Sensor Design and Interaction Techniques for Gestural Input to Smart Glasses and Mobile Devices

Andrea Colaço* MIT Media Lab

Figure 1. 3D Sensor Design and Interaction Scenarios: Mime is our compact, low power sensor for 3D gestural control of smart glasses. It comprises an active illumination three-pixel time-of-flight (TOF) sensor, and a 2D RGB camera. Mime's real-time signal processing pipeline enables fast and precise 3D gestural control in cluttered environments, and in strong daylight and dynamic light conditions.

ABSTRACT

Touchscreen interfaces for small display devices have several limitations: the act of touching the screen occludes the display, interface elements like keyboards consume precious display real estate, and even simple tasks like document navigation which the user performs effortlessly using a mouse and keyboard require repeated actions like pinch-and-zoom with touch input. More recently, smart glasses with limited or no touch input are starting to emerge commercially. However, the primary input to these systems has been voice.

In this paper, we explore the space around the device as a means of touchless gestural input to devices with small or no displays. Capturing gestural input in the surrounding volume requires sensing the human hand. To achieve gestural input we have built Mime [3] – a compact, low-power 3D sensor for short-range gestural control of small display devices. Our sensor is based on a novel signal processing pipeline and is built using standard off-the-shelf components. Using Mime we demonstrated a variety of application scenarios including 3D spatial input using close-range gestures, gaming, on-the-move interaction, and operation in cluttered environments and in broad daylight conditions. In my thesis, I will continue to extend sensor capabilities to support new interaction styles.

Author Keywords

Hand tracking, 3D sensing, gesture sensing, time-of-flight imaging, mobile, head mounted displays, wearables.

ACM Classification Keywords

H5.2 [Information interfaces and presentation]: User Interfaces. - Interaction Styles.

*acolaco@mit.edu

UIST'13, Oct 08-11 2013, St Andrews, United Kingdom
ACM 978-1-4503-2406-9/13/10.
http://dx.doi.org/10.1145/2508468.2508474

INTRODUCTION

Input to mobile devices is a rich design opportunity because of their pervasive use. Modern mobile devices exist in several different forms each with its own unique interface. These smart devices have become our best digital swiss army tool; users can perform a wide variety of computing and communication tasks through these devices.

An input technology intended for mobile device control and interaction ideally possesses the following characteristics:

- **Technical**: High accuracy, low power, low latency, small size, daylight insensitivity, and robust performance in cluttered, noisy and fast changing environments.
- **User experience**: Interaction experience should be intuitive and should not induce fatigue upon prolonged use.
- **User convenience**: The sensor should be embedded within the device to enable unencumbered user interaction. The user should not be required to wear markers [11] or external sensors [7, 13] or carry additional touch pads.

Currently, touch-screen input is the primary interaction modality for smart devices which require a display – no matter how small. For wearables, such as smart glasses, voice is the input of choice; these upcoming devices do not have an touch-screen display which can double as input device.

Flat screen touch interfaces do not fully take advantage of human dexterity and has its own set of limitations – it requires the user to be in constant contact with the device, touching the screen for input occludes the display, and even simple tasks like menu navigation require tedious, repetitive actions. Equipping users with better input tools for more complex and visually demanding tasks will be important in enabling new applications and making the interaction experience more intuitive and efficient.

We explore the use of 3D volume around the device as a means of interaction via touchless 3D hand gestures. The space around the device is unused and often unoccluded. Close range 3D gesture sensing introduces a new interaction paradigm which goes beyond touch and alleviates its current limitations. The implementation of 3D gestural input requires depth sensors in mobile and wearable devices. However,

existing state-of-the-art 3D sensors cannot be embedded in mobile platforms because of their prohibitive power requirements, bulky form factor, and hardware footprint.

As a part of my thesis, I built Mime, a compact and low-power sensor for volumetric interaction techniques for mobile devices via touchless 3D gestures. Mime operates using a real-time signal processing framework that combines a three-pixel time-of-flight (TOF) module with an RGB camera module. The use of TOF enables 3D hand-motion tracking, while the fusion with an RGB camera provides finer gesture identification. Our Mime hardware prototype achieves fast and accurate 3D gesture tracking. Compared with state-of-the-art 3D sensors like TOF cameras, the Microsoft Kinect and the Leap Motion Controller, Mime offers several key advantages for mobile applications and mobile use cases: very small size, daylight insensitivity, and low power consumption. As shown in Figure 1 (left), Mime is built using standard, low-cost opto-electronic components and promises to be an inexpensive technology that can either be a peripheral component or be embedded in the mobile device, thereby eliminating the need for markers, hand-worn sensors, or mobile controllers.

