3

Stereo Vision for Unrestricted Human-Computer Interaction

Ross Eldridge and Heiko Rudolph
RMIT University
Melbourne,
Australia

1. Introduction

Since the advent of the electronic computer there have been constant developments in human-computer interaction. Researchers strive towards creating an input method that is natural, compelling and most of all effective.

The ideal computer interface would interpret natural human interactions in much the same way another human would. As humans we are used to giving instructions with our voice, body language and gestures: i.e. pointing at objects of interest, looking at things we interact with and using our hands to move objects. The ideal visual input device, the "holy grail" of human computer interfaces would be capable of accepting visual cues and gestures much as another human being would. Although this theoretical ideal is not feasible with current technologies it serves to guide research in the field.

Stereo vision has been researched for the purpose of studying human motion for many years (Aggarwal & Cai 1997; Cappozzo et al. 2005). Much effort has gone into achieving full body motion capture with multiple cameras, in order to overcome the need for artificial markers. This allows for unrestricted movement for the user, improves flexibility and overcomes occlusion problems with marker based approaches. However the processor time required to locate and track arbitrary objects in 3D has until recently been prohibitive for use in interactive scenarios.

Faster processors, memory, bus/interface speeds are helping overcome a significant constraint in using stereo vision for human-computer interaction (HCI), that of: producing real-time refresh rates and low latency response.

Now that interactive 3D human motion capture is becoming a reality, it's necessary to take stock of what can be achieved with this approach.

This chapter will not cover the theory behind stereo vision. Nor will we be considering technologies such as motion capture, as these are not intended for everyday computer usage. We are interested in applications that allow unencumbered interactions and operate without requiring the user to attach special devices to their body. We will be looking the current state of the art and where the next could be taken.

Open Access Database www.i-techonline.com

Source: Stereo Vision, Book edited by: Dr. Asim Bhatti,
ISBN 978-953-7619-22-0, pp. 372, November 2008, I-Tech, Vienna, Austria

2. History

2.1 Computer input devices

The majority of computer input methods are tactile in nature. From the humble switch to the joystick or touch pad, these are all based on the user physically interacting with a device, generally with the hand and fingers. This represents the most straightforward way for a computer to obtain information: The user presses a switch and the signal changes in the circuitry. The advantages of this method of interacting with a computer are:

- Input is essentially digital, i.e. 'unambiguous'.
- Low processing of input data when compared to more complex systems.
- Standard method applies across languages, cultures and computer systems.

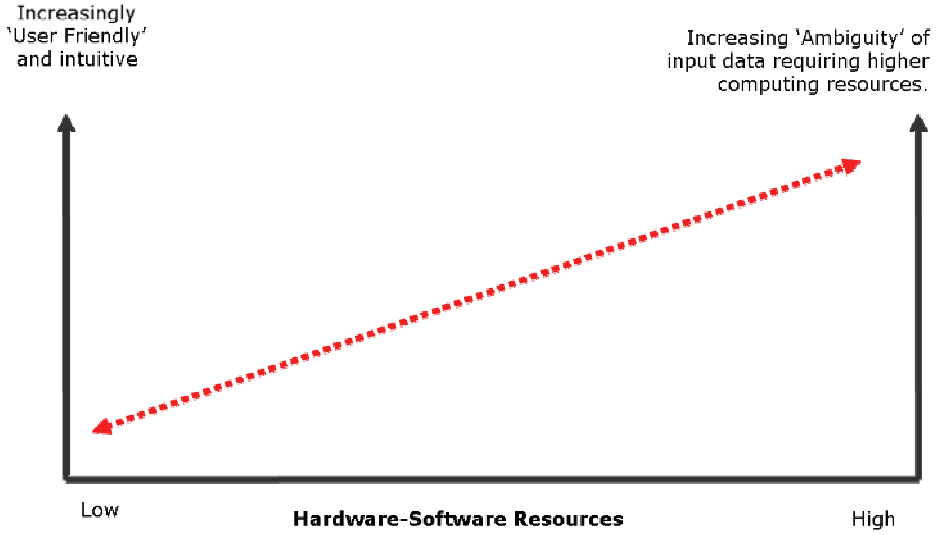


Fig. 1. Conceptual graph illustrating general relationship between 'ambiguity' of input, 'user friendliness' and computing resource requirements. Note that although show linearly, the relationship is likely not linear.

