1
Introduction

The author's original intention, a couple of years ago, was to develop a kind of an intuitive, dataglove-based interface for Computer-Aided Design (CAD) applications. The idea was to interact with 3D geometry *directly*, i.e. using hands, just as we interact with physical 3D objects in the real world. However, while undoubtedly there exist application areas where datagloves are the best and optimal choice, they (datagloves, sometimes also called *cybergloves*) continue to be expensive, invasive, and essentially a niche technology.

Subsequently, it was in 2006 when the author came into contact with advanced computer-vision techniques (through the graduate-level course "Visão Computacional e Realidade Aumentada", held by professor Marcelo Gattass), that the idea to use vision-based hand tracking for the same type of a CAD interface, instead of using datagloves, was born.

However, using computer vision (CV) techniques to detect and track human hands is difficult. Although in recent years many advances in the field of vision-based hand tracking (and tracking of articulated structures in general) have taken place, a lot of theoretical and practical problems remain. For example, while detecting an arbitrarily oriented and illuminated human hand in an digital image reliably, robustly and quickly is a difficult problem, it is simply a non-issue with datagloves. Furthermore, tracking a hand using CV techniques is even harder, while datagloves can do the same tracking with almost exact precision. On the other hand, CV techniques adopted in this work don't require the user to don any device, and offer a complete freedom of movements.

That said, and as I have already mentioned at the beginning of this preface, the central theme and objective of this dissertation is actually the *manipulation of 3D objects using hands*, and computer vision is "merely" our vehicle to achieve that end. Put differently, the main motivation for doing this work was to try to map manipulation operations as we know them in the real, physical world (when we manipulate real objects), to a set of corresponding manipulation operations in the virtual environments, so that we can manipulate virtual 3D objects.

1.1 Historical context

As of time of writing (second half of the first decade of the 21st century), the field of computing is as alive and active as ever. Just in the last couple of years, we've witnessed the meteoric rise of Google, Inc. ("organizing the world's information and making it universally accessible and useful"), the widest possible dissemination of mobile computing platforms, and last but not least a slow but steady switch (actually, a sea change) to many-core and multi-core computing, which will soon have deep repercutions on how we conceptualize, develop and use computer programs.

Taking a look at the hardware interfaces of our good old standard Personal Computer (PC), apparently we cannot find the same level of dynamism. The peripherals we use to interact with our PCs practically haven't changed since the original IBM PC was ushered into the IT scene in 1981 — granted the data buses have become wider and faster, processors tick at 3 GHz instead of at 4.77 MHz, RAM and disk sizes are much more plentiful and the operating systems that drive this hardware are much more complex and capable — but the keyboard stayed almost the same as the one featured by the original IBM PC, and the mouse is conceptually equivalent to the one Douglas Engelbart invented and perfected in the 1960s. Furthermore, the technologies that were predicted to revolutionize human-computer interfaces, like for example speech recognition, haven't come to realize their full potential.

Yet, exactly in the last couple of years preceding this work, we have witnessed some interesting developments in the field of HCI[1] (specifically, launches of commercial products, which were of course preceeded by years and even decades of academic and corporate research); characteristically, all these developments try to provide more *natural* ways for users to interface with computers, like for example *touch* and *hand gestures*.

For starters, in 2007 Microsoft Inc. introduced[2] the *Microsoft Surface*^{tm}, a computerized table whose tabletop (a touch-sensitive display) can detect user's touches and recognize physical objects by means of five infra-red cameras situated beneath the surface. The device itself is built around the principles of **direct interaction** (the user manipulates virtual objects using hands and/or fingers), and **multi-touch interaction** (the user can apply one, several or all his fingers to interact with the device; also, many users can interact with the device at the same time).

Further, also in 2007, Apple Inc. launched a commercially successful

[1] HCI is the acronym for "Human-Computer Interaction".
[2] www.microsoft.com/surface/

product line which includes the iPod[3] and the iPhone[4] whose interfaces also feature a touchscreen able to detect touching and dragging finger gestures. Similarly to Microsoft's Surface, it's possible to stretch a photo by placing two fingers on two opposite corners of the image, then spreading the fingers thus enlarging or shrinking the image.

In an somewhat earlier development, Nintendo Inc. introduced in 2005 the gaming console *Wii*[5] and the associated *Wiimote*, which actually acts as an accelerometer and 3D position tracker; in conjuction with the associated software, this system can recognize certain hand gestures. This way it's possible, for example, to simulate hitting the tennis ball by doing the equivalent, *natural* hand movement and gesture.