PRIOR ART ON 3D GESTURAL INTERACTION

The use of gestures for human-computer interaction is currently an active area. From a user-experience viewpoint, gestural control using 3D cameras has been demonstrated to be an intuitive, robust, and widely-popular input mechanism in gaming applications (for example, Microsoft's Kinect Sensor. [1]). Several new technologies, like the Leap Motion Controller[2] and compact TOF cameras[3], are still being explored for gesture-controlled interfaces in the context of personal computing spaces. Recent user studies have recently demonstrated that 3D gesture input is at least as effective as touch input for mobile devices [6]. This finding raises interesting possibilities for smart wearables, like head mounted displays (HMD), which lack a dominant input interface like touch. Several different forms of gestural input exists:

Close-to-body interactions

Several researchers have showcased the design of on-body and close-to-body gestural interfaces using a 3D camera mounted in different regions close to the user; Omnivision [5] mounts the 3D camera on the shoulder of the user, and Shoesense [1] mounts a 3D camera on a user's shoe facing up. Another class of examples [4, 2] uses motion capture systems such as the Vicon to create applications that track free-form hand movement, position, and orientation. The gesture pendant [12] is also an early example of gesture recognition in a wearable device. All the above examples clearly demonstrate the need for 3D gesture sensing in scenarios where the user is mobile, corroborating our vision for building sensors that can eventually be embedded in mobile and wearable devices.

Gestural control with 2D cameras

Computer vision techniques allow the use of embedded 2D cameras to recognize hand gestures [10]. These gestures have been widely used for unencumbered line-of-sight interaction with mobile devices. Standard RGB-image based gesture recognition suffers from several technical problems: it is not robust in cluttered environments, is computationally complex for mobile processors, and supports a small dictionary of simple motion-cued gestures like waving. RGB image-based gesture recognition can be made more precise, robust and generic at the expense of using of additional elements like color markers [9] or infrared trackers [12].

Gestural control with 3D cameras:

Stereo cameras, TOF cameras [8] and structured light sensors such as Microsoft's Kinect capture 3D scene structure in a depth map, which is then processed to identify gestures. 3D cameras retain all the advantages of standard 2D cameras but offer better performance and support complex gestures through acquiring position and motion cues in real time. The main limitations in using 3D sensors for mobile devices and smart wearables are high power requirements, bulky form factor, heat dissipation, and computational complexity.

In addition to the aforementioned gesture recognition systems, wearable sensors such as Digits [7] have also been proposed for precise full-hand tracking. Similar to hand-held controllers and voice input, wearable sensors allow non-line-of-sight user interaction.

Comparison of Mime with Prior Art

Mime compares favorably on accuracy vs. power trade-offs with other real-time sensors useful for mobile device interaction that are compact and enable unencumbered, free-form, interaction. Compared with these other sensing modalities, Mime offers fast and precise 3D gesture sensing at low-power. Also, as the display fidelity increases and size of mobile devices decreases, we expect that unencumbered input will be an important user interface design consideration for consumer applications and daily use cases. Mime also compares favorably with other input techniques on performance vs. encumbrance axes. Along with other 3D gestural control techniques, Mime offers high performance and unencumbered interaction, with the added advantage of being embedded in the bezel of the mobile device.

MIME: SENSOR OVERVIEW

In this section, we briefly overview Mime operation and highlight key technical distinctions from RGB cameras and depth sensors used in mobile device input and control. We describe the hardware implementation and performance of the existing system. The Mime hardware comprises two modules:

1. A low power, time-of-flight triangulation module built using a pulsed LED and a linear array of three photodiodes.
2. A standard 2D RGB camera.

Mime operates by first using TOF triangulation for accurately localizing the 3D hand position in the sensor field-of-view (FOV) (see Figure 2(a)). Once this region of interest (ROI) is identified, the corresponding RGB image patch is processed to obtain detailed hand gesture information using well-known computer vision techniques (see Figure 2(b)).

[1] `http://www.xbox.com/en-US/kinect`

[2] `https://www.leapmotion.com/product`

[3] PMD Camboard Nano. `https://www.pmdtec.com`

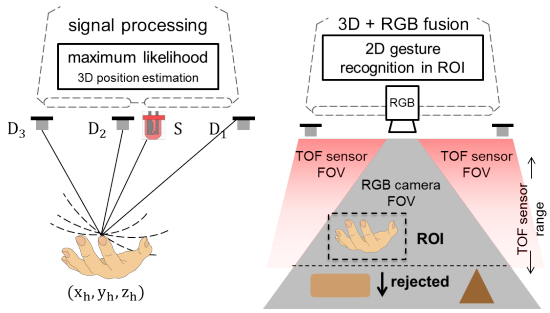

Figure 2. Mime sensor operation.