The standard input devices used with modern computers are the keyboard and mouse. Keyboards with physical keys that allow the input of natural language and computer commands, and a mouse to select and manipulate graphical user interface elements. The computer mouse has been available commercially since 1981, and the keyboard can be traced to the mechanical typewriter. Other devices are certainly used for many applications (e.g. touch screens, joysticks, voice input), however the core interaction with a computer is still via these two venerable devices.

There is good reason for the longevity of these interaction methods. The keyboard is one of the simplest ways for a human to enter information into a machine. The keyboard forces the human operator to enter unambiguous information in digital form which a computer can process with minimal effort. The even longer history of keyboard layouts within typewriters has meant that those familiar with traditional typewriters can make the transition easily.

The Graphical User Interface (GUI) and pointing devices were a radical and significant milestone in making personal computers accessible to the general population. The GUI reduced the need for specialized command line input via the keyboard, and reduced the amount of computer knowledge required by average users, while maintaining a clear and unambiguous input structure for the personal computer.
It should however be noted that GUI's require significant more processing power and hardware and software resources than its previous purely text-based interfaces. As a general principle: as the holy grail of human computer interfacing is approached, more and more computing resources are required.

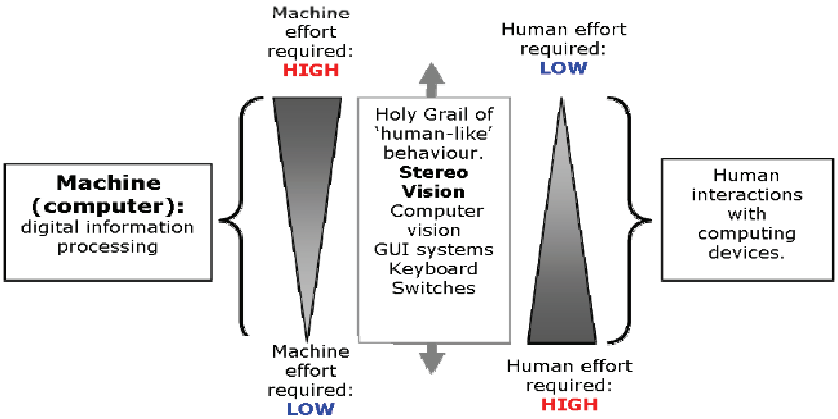


Fig. 2. Placing human computer interfaces into context regarding effort required by both sides.

This can be broken down into two directions:

1. Seen from the direction of human-computer interface, this principle can be understood in terms of degrees of 'ambiguity'. Switches and keyboards are a relative simple input for a machine to process, with GUI's becoming more resource demanding. As we approach computer vision systems the degrees of ambiguity increase and more resources are required to break down the input into information which can be digitally processed unambiguously.
2. Moving in the direction of machine to human communication: the issue is one of relevant feedback for the user. Information from the machine needs to be organized and presented to the user in ways which are easy for the user to process and understand and modify.

Within the past few years several new interaction technologies have gained commercial success, such as touch-screens which can detect multiple inputs simultaneously (Apple iPhone, Microsoft Surface) and wireless controls that use accelerometers to obtain free human movement as input (Nintendo Wii).

2.2 Stereo vision

The applications that have driven the development of computer stereo vision have varied greatly since its inception. The first major use was for mapping the topography of the land

by performing calculation disparity in satellite imagery (Barnard & Fischler 1982). Stereo vision later saw applications in human motion capture and allowed a computer to better animate humanoid models by capturing 3D human motion.
A further key area was robotics, especially in autonomous mobile systems. Stereo vision allows the robot to calculate its distance from objects in the environment: enabling it to calculate the dimensions of objects and spaces in the surroundings with greater accuracy.
More recent applications range from cameras in cars to judge distances to obstacles, to tracking people in open areas for surveillance (Cai & Aggarwal 1996).

