The State of Gesture Recognition

A seminar essay

Adrian Kündig
Student
ETH Zürich
adkuendi@student.ethz.ch

ABSTRACT
Devices like musik players, phones, and other consumer electronics tend to get smaller and smaller. This imposes a challeng to designers who have to create interfaces for an always decreasing form factor. Recent examples like the iPod Shuffle show that the user interface is sometimes already determining the size of the device.

This trend leads to designers looking for alternative approaches to controlling a device. And Gesture Based interfaces could provide a solution.

In this essay I will discuss 6 different papers that all take a different approach on providing hand pose, hand model or hand gesture recognition. Comparing them on different categories like mobility and instumentation requirements will lead to an overview on the field of gesture recognition and propose appropriate solutions for different scenarios.

ACM Classification: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

General terms: Human Factors, Design

Keywords: Evaluation, Gesture Recognition.

INTRODUCTION
Gesture based interfaces described in literature predominantly use one particular input device - the Data Glove. Data Gloves are used to capture the hand pose of the user, the position of the hand in the room, and sometimes even give feedback to the user through vibration.

But the Data Glove is too obstrusive for daily use. It was originally designed to be used in an instrumented environment were cameras tracked the gloves position by following markers attached to the gloves. Inside the gloves are mechanics that try to infer the pose of the hand by looking at the joint angles. But this mechanics have to be reconfigured every time the user puts on the gloves to ensure accurate reconstruction of the pose.

This heavy instrumentation is not fit for the use in daily live where the hand of the user is often occupied by other objects. Furthermore does a user not want to put on a glove to controll his smartphone or his MP3 player. Additionally the use of cameras to track the users hand position is not feasible in a mobile setting. Finally there is not always a need to know the position of the hand because the interaction with a device can be handled by simple in-air gestures.

HISTORY
Gesture based interfaces made their first appearance in the Virtual Reality setting called Video Place[1]. VideaPlace was a project by Myron Krueger from 1969. Multiple cameras were used to record the user. The recorded movies where then processed on a computer which then in turn projected a 2D shadow of the user onto a screen. On this screen the user could interact with artificial objects like figures and sticks. Yet this project was an art project and not a user interface study.

Later came a paper called Charade[4]. It was published in 1993 by Baudel and Beaudouin-Lafon and was one of the first papers to formally define a set of gestures to control applications like PowerPoint. The user had to put on a Data Glove which was connected to a computer through a wire. Their gestures all had a starting pose, then a direction of the arm motion, and finally an end pose. This paper marked the beginning of formal research in gesture controlled user interfaces.

More recently Steven Spielberg released a Hollywood movie called Minority Report[2]. The movie is set in the USA in the year 2054 where a team of police men fight crime using foreknowledge provided by three physics called precogs. To make the scenes as realistic as possible Steven Spielberg had a team of technical advisors that envisiond future technologies. One of these technologies was a gesture controlled computer interfaces that the main actor would use to browse through information. Spielberg wanted the interface to be used as if the user was conducting an orchestra. This interface was often picked up by the press as one of the possible computer interfaces of the future. John Underkoffler was the lead designer of the user interface developed for this movie and also published a research paper where he described the thoughts behind the design.

John Underkoffler later also founded a company called Oblong industries that has developed a commercially available product called G-Speak[3]. It uses Data Gloves to implement

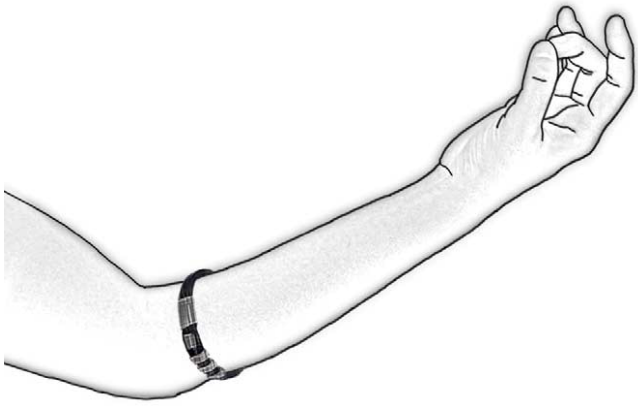

Figure 1: A picture showing the Muscle Computer Interface

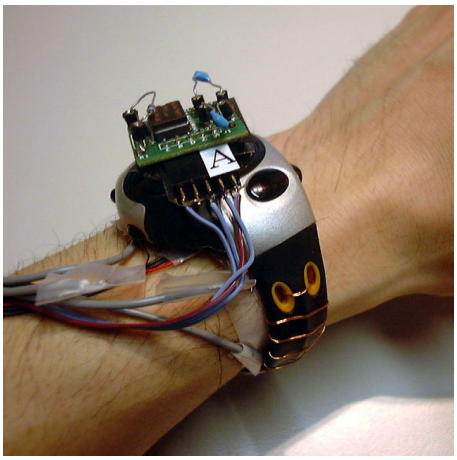

Figure 2: A picture showing the Gesture Wrist device

the very interface that the actor in Minority Report uses and is intended for Big Data analysis and navigation in a virtual 3D environment.

REPLACING THE DATA GLOVE

What most of the historic technologies have in common is a Data Glove. But as mentioned in the introduction designers can not require the user to put on a Data Glove to interact with their devices. Thus the first four papers I am discussing are intended as replacements for the Data Glove. Some of them try to focus on the issue of occupied hands by making it possible to interact with gestures even if the hads are already holding an object. Others try new technical approaches to reconstruct a full 3D model of the hand.

