Interaction via motion observation

M A Foyle[1] and R J McCrindle[2]

School of Systems Engineering, University of Reading,
Reading, UK

mfoyle@iee.org, *r.j.mccrindle@reading.ac.uk*

www.sse.reading.ac.uk

ABSTRACT

The main method of interacting with computers and consumer electronics has changed very little in the past 20 years. This paper describes the development of an exciting and novel Human Computer Interface (HCI) that has been developed to allow people to interact with computers in a visual manner. The system uses a standard computer web camera to watch the user and respond to movements made by the user's hand. As a result, the user is able to operate the computer, play games or even move a pointer by waving their hand in front of the camera. Due to the visual tracking aspect of the system, it is potentially suitable for disabled people whose condition may restrict their ability to use a standard computer mouse. Trials of the system have produced encouraging results, showing the system to have great potential as an input medium. The paper also discusses a set of applications developed for use with the system, including a game, and the implications such a system may have if introduced into everyday life.

1. INTRODUCTION

The invention of the electronic computer in the 20th Century has probably changed the way in which we live in more ways than any other invention. Whilst it may be ironic that Thomas Watson, chairman of IBM famously said in 1943 "I think there is a world market for maybe five computers", computers are now so widespread that considering a modern society without them is almost impossible.

Advances in chip fabrication have seen processors become smaller and faster, storage capacities have increased phenomenally, and we have seen the computer merge with other household devices, producing so called convergence devices. Yet today, over 20 years on from the first commercial graphical computer operating systems, the principal mechanisms for data input are still the humble keyboard and mouse. In fact, the modern computer mouse stems from research developed over 40 years ago, such as the Sketchpad (Sutherland 1963) and work undertaken at Stanford (Engelbart 1967) in the late 1960s.

The aim of this project was to develop a novel and exciting method of interacting with computers, which would take advantage of these technological advancements, and which would also aid disabled people whose condition may restrict their ability to use a standard computer mouse.

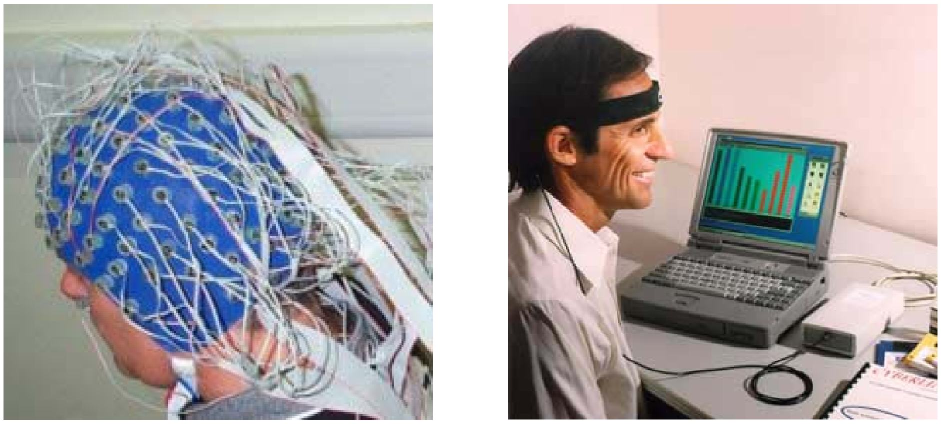
The system that was produced, known as the IMO (Interaction via Motion Observation) system, is based upon a standard computer web camera, generally used for video conferencing. In conjunction with the camera, a software system was developed that is used for interpreting the images captured by the camera, so that they can be processed and turned into useful input data. The system allows the user to interact with the computer in a visual way, without needing to make physical contact with an input device, or wear any additional tracking equipment (e.g. special gloves or head gear).

2. BACKGROUND

A majority of the research being conducted into developing new input devices is being performed in order to develop systems which can been used as assistive devices for disabled persons. In some of these cases, the computer provides the user's only channel of interaction with the real world. One example of this type of technology is the Brain Computer Interface (BCI), which has been developed for use by people with severe

©2004 ICDVRAT/University of Reading, UK; ISBN 07 049 11 44 2

motor disabilities. The use of electroencephalograms (EEGs) allows brain activity to be observed, and in turn the signals extracted can be used to operate a computer. The EEGs are captured from the person's brain by the attachment of an array of nodes to the outside of their head, as shown in Figure 1(a). Work carried out at the Wadsworth Center in New York has been used to drive both computers and simple prosthetic devices.