2.3 What stereo vision brings to HCI

What does stereo vision bring to human computer interactions that can't be achieved using single camera approaches? Stereo vision can:

- Create active/inactive spaces for interaction. Just as we can lift our hands off a mouse to stop interaction, gestures and other interactions can be disabled based on distance from the camera.
- Help distinguish interactive parts of captured image (objects that are valid input) and background parts to be ignored.
- More accurate and reliable 3D position data than single camera (distance approximation) approaches.
- Better face tracking/matching by perceiving a disparity map of a person's face, giving another dimension to matching algorithms.

3. Computer vision for unrestricted human motion tracking

Unlike the previously mentioned types of human input, limb and joint tracking has been researched and used for many years before it became feasible for real-time applications. Early computer based human motion tracking used physical sensors and were primarily for bio-mechanics research. Other researchers began analyzing human movement in prerecorded video footage. Multiple cameras are used to track movements in 3D and to reduce occlusion problems, however this could not be achieved in real-time and required manual post-processing to correct errors (Sturman 1994).
Once real-time analysis became feasible, optical motion capture began seeing extensive use in computer graphics animations. The predominating system for this requires the person to wear a special suit covered in markers, as this improves accuracy and reduces computation requirements.
This required specialized and expensive hardware, not possible on a consumer level computer until more recently (Brown et al. 2003). Goncalves researched estimating 3D arm position without markers using a single camera, to reduce cost and make the system more practical (Goncalves et al. 1995), however computation speed limited the real-time application of such a system.
Within the past decade gesture, fingertip and arm tracking have reached a point where real-time interactions are possible. Various applications have since developed. These range from using 3D arm tracking as input to a robotic arm, achieving an untethered and natural remote interface (Verma et al. 2004), to using stereo vision for rehabilitation and human motion analysis (Cappozzo et al. 2005).

4. Relevant HCI considerations

4.1 Interaction models

A main facet of HCI in this area is how to structure the computer interactions to suit the new input device. Existing modes of interaction may not be best suited to unrestricted HCI, there are new areas where it will excel and different limitations. Will discuss viability of using stereo vision input for:

- Everyday use: Limitations of the standard WIMP (Window, Icon, Menu, Pointing device) model, new interaction models to achieve everyday computing tasks.
- Gaming: Achieve more life-like interactions by having the user physically perform a series of motions to interact with the game.
- Rehabilitation/training: Track limb movement in real-time for analysis and give immediate sense of achievement.
- Alternative input for the disabled: e.g. Gaze direction instead of mouse movement.

4.2 Direct input

Most vision based input devices can still be generalized as cursor devices, as we are interpreting an area of interest or intention, much like mouse input.

A major UI decision in vision based user interaction is the use of direct input. Direct input is any system where the interaction takes place in the same area as the response. For example, a touch pad (used in many portable computers) is an indirect device, as your movements control a virtual cursor on screen. A touch screen however is a direct input devices, as your finger is the cursor itself.

One of the key advantages to a direct input system is its ease of use. We are accustomed to directly interacting with the environment using our hands, so a direct input device is more natural. However, there are some disadvantages. Direct input methods tend to be less precise. A person's finger is larger than average mouse cursor, so interactions with a touchscreen will be less precise than, for instance, mouse input. The user's finger will also obscure part of the displayed image.

Direct input devices usually require a different GUI paradigm to the standard WIMP model, as they are two state devices. A cursor device generally consists of the following three states:

1. Hover
2. Active
3. Active and Moving

For a mouse: State 1 is moving the mouse around. State 2 is a mouse click. State 3 is click and drag. A number of alternate input devices lack one of these states, or have to infer the state from secondary information. A touch screen only has states 2 and 3, as there isn't a way to move the cursor without touching the screen. The only method to determine cursor position (touching) is also the only method to determine click activation.

It is often argued that this isn't relevant in direct input devices such as the touch screen, as your finger is the cursor and hence you don't need visual feedback to see where the cursor is. However there are many GUI conventions that stem from mouse usage which are impossible to achieve with a touch screen alone. (i.e. hovering over an area of interest to see a tool tip describing what the area does).