Key Technical Advantages

- **Application specific sensing:** Mime is intended for single-handed gestural interaction. It sacrifices generality of use and working range (0-4 feet) to achieve high frame-rate and low power, accurate performance.
- **Hybrid RGB and TOF sensing:** Mime accomplishes a unique combination of 3D and RGB information – it effectively robustifies standard computer vision algorithms used to recognize gestures from RGB information.
- **Disrupting the computer vision pipeline:** Mime's RGB image processing pipeline effectively rejects most of the pixels captured in the RGB image, focusing only on the region of interest (ROI). This fusion drastically improves the robustness of the system in cluttered, complex and fast changing environments, where false positives often render the standard vision algorithms useless.
- **Depth super-resolution:** Mime has a small baseline of $5 - 7$ cm compared with the working range of $0 - 1.2$m and achieves centimeter-accurate 3D localization through physically-accurate signal modeling.
- **Daylight insensitivity:** Mime's signal processing only makes use of high frequency information enabling it to be robust to daylight and light fluctuations by rejecting low frequency ambient noise.

Hardware Implementation and Evaluation

The Mime sensor (see Figure 1 (a), (b)) was implemented using standard off-the-shelf opto-electronic components. The illumination source is an LED (OSRAM 4236), the photodetectors are silicon PIN diodes (FDS100), and the received signals are sampled through a 4-channel USB oscilloscope using a 5MHz sampling bandwidth (low compared with pulse modulation of 200ns at 10kHz).

We evaluated the performance of the Mime implementation on a number of different metrics including resolution, power, working range, latency and daylight insensitivity. We provide details in our paper [3].

Gestures implemented using Mime

We implemented 4 motion-cue gestures using Mime's TOF sensor only. These are swipe in a straight line (left to right, up to down, diagonally), point-and-click, zoom in and out using depth, circle gesture (see Figure 3). We implemented a few example use cases that use the fusion approach (see Figure 1).

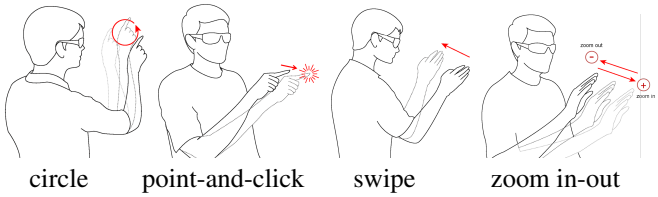

Figure 3. Motion-controlled gestures implemented using only the 3D coordinate data acquired by Mime's TOF sensor.

FUTURE WORK: SENSOR DEVELOPMENT

The rest of my thesis work will focus on advancing the sensor hardware, extending the algorithmic framework to detect and track multiple fingers and hands, and finally develop applications for smart phones and smart glasses that utilize the sensor information. To evaluate the performance of the Mime sensor and its efficacy for input to applications, my goal will be to conduct evaluation experiments that study performance.

Signal Processing and Hardware Improvements

Currently, the Mime system is limited to single-handed gestural input. System extension to more complex input capabilities requires developing two separate aspects: algorithmic extension to the TOF module to detect and track two hands in 3D space; RGB region of interest processing will be used to identify finer features like multiple fingers. The former will require careful optical analysis of the source and sensors to understand bandwidth and power trade-offs in acquiring multiple target information in the field of view.

FUTURE WORK: APPLICATIONS

With the existing Mime system we have demonstrated simple styles of interaction that mimic mouse pointer input, navigation and shape based interactions. With the hardware and processing extensions described above, the Mime system will be capable of supporting more complex interaction styles. Here, I describe two specific applications that attempt to go beyond the input capabilities of smart glasses and mobile phones.

Back to the desktop

In this application space we are interested in making small display devices capable of supporting complex actions desirable for desktop style input configurations. We target tasks that have a high density of input actions to content such as text composition, navigating large documents, target acquisition in high dimensional data sets. Currently, the limitation in display size and touch input limits complex input to smartphones and tablets without an external keyboard. Here, we propose constructing a virtual desktop centered around the mobile display and use the surface around the display for opportunistic input. The Mime sensor provides this input through hand gesture tracking and motion sensing. We propose the following new applications for smartphones and smart glasses.

For smart phones:

The Mime sensor on the phone allows the table surface next to the phone to be mapped to conventional desktop windows, and the phone's display is a small viewport onto this desktop. Moving the hand is like moving the mouse, and as the user shifts into another part of the desktop, the phone viewport display moves with it. Instead of writing new applications to use

Figure 4. Mobile application scenarios with Mime: volumetric input for mobile interaction

smart surfaces, existing applications can be readily controlled with the hands. An on-demand keyboard, mouse and application space is sensed on the surface around the device (Figure 4 left). The keyboard position is virtually placed at the location of the users hands and finger movements constitute relative keystrokes. This configuration introduces a new combination of existing surface gestures (pinch-to-zoom, scale) with traditional keyboard style input without using the touch surface.