(a) *(b)*

Figure 1. *(a) A person's brain activity being measured through EEG signals, and (b) The Cyberlink Mindmouse*

In addition to current research projects there exist several commercially available devices, one example of which is the Cyberlink Mindmouse. The device consists of a headband that uses 3 sensors to detect electrical activity from the forehead and convert them into signals to drive the computer. Figure 1(b) illustrates a typical setup of the Mindmouse showing a user wearing the headband, which is in turn connected to the computer via a processing and control box. However, the problem with these devices is that they are very expensive and not readily available. For example, the Mindmouse retails at over $2000 (US), and devices using EEGs are typically experimental or hospital based.

In addition to these EEG-based interfaces, other methods of interaction are also being developed such as eye tracking. These systems typically use small cameras mounted onto the frames of spectacles, with the cameras positioned such that they can monitor the position of the user's pupil.

Our study revealed that in general the systems being developed are expensive, require special application of specific technology and generally require the user to wear some kind of monitoring device. These factors mean that devices based on these technologies are generally expensive to manufacture, and as a result are unlikely to be available to the mass market. It became clear that for any novel input device to be widely used, it would need to meet the following criteria:

- The hardware required for the device should be readily available, and relatively low cost.
- The device should not require any kind of special monitors or devices to be worn by the user.

These criteria became important influences in the development of this project, as it was felt that they were important in ensuring that the resultant input device would be both novel and accessible to all potential users.

In order to satisfy the first criterion, it was decided that using a piece of hardware which is available "off the shelf" would be desirable, particularly if the hardware already incorporates a computer connection interface. A study was conducted into the different types of computer peripherals available that might be suitable for use as a novel input device. Clearly there exist a large number of devices that are designed for user interaction in one form or another, such as gamepads and joysticks. However, these devices are related by concept to computer mice, and it was felt, do not represent a novel approach to user interaction.

The study highlighted that two "non-mechanical" devices exist which have the potential to be part of a novel input system. The first is the computer microphone, which is known for its use as a component in vocal recognition systems, and the second is the 'web camera' – a small, low resolution video camera used for video conferencing. Since voice recognition packages are commercially available, it was decided that the use of a web camera as part of an input device was both novel and interesting, since visual input systems are generally not very widespread.

3. HARDWARE

The main aim of the work was to implement the web camera as an input device in the following way. The camera is positioned such that it is able to monitor the movement of a user's hand. The user then moves their hand around in the visual area of the camera, and a continuous stream of images is captured. The software part of the system then processes the images in real time, finds the position of the user's hand and then uses this information for whatever purposed is selected – such as a custom interface, game, etc. Thus the system is able to see how the user moves, and respond accordingly, providing interaction through the observation of the user's motion. The system developed is known as Interaction via Motion Observation (IMO).

Traditionally web cameras are used for video conferencing applications, so using a web camera as a computer input device is quite a radical concept. Several choices are available in the computer web camera market, ranging from cheap, low-resolution cameras, right up to more expensive, high-resolution and high bandwidth cameras. For the purposes of this project, a mid-range Logitech Quickcam Pro 4000 camera was used (see Figure 2), costing approximately £50. This camera was chosen for several reasons – the camera is a popular and readily available camera, and the manufacturer provides a free Software Development Kit (SDK).

Figure 2. *Logitech Quickcam Pro 4000*

The camera was connected via a standard USB port to a PC, running Microsoft Windows XP. All software developed to work with the camera was designed for, and tested on the Windows XP Operating System.

4. SOFTWARE

The main task of the project was creating the system that would be able to analyse the video stream from the camera and extract the location of the user's hand, so that its position could be used to drive an interface. Whilst the system was designed primarily to be operated by the user's hand, it was decided that the image processing stage should not look specifically for a hand shaped object. Instead, it would be preferable if the movement of any hand-sized object, such as a foot, or even a small book, could operate the system. This approach would ensure that the system would be accessible and usable by a greater range of disabled people.

The detection of an object in the captured video stream was one of the most challenging parts of the project, especially due to the main issue of visual accuracy versus system response. For the system as a whole to work well, it needs to be accurate at detecting the user's hand within the captured stream of images. Many image-processing algorithms exist for this very purpose, and generally they perform a good job at finding the required object within an image. However, these algorithms can be very complex and computational (Davies 1996, Gonazalez 2002), and so may take a few seconds to process the image. Clearly this is acceptable when dealing with static images, but for a stream of images that are being used to operate a computer interface, this renders the interface unusable, regardless of the accuracy. In addition, it is important to bear in mind that the system is an input system, and so should have a low processing requirement, as it is likely to be used with other programs which may be more processor intensive. Thus it was important to ensure that the system was accurate enough to locate the user's hand, but without the lag caused by complex algorithms.