By tracking movement when the person isn't touching the surface we can create the Hover state in the standard cursor model, adding another degree of freedom. This enables a more

direct interpretation of standard mouse actions, given the three state interaction, and hence the user can use existing GUI functionality. It also allows for a virtual cursor when not touching the screen. This gives the user a constant sense of where their interactions will take place within the GUI environment, before they commit to activating UI elements.

5. Stereo vision for interactive surfaces

5.1. Interactive surfaces

An interactive surface generally consists of standard computer screen combined with some form of direct input. The most ubiquitous of these is the touchscreen kiosk, often used in shopping centers or for library catalogs.

Another type of interactive surface is the interactive tabletop environment, also known as an augmented desk.

Traditional interactive surface environments take one of three approaches. The oldest and most common approach is a touchscreen. These devices are usually "resistive" screens. They detect changes in electrical resistance across the screen to determine where the interaction will take place. They therefore require a conductive material (such as the human hand) for interaction, and output a single overall position for the cursor.

Another approach that has seen success recently is "Multi-touch". There are several technologies that can detect multiple touches on a screen surface. Jefferson Han's system uses infrared light that is internally reflected within the surface (Han 2005). When the user touches the surface this breaks the internal reflection and the IR light can be detected by a rear mounted camera.

This and other rear mounted camera approaches greatly increase the space required by such a system when compared to conductive technologies. SmartSkin (Rekimoto 2002) uses a capacitance based approach to achieve multiple input detection in a flat surface. Phillips Entertaible (Hollemans et al. 2006) achieves multiple finger interaction as well as basic shape matching with an LCD panel. It uses a proprietary method based on IR emitters and photo diodes.

The touch screen interface is more traditional. Interaction takes place when the user physically touches the screen. Methods of interaction include clicking and dragging objects, similar to the standard WIMP module. These may be expanded into gesture based interaction, and multiple touch allows for more dynamic gestures: rotation, scaling, etc.

Vision based tracking in these environments tend to not require physical contact. The user does not need to touch the screen as interaction is detected by a camera above or behind the screen, and clicks can be trigger by having the user 'dwell' on a particular spot or using gestures. The University of Tokyo have developed a natural hand tracking augmented desk (Oka et al. 2002) using a far-infrared camera to obtain clear hand silhouette, and user interaction is determined by finger gestures.

The use of stereo vision systems can bring the benefits of a multi-touch system and vision based tracking approaches. In the case of the work by Wilson, A.D (Wilson 2004), a stereo camera rig is placed behind the transparent screen. Distance data is used to determine when the user is interacting with the screen (effectively a multiple-touch screen), but also to morph images of the user for video conferencing. This new dimension can also expand the

interactive surface area into an interactive box above the screen, with many possibilities for interaction.

5.2 Depth and its usefulness

The third dimension (e.g. distance from the camera/screen) is generally used to determine whether an object is touching the screen. As such this data is effectively thresholded and binary, touching or not touching.

Researchers at the University of Sydney developed an augmented desk system using stereo cameras (Song & Takatsuka 2005). By thresholding the depth data of their fingertip tracking system just above screen level they achieved a virtual button press, which acts in much the same way as a touchscreen. An earlier system placed the cameras at non-aligned positions (above and to one side) to achieve 3D position tracking of the human arm (Leibe et al. 2000).

This effect has also been achieved with a single camera and a touch screen interface. Dohse uses one camera below the table for touch (using infrared reflection), one above for above-screen hand movement (Dohse et al. 2008). Whilst this system uses two cameras, they aren't able to see the same objects and hence can't determine 3D position.

There are many possibilities to enhance user interaction by using depth data in a more granular way. A relevant use of this is seen in (Franco 2004), a gesture based infrared instrument that uses height data to control musical effects.

Another possibility is to track a user's finger and obtain a 3D vector to determine where the user is pointing. Rather than reaching across a large surface to interact, the user need only point with a finger. Researchers have used stereo cameras to extract 3D pointing gesture from a dense disparity map (Jojic et al. 2000; Demirdjian & Darrell 2002).