For smart glasses:
In this case we will experiment with new forms of input to novel virtual interfaces for head mounted displays. The users hand position and finger orientation is accurately tracked over time and the fingertips are overlaid with virtual menu items for frequently accessed applications (Figure 4 center). We will experiment the option to carry apps at your finger-tips. Also a set of gestures needs to be developed to mimic the mouse in order to interact with these applications once they are selected.

Interactive capture tools for photography

In this application case we use the Mime sensor to allow manipulation of visual content as we capture it. While capturing pictures from a vantage point the user typically cannot manipulate it on the display screen itself because of the smaller degrees of freedom of control available with a flat touchscreen and the inherently limited display size. Our approach allows manipulation visual content during the capture pipeline through gestures that map to the scene or region of interest being captured. The key concept is the use of gestures to interact with the image while the photo is being taken and the scene and desired view is fresh and alive. For smart phones: An example is shown in Figure 4 (right), where the mobile user is photographing the objects of interest, modifying the images using gesture shortcuts and finally integrating them in the powerpoint presentation. Selecting a region of interest, rotation and cropping are some other simple gestures that can be implemented. We will extend this app to support frequently used photography tools. For smart glasses: In a head-mounted display with a camera seeing the world, we propose to experiment with methods of selecting a region of interest that maps to the users viewpoint, and overlaying the image with hand annotations in the form of text or emoticons.

CONCLUSIONS

The use of computing systems is defined and limited by the set of input actions available to the user. Desktop based systems limit the interaction space to the keyboard and mouse. The use of touch devices confines the user to the boundaries of the display itself. Emerging wearable displays like Google Glass attempt to dissolve the boundary between display and the computing platform behind it, thereby making the user experience truly mobile by limiting the need to pull out an additional device. While output to the user is achieved through information overlaid on the display, designing responsive input to the system is of paramount importance for an effective on-the-go computing experience. This paper presents a compact, low-power 3D gesture tracker that is amenable to mobile size and power constraints and supports new interaction possibilities.

ACKNOWLEDGEMENTS

The author would like to thank her thesis committee members Chris Schmandt and Vivek K Goyal, MIT.

REFERENCES

1. Bailly, G., Müller, J., Rohs, M., Wigdor, D., and Kratz, S. Shoesense: a new perspective on gestural interaction and wearable applications. In *ACM Conf. Human Factors in Comput. Syst.* (2012).
2. Chen, X. A., Marquardt, N., Tang, A., Boring, S., and Greenberg, S. Extending a mobile device's interaction space through body-centric interaction. In *Proc. 14th ACM Int. Conf. Human-Computer Interaction with Mobile Devices and Services* (2012).
3. Colaço, A., Kirmani, A., Yang, H. S., Gong, N.-W., Schmandt, C., and Goyal, V. K. Mime: Compact, low-power 3d gesture sensing for interaction with head-mounted displays. In *ACM UIST* (2013).
4. Gustafson, S., Bierwirth, D., and Baudisch, P. Imaginary interfaces: spatial interaction with empty hands and without visual feedback. In *Proc. ACM Symp. User Interface Softw. Tech.* (2010).
5. Harrison, C., Benko, H., and Wilson, A. D. Omnitouch: wearable multitouch interaction everywhere. In *Pro. Ann. ACM Symp. User Interface Softw. Tech.* (2011).
6. Jones, B., Sodhi, R., Forsyth, D., Bailey, B., and Maciocci, G. Around device interaction for multiscale navigation. In *Proc. Conf. HCI with Mobile Devices and Services* (2012).
7. Kim, D., Hilliges, O., Izadi, S., Butler, A. D., Chen, J., Oikonomidis, I., and Olivier, P. Digits: freehand 3d interactions anywhere using a wrist-worn gloveless sensor. In *ACM UIST* (2012).
8. Lange, R., and Seitz, P. Solid-state time-of-flight range camera. *IEEE J. Quant. Electron.* (2001).
9. Mistry, P., Maes, P., and Chang, L. Wuw-wear ur world: a wearable gestural interface. In *CHI'09 Ext. Abs. Human Factors in Comput. Syst.* (2009).
10. Mitra, S., and Acharya, T. Gesture recognition: A survey. *IEEE Trans. Syst., Man, Cybernetics, Part C: Appl. Rev.* (2007).
11. Smith, R., Piekarski, W., and Wigley, G. Hand tracking for low powered mobile ar user interfaces. In *Proc. Sixth Australasian Conf. User Interface* (2005).
12. Starner, T., Auxier, J., Ashbrook, D., and Gandy, M. The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In *IEEE Int. Symp. Wearable Comput.* (2000).
13. Thomas, B. H., and Piekarski, W. Glove based user interaction techniques for augmented reality in an outdoor environment. *Virtual Reality* (2002).