The first stage was to filter the captured images to leave behind only the objects in the scene that were of interest. Since the system is designed to be driven by a user's hand, the object that we are primarily interested in detecting is therefore a hand. However, forcing the system to look specifically for hand shaped objects could cause potential issues. For example, there is no such thing as an 'average hand', hands can vary significantly in shape and size, and if it were possible to operate the device with objects other than hands, such as feet, then the interface may be usable by a greater audience. Systems that have been developed to work specifically with hands generally build a model of the human hand (Lin 2000), use a template image (Lewis 1995) or look for particular tones of skin in the image (Lenman 2002).

Since the system is not to be restricted to just hand detection, the object recognition system needed to look for general object movement. This required knowing what has changed in the image, and then determining where the object is located and its direction of movement. Early prototypes of the system used image-differencing techniques to detect motion. However, it was found that detecting the direction of the largest movements was fairly computational, and as a result the system was unresponsive and difficult to interact with.

As a result of these attempts, it was decided that a better approach was to consider what objects were 'new' to the scene, and to assume that the biggest object was the object that was to be tracked. This approach has been used in many systems, such as in hand gesture recognition (Ziemlinski 2001), where a plain background and static lighting are assumed. With this approach, when a user places his or her hand in front of the camera, the system recognises this as a new object and derives its location. The technique employed for finding motion was to take a snapshot of the background scene and to then subtract it from each image captured.

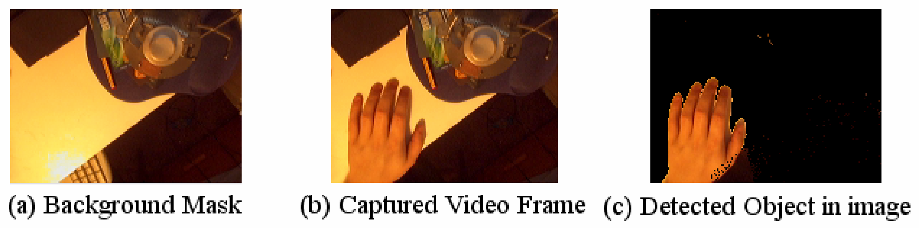


Figure 3. *A demonstration of the extraction system in operation*

This can be seen in Figure 3, which shows the background (image a), the captured image (image b) and the result of background subtraction (image c). In conjunction with a threshold technique that works relative to the intensity of the pixels in the captured image, it is possible to extract any large changes, such as hand motion. The system was tested with several different objects, including human hands, rulers and blocks of wood, and it was able to extract each of these objects relatively well. Whilst other objects, background changes and jolts to the camera could sometimes be detected in the processed image, these were "removed" by the next stage of the process, which determines the location of the user's hand (or whatever object is being used to control the system).

Once an image containing the objects has been obtained, it is then necessary to identify which of the clusters of pixels form the object of interest and the precise location of that object. Whilst techniques exist to perform this operation, due to the system response constraint, it was desirable to use a method that would work quickly, and which could be performed as the image was read from the camera. Initially, a image projection was performed, to count the frequency of pixels in each row and column. The theory behind this concept was that the object being used to control the system was generally large in comparison to any "noise" detected, and therefore the column and rows with the highest pixel counts would pass through the object. In addition, the system incorporates a weighting system, such that connected areas hold a higher value than non-connected areas, and as a result the largest object in the image will be located correctly. Tests applied during the development of the system proved this method to work successfully, although the system may be developed in the future to use more sophisticated techniques.

5. EXAMPLE APPLICATIONS

Once the basic IMO system had been created, and it was possible for the system to locate and track the movement of the user, the uses of the system in terms of computer interaction could be explored.

The first application developed was a simple program which allowed the on-screen mouse cursor to be moved around by the user moving their hand. The system performed well at this task, and it was possible to move the cursor in the same direction as the motion of the user's hands. Using the system as a replacement for the desktop mouse shows great potential, and would allow the system to be used with a large number of existing applications. Due to the difference between the resolution of the camera and the display, the motion was fairly jerky and as a result it was particularly hard to move the cursor to a precise location. However, as the specification of cameras improve and prices drop, it may be possible to improve this application in future work, by using a higher specification camera.