The 3D information may also be useful in a non-immediate sense. The paths that a user takes when using the interface can be analyzed after the fact to examine how people use the device, in order to improve the system. For example, if it is found that users only ever interact within a subset of the area that the system covers, the system may be altered for increased accuracy or speed.

7. Challenges facing stereo vision in HCI

Over the last few years many new human-computer interaction devices have gained commercial success. (e.g. Nintendo's Wii gaming console, Apple's iPhone) The concepts that we have covered here are making headway in mainstream computing applications, such as multi-touch and gesture based interaction. However the success of stereo vision based approaches has been limited.

There has yet to be a 'killer app' for stereo vision computer interfaces (i.e. an application that would accelerate mass adoption of the technology). It has seen success in niche markets and with more commercial, industrial and military systems. And whilst there are low cost devices that can achieve stereo vision available to the home user, there isn't a major drive for use in the consumer market.

The true advantage of stereo vision over other technologies in this field is the acquisition of real-time and reasonably accurate 3D position data for arbitrary objects in the physical environment. In order to achieve this, the cameras much be able to see the subject in

question. This generally means placing the cameras in highly visible and often awkward positions (e.g. pointing down from above a desk). This makes it nearly impossible to integrate into a self-contained device. The system requires setup and some form of calibration.

Another issue is privacy and comfort of the users around active video capture. In order to obtain 3D arm position or determine gaze direction, the computer needs to 'see' the end user. People may have become used to computer's seeing them through a webcam, but this would be an always-on approach: all interactions effectively require video 'surveillance' after a fashion.

Infrared based multi-touch systems (Han 2005) use cameras to detect IR reflections, however these are placed behind the screen: they can't see the end user. Nintendo's Wii system has a camera (within the game controller), however it can only see IR hotspots and is generally pointed away from the user. This is a fundamental issue with the technology, and indeed any technology that wishes to achieve the goal of untethered and unrestricted human movement input.

8. Conclusion

Human computer interfaces have come long way in recent years, but the goal of a computer interpreting unrestricted human movement remains elusive. The use of stereo vision in this field has enabled the development of systems that begin to approach this goal. As computer technology advances we come ever closer to a system that can react to the ambiguities of human movement in real-time.

In the foreseeable future stereo computer vision is not likely to replace the keyboard or mouse. There is at this point no clearly identifiable mass market for stereo vision in the human computer interaction field.

However in this regards stereo vision may be in a similar position to personal computing: in the late 1980's, and early 1990's. In that period personal computers were a somewhat specialized technology with a slowly growing application. The main driving force for personal computers outside research and certain business applications were video games for young people.

Similarly it is still the games industry, and the entertainment industry which drives the adoption of new human computer input systems in general and computer vision systems in particular. Thus it is quite possible that stereo vision is at this point in a similar position personal computers once were.

Certainly the foundations for any expansion into computer vision are increased processing power and software intelligence to drive the systems. Both are proceeding at a rapid pace.

However even as stereo vision advances it is would most likely be used in applications where its strengths are required. Stereo computer vision is at this point a technology in its pre-teenage years, which given the advances in computing, can be expected to mature rapidly from here on.

10. References

Aggarwal, J.K. & Cai, Q. (1997). Human motion analysis: a review. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, (pp. 90-102)

Barnard, S.T. & Fischler, M.A. (1982). Computational Stereo. *ACM Comput. Surv.*, 14, 4, (pp. 553-572)

Brown, M.Z., Burschka, D. & Hager, G.D. (2003). Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 8, (pp. 993-1008)

Cai, Q. & Aggarwal, J.K. (1996). Tracking human motion using multiple cameras. In *Pattern Recognition, 1996., Proceedings of the 13th International Conference on*, (pp. 68-72 vol.3)

Cappozzo, A. et al. (2005). Human movement analysis using stereophotogrammetry: Part 1: theoretical background. *Gait & Posture*, 21, 2, (pp. 186-196),

