On-line Documents. In *Proceedings of ACM CHI'97 Conference on Human Factors in Computing Systems*, Atlanta, GA.

Pentland, A. P. (1996). Smart Rooms. *Scientific American 274*(4), 68–76.

Saund, E. (1996). Example Line Drawing Analysis for the ZombieBoard Diagrammatic User Interface. http://-www.parc.xerox.com/spl/members/saund/lda-example/-lda-example.html.

Sellen, A. J. & Harper, R. H. R. (1997). Paper as an Analytical Resource for the Design of New Technologies. In *Proceedings of ACM CHI'97 Conference on Human Factors in Computing Systems*, Atlanta, GA.

Stafford-Fraser, Q. (1996). BrightBoard: A video-augmented environment. In *Proc. of CHI'96*.

Szeliski, R. (1994). Image mosaicing for tele-reality applications. In *Second IEEE Workshop on Applications of Computer Vision*, Sarasota, Florida, pp. 44–53.

Taylor, M. J. & Zappala, A. (1997). Documents through Cameras. *submitted to Image and Vision Computing - special issue on Document Image Processing and Multimedia Environments*.

Turk, M. (Ed.) (1997). *Workshop on Perceptual User Interfaces (PUI'97)*, Banff, Canada.

Ullman, S. (1983). Visual Routines. *Cognition 18*, 97–159.

Waters, K., Rehg, J., Loughlin, M., Kang, S. B., & Terzopoulos, D. (1996). Visual Sensing of Humans for Active Public Interfaces. Technical Report CRL 96/5, Digital Equipment Corporation, Cambridge Research Lab.

Wellner, P. (1993). DigitalDesk. *Communications of the ACM 36*(7), 86–96.