The second application was developed as an investigation into custom graphical input interfaces for the system. Instead of using a mouse pointer, the interface uses a concept of “zones”. Essentially the camera’s field of view is split into a grid of zones (6 for the test application described), and the system observes which zone the user’s hand is located in. The first application to use this approach split the viewable workspace of the camera into 6 zones each of an equal size. Thus, if an object was detected in the top left corner, it would be classified as zone 2. Figure 4 shows an example illustration of how the workspace is divided.

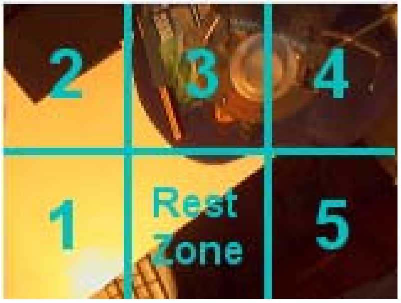
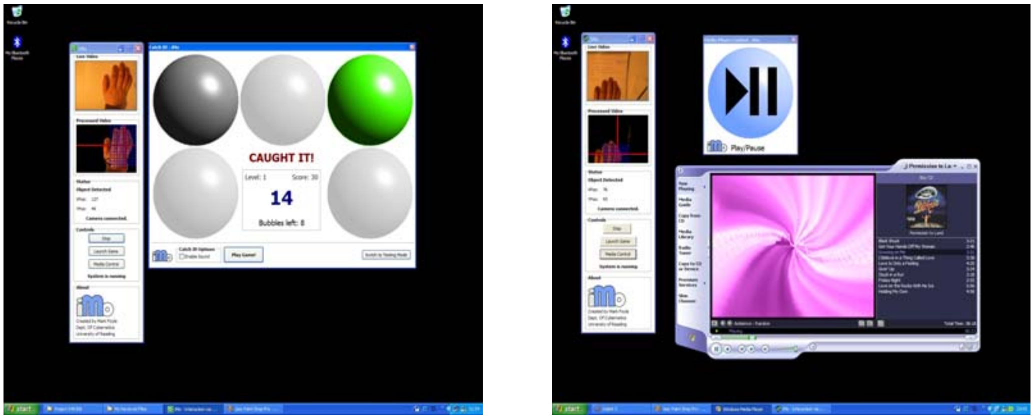


Figure 4. *How the camera workspace is divided into ‘Zones’*

An onscreen interface shows a set of objects, one for each of the camera’s visual zones. Thus if the user’s hand is located in the top right zone, this is echoed on screen by the illumination of an object in the top right area of the interface. An example of this can be shown in Figure 5(a), which shows the user selecting the top right object, by placing their hand in the top right zone. This system has the benefit that it is easier to select an object than by positioning a mouse cursor over a specific part of the screen.

Since there is no visual equivalent of a mouse button press, to signify that the user wishes to activate a command, the user simply selects an object by moving into the corresponding zone, and then pausing for half of a second. Trials of the interface showed this system to work well, and over time a user could become quite competent with the interface and selecting objects.



(a) *(b)*

Figure 5. *(a) The demonstration “Catch It!” game and (b) The Media Control program, controlling Windows Media Player*

One of the first applications developed using this interface was a reaction game, in which the computer picks at random one of the objects, and the user has to select the object as quickly as possible. Once the user has selected the object, the computer then picks another at random, and so it continues. In addition, the system records the time taken for the user to respond to the changes. An example run of the system can be seen in Figure 5(a).

In addition to the game, an application was developed which allowed interaction with a Media Player application. This application, shown in Figure 5(b), allowed the user to skip backwards and forwards through music tracks on a CD, and to also play and pause the music. The application was developed as an example of how the system could be used for a non-computing application.

6. OBSERVATIONS

Initial trialling of the IMO system indicated that it could become quite tiring to operate the device when the camera is mounted on a desk aimed horizontally at the user, as is the standard set up configuration for a web camera. This was especially true if the user was trying to reach far points of the screen. Indeed, trials of other similar systems (Lenman 2002, Freeman 1995) have also shown this to be the case.