Demirdjian, D. & Darrell, T. (2002). 3-D articulated pose tracking for untethered diectic reference. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, (pp. 267-272)

Dohse, K.C. et al. (2008). Enhancing Multi-user Interaction with Multi-touch Tabletop Displays Using Hand Tracking. In *Advances in Computer-Human Interaction, 2008 First International Conference on*, (p. 297—302)

Franco, I. (2004). The AirStick: a free-gesture controller using infrared sensing. In *NIME '05: Proceedings of the 2005 conference on New interfaces for musical expression*, Singapore, Singapore: National University of Singapore, (pp. 248-249)

Goncalves, L. et al. (1995). Monocular tracking of the human arm in 3D. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, (pp. 764-770)

Han, J.Y. (2005). Low-cost multi-touch sensing through frustrated total internal reflection. In *UIST '05: Proceedings of the 18th annual ACM symposium on User interface software and technology*, New York, NY, USA: ACM Press, (pp. 115-118)

Hollemans, G. et al. (2006). Entertaible: Multi-user multi-object concurrent input. *Adjunct Proceedings of UIST*, 6, (pp. 55-56)

Jojic, N. et al. (2000). Detection and Estimation of Pointing Gestures in Dense Disparity Maps. *IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France*

Leibe, B. et al. (2000). The Perceptive Workbench: toward spontaneous and natural interaction in semi-immersive virtual environments. In *Virtual Reality, 2000. Proceedings. IEEE*, (pp. 13-20)

Oka, K., Sato, Y. & Koike, H. (2002). Real-time fingertip tracking and gesture recognition. *Computer Graphics and Applications, IEEE*, 22, 6, (pp. 64-71)

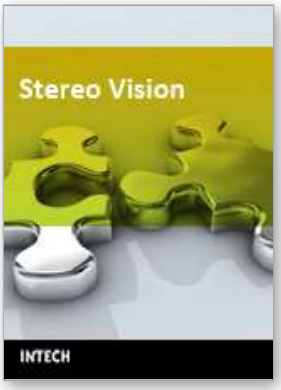
Rekimoto, J. (2002). SmartSkin: an infrastructure for freehand manipulation on interactive surfaces. *Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves*, (pp. 113-120)

Song, L. & Takatsuka, M. (2005). Real-time 3D finger pointing for an augmented desk . In *Proceedings of the Sixth Australasian conference on User interface - Volume 40* , Newcastle, Australia : Australian Computer Society, Inc., (pp. 99-108)

Sturman, D.J. (1994). A Brief History of Motion Capture for Computer Character Animation. *SIGGRAPH 94, Character Motion Systems, Course notes*

Verma, S., Kofman, J. & Wu, X. (2004). Application of markerless image-based arm tracking to robot-manipulator teleoperation. In *Computer and Robot Vision, 2004. Proceedings. First Canadian Conference on*, (pp. 201-208)

Wilson, A.D. (2004). TouchLight: an imaging touch screen and display for gesture-based interaction. *Proceedings of the 6th international conference on Multimodal interfaces*, (pp. 69-76)



Stereo Vision

Edited by Asim Bhatti

ISBN 978-953-7619-22-0
Hard cover, 372 pages
Publisher InTech
Published online 01, November, 2008
Published in print edition November, 2008

The book comprehensively covers almost all aspects of stereo vision. In addition reader can find topics from defining knowledge gaps to the state of the art algorithms as well as current application trends of stereo vision to the development of intelligent hardware modules and smart cameras. It would not be an exaggeration if this book is considered to be one of the most comprehensive books published in reference to the current research in the field of stereo vision. Research topics covered in this book makes it equally essential and important for students and early career researchers as well as senior academics linked with computer vision.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Ross Eldridge and Heiko Rudolph (2008). Stereo Vision for Unrestricted Human-Computer Interaction, Stereo Vision, Asim Bhatti (Ed.), ISBN: 978-953-7619-22-0, InTech, Available from: http://www.intechopen.com/books/stereo_vision/stereo_vision_for_unrestricted_human-computer_interaction

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2008 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.