Finally, we have earlier devices that also support direct manipulation, like MERL's DiamondTouch[6] [1], a multi-user, touch-and-gesture-activated screen for supporting small group collaboration, and the Responsive Workbench [2], a virtual work environment.

As a conclusion, there seems to exist a certain momentum towards more natural and intuitive ways to interact with computers, although only time will tell how successful this push for more intuitive and "natural" interfaces will ultimately be.

1.2 The motivation

The motivation for this work is simply to try to use our own hands to interact with 3D geometry, and also due to the author's deep insatisfaction with the current state of affairs in the field of 3D user interfaces. While the mouse (and a number of specialized devices like 3D mice, SpaceBall^{TM} and SpaceNavigator^{TM} by 3Dconnexion Inc., and similar devices), have proved their value in various 3D application contexts along the last *several* decades, this work is an attempt to offer an arguably more *natural* and *intuitive* method to interact with a 3D computer model, especially having certain types of user communities in mind (for example, architectural and graphic designers, sculptors, and artists in general).

[3] www.apple.com/ipod/
[4] www.apple.com/iphone/
[5] www.nintendo.com/wii/
[6] www.merl.com/projects/DiamondTouch/

1.3 The scope covered by this dissertation

The title of this MSc dissertation is **Direct spatial manipulation of virtual 3D objects using vision-based tracking and gesture recognition of unmarked hands**, which implies the following:

- **Direct spatial manipulation** — the expression "direct manipulation" (without the adjective "spatial" or "3D") refers to the *technique of making user interfaces more intuitive by representing the objects of interest visually and letting the user manipulate them directly with an input device like a mouse* [3]. Consequently, "direct spatial manipulation" or "direct 3D manipulation" can be considered to be a specialization of direct manipulation, in the following way:
  - we deal with the manipulation of virtual 3D geometric objects, as dinstinguished for example from manipulation of 2D icons.
  - we use free-form hand movements for spatial input, and
  - there is a minimal (or equal to zero) spatial displacement between the user's physical hand (and of its virtual representation) and the manipulated virtual 3D object.
- **Virtual 3D objects** — here the fact that we manipulate virtual (computational) 3D models, instead of physical objects, is emphasized.
- **Vision-based tracking of (unmarked) hands** — "tracking", in our context, refers to the process and techniques for future position prediction of a target object. "Hand tracking", therefore, refers to the tracking of human hand. Consequently, tracking of *unmarked* (i.e. uninstrumented, unadorned, bare) hands refers to hand tracking which does not try to instrument the hands in any way, like for example, by placing a marker on the hand. Finally, we perform tracking using passive computer vision techniques, that is, we do not consider active computer vision techniques like for example projecting a pattern onto the object of interest.
- **Vision-based gesture recognition of (unmarked) hands** — again, we use use passive computer vision techniques to recognize various hand gestures (in our case, static gestures, that is, views of hand postures) which modulate the movements of human hands in the workspace.

Therefore, according to the definitions above, this dissertation describes an approach to direct spatial manipulation of virtual 3D objects, using passive computer vision techniques to detect and track user's hands in the workspace, as well as recognize hand gestures made by the user in the workspace.

1.4 The structure of this MSc thesis

This MSc thesis consists of two parts: the first part describes related work, and the second part describes the prototype software application.

The first part, **Related Work**, describes prior work done in all the areas that are relevant to this MSc thesis, and consists of the following chapters:

- Chapter 2 describes the anatomy and biomechanical properties of the human hand, as well as gives an overview of existing biomechanical models of the human hand.
- Chapter 3 describes one-handed and two-handed gestures for manipulation, as well as gives the theoretical framework for hand gestures and hand gesture recognition.
- Chapter 4 describes interaction techniques for direct 3D manipulation.
- Chapter 5 gives an overview of computer vision topics for hand detection, recognition and tracking.
- Finally, Appendix A gives a timeline of research in manipulation of virtual geometrical objects.

The second part, **Prototype Application**, consists of the following chapters and appendices:

- Chapter 6 describes all the aspects of the prototype application.
- Chapter 7 gives conclusions and future work.
- Appendix B describes the Viola-Jones detection method, which is used in the prototype for hand detection.
- Appendix C describes KLT features, which are used in the prototype for hand tracking.
- Appendix D describes the Hartley-Sturm triangulation method, which is used in the prototype to perform 3D reconstruction of the tracked hand's position in workspace.