A solution to this problem was evolved from some consideration as to how the IMO system may be used in public environments in which there is generally a lot of background motion. Due to the way in which the object recognition system was implemented (to allow users to use objects other than hands as input), any significant motion observed by the camera is recognised. Thus if there is a lot of movement in the background behind the user, for example many people walking past, then the system may become confused and recognise the motion of people walking past as the intended movement of the user. The solution to this problem was to try and cut down the background noise, preferably using some kind of static background. Since it would not be feasible or desirable to have a static background (such as a curtain) mounted behind the user, it was decided that the best approach was to rotate the camera and orient it such that it pointed down onto a surface, such as a desk. This was implemented in the system by attaching the camera to the side of the computer monitor such that it pointed downwards, with its field of view encompassing the area of desktop in front of the monitor, as shown in Figure 6.

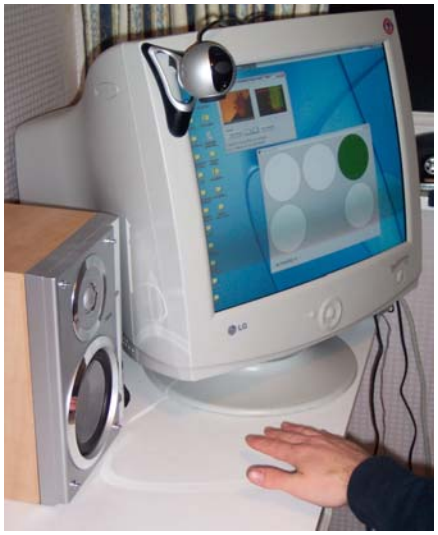


Figure 6. *Example setup of the IMO camera*

7. TESTING AND RESULTS

After the development of the applications, a series of tests were performed to provide a measure of the effectiveness of the input system. The tests were performed using an extension to the game application, such that the response time (the time taken for the user to react and select the appropriate bubble) could be recorded. A test sequence was generated, requiring the user to select 30 random bubbles in sequence. This sequence was saved so that it could be repeated for tests with additional users. The time between the successful selection of each bubble was recorded, allowing the data to be analysed later.

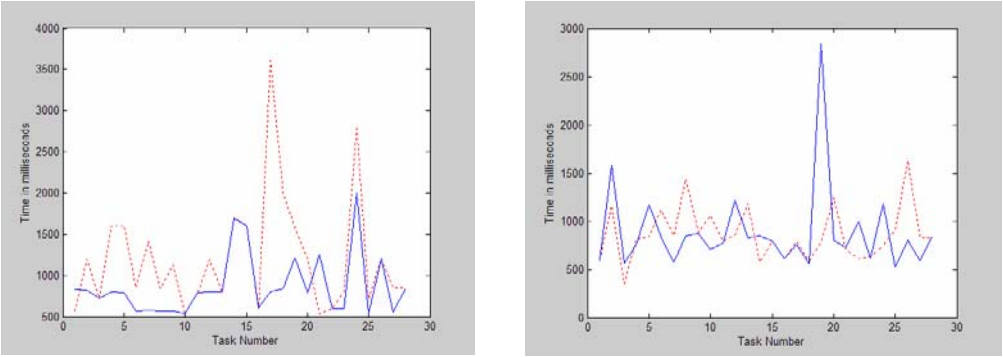
A total of 10 candidates undertook the tests, and each candidate performed the test twice, to determine whether users would become more proficient with the system as they used it more. Whilst the users were aware that they were to be tested twice, they were unaware that the second test would be identical, to ensure that they did not memorise the sequence for the second test. All times recorded were adjusted to take into account the 400 milliseconds required to trigger the target, and the different distances travelled. The final times therefore reflect the time taken by the user to respond to the change in target. Table 1 shows the average response times, per candidate, for both the first and second trials. In addition, it also shows the percentage change between the two trials.

Table 1. *Table of results*

Candidate ID	First trial – average response time (ms)	Second trial – average response time (ms)	Percentage change (%)
1	863.8	885.9	-2.6
2	1187.6	845.2	28.8
3	1295.9	1766.2	-36.3
4	1139.0	1585.8	-39.2
5	1210.6	868.5	28.
6	946.3	709.2	25.1
7	1578.7	938.7	40.5
8	1244.8	950.8	23.6
9	797.7	693.6	13.0
10	1128.5	974.3	13.7

The results obtained showed that in all cases the response time was generally low (around 1 second), and that in the majority of cases, users performed better in the second trial. This would suggest that the test candidates became more competent with the system over time, and as a result were able to adjust to this method of interaction. In addition, the relatively low response time for the candidates indicates that the system responds at an acceptable speed, and is therefore suitable for this type of activity. Figure 7(a) shows the results for candidate 5 for both the first trial (dotted line) and the second trial (solid line), illustrating the improvement of the user in the second trial.

The results table shows that in three cases, the candidate's performance actually deteriorated in the second trial. Upon inspection of the results, this was due to one or two abnormal results. Figure 7(b) shows an example of one of these occurrences, taken from the results from candidate 1. It transpires that most of these abnormal results were due to loss of concentration of the candidate or a glitch in the prototype system, and thus in conclusion the system performs well as an alternative input system. If these abnormal results are ignored, then it can be seen that the performance in the second trial was either the same or better.



(a)

(b)

Figure 7. *Graphs showing the response time for each task in the 30-task trial. (a) Shows the results for candidate 5 and (b) shows the results for candidate 1. In both instances, the dotted line corresponds to the first trial, the solid line corresponds to the second trial.*

8. FURTHER WORK

As discussed previously, since the system does not require any contact forces (e.g. no pushing or pulling), it may be particularly useful for use by disabled people. If the system could be adapted such that it was built into standard household appliances, then it may be possible to make many awkward to use household appliances more accessible. As an example, the system could be embedded into current light switches, so that it would be able to turn on and off the light in a room, without having to apply the usual pressure required for a standard push button switch. This would benefit those people who may find such a procedure troublesome, due to a disability.

The prototype user interface that has been developed shows a lot of potential as a platform for a range of applications including systems that are used for serving information, such as information kiosks. Such systems generally use expensive touch-screens, but with the IMO system these could be replaced with relatively cheap cameras.

As a proposal for further study, it would be interesting to develop the system for use in Computer Supported Collaborated Work (CSCW). Users could be placed at different computers around the world, each using the IMO system, and interact within a single application. However, unlike most collaborative work systems, this would allow the users to interact through direct physical movement.

One final area for investigation is the potential application of this system in the field of rehabilitation. As the system requires physical motion for input, it could be possible to devise a virtual task, which would require the user to make a series of specific movements in order to complete the task.

9. CONCLUSIONS

This paper has discussed the development of a novel human computer interaction device, and the potential such a system has for use by disabled people, and for novel methods of computer interaction.

Since the start of the project, similar systems have started to appear in the shops for the gaming market. Systems such as the Sony EyeToy allow users to interact with basic games on a Playstation games console. Generally, these systems have only been used for gaming purposes – no mainstream system yet exists which is aimed at general computer interaction for disabled people.

The results obtained from the tests have indicated that the system performs well. In order to obtain more detailed results, we aim to develop the test applications further, and to perform more thorough testing. We also aim to test the system with people possessing limited mobility, so that the suitability of the system for disabled people can be assessed.

Whilst the system developed is primarily aimed at being used as a computer input system, there is no reason to suggest that such a system could not be used for other purposes. With the rise in convergence devices in the home, it could be possible to see such technology in consumer appliances, such as televisions and home stereos, in the very near future.

10. REFERENCES

E R Davies (1996), ‘Machine Vision: Theory, Algorithms, Practicalities’, Second Edition, Academic Press

D C Engelbart (1967), ‘X-Y Position Indicator for a display system’, USA, Reference 3,541,541, Available: http://sloan.stanford.edu/MouseSite/Archive/patent/

W T Freeman & C D Weissman (1995), Mitsubishi Electric Research Labs, ‘Television Control by Hand Gestures’, IEEE International Workshop on Automatic Face and Gesture Recognition, Zurich

R C Gonzalez & R E Woods (2002), ‘Digital Image Processing’, Second Edition, Prentice Hall

S Lenman, L Bretzner, B Eiderbäck (2002), ‘Computer Vision based recognition of Hand Gestures for Human Computer Interaction’, Kungll Tekniska Högskolan

J P Lewis (1995), ‘Fast Normalized Cross-Correlation’, Vision Interface

J Lin, Y Wu, T S Huang (2000), ‘Modelling the constraints of Human Hand Motion’, University of Illinois

I Sutherland (1963), ‘Sketchpad: A Man-machine Graphical Communications System’, Ph.D. thesis, Massachusetts Institute of Technology

R Ziemlinski & C Hynes (2001), ‘Hand Gesture Recognition’, Available from http://www.via.cornell.edu/ece547/projects/g13/exp.htm