Muscle Computer Interface

The Muscle Computer Interface[10] uses a technology called electromyography to sense muscle activity. Electromyography is used in the field of medicine to assess muscle function and control prosthetics.

It works by measuring the Action Potential generated by the brain to contract a muscle. This Action Potential is an electrical potential. It travels down the nerve to the muscles where it can be measured either invasively or none invasively. The invasive method inserts needles into the muscle, the none invasive method attaches electrodes onto the skin.

The Muscle Computer Interface uses the none invasive method by attaching electrodes in a ring like manner around the back of the forearm.

The data collected by the six electrodes is then further processed to extract a set of features. This features are used by a Support Vector Machine (SVM) which in turn outputs the hand pose it recognized.

The features extracted can be divided into three categories:

- Root Mean square (RMS) of the amplitude per channel and ratio between pairs of RMSs between two channels. A channel represents the output of one electrode.
- Frequency energy of each channel calculated by taking the Fourier Transformation on the output data of a electrode and then summing up the amplitude of all frequencies.
- Phase Coherence between pairs of channels which intuitively can be seen as how strong the data of the electrodes correlates. The Phase Coherence is stronger the more two waves overlap.

The system recognizes fingers touching the thumb or individual fingers being bent more strongly if a hand is occupied holding an object. It is required that the user contracts his muscle by pushing his fingers together hard or by squeezing the object stronly. Otherwise the system does not recognize the interaction event because no muscle activity has been recorded.

Mean accuracy for recognizing free hand gestures ranges from 57% to 79%. While executing free hand gestures the hand is not occupied by an object. The big difference between the lower and the upper bound can be explained by the fact that rotating the arm imposes a big problem to classification. Thus 79% accuracy was achieved when the system was trained in the same arm posture as the test was executed later. The worst recognition rate happened when the arm was rotate very much to the left while training and the very much to the right when testing.

Mean accuracy for recognizing hands busy gestures while holing a coffee mug ranges from 65% to 85%. 65% accuracy was achieved when no feedback about the recognized gesture was provided to the user and the system tried to recognize the activity of four fingers. 85% accuracy was achieved when feedback was given to the user and he/she had to accept the recognized gesture.

Gesture Wrist

The next paper in this discussion is GestureWrist and GesturePad: Unobtrusive Wearable Interaction Devices[9]. The presented Gesture Wrist device uses a technology called capacitance sensing at the wrist to infer the shape of the hand.

Capacitance sensing is done by putting a transmitter at the top of the wrist and multiple electrodes at the bottom. The transmitter then sends out a wave like electrical signal through the wrist. The receiver electrodes capture this signal and calculate the signal strength. To minimize noise coming from

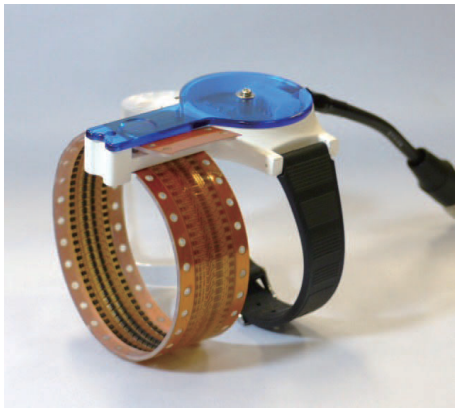

Figure 3: A picture showing the prototype of the hand pose reconstruction device using photoreflectors

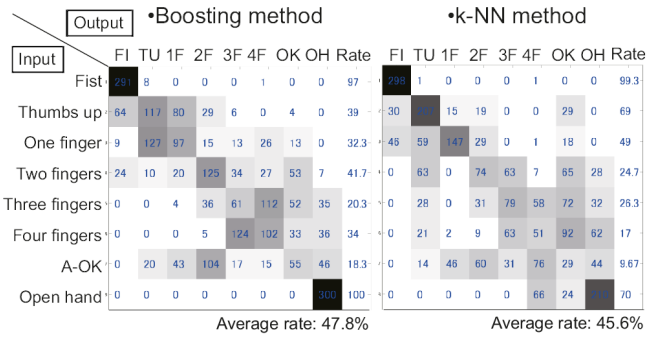

Figure 4: The confusion matrix that shows the test results of hand pose reconstruction using photoreflectors

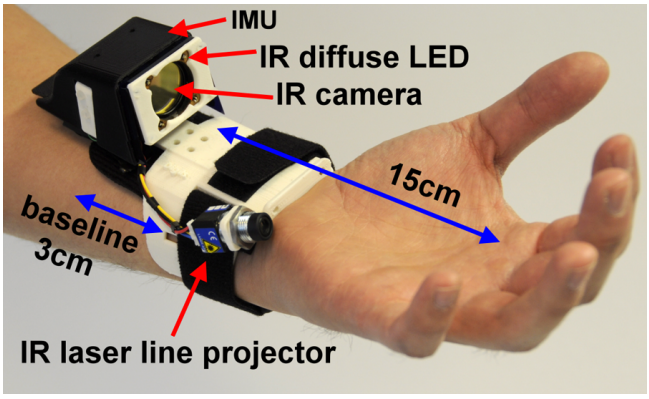

Figure 5: A picture showing the Digits device

the environment the transmitter and the receivers synchronize. This means that the bottom electrodes only sense a signal when the transmitter is sending.

Three sources of resistance can influence the strength of the signal at the bottom. First the resistance between the transmitter electrode and the skin at the wrist top. Second the resistance that the wrist itself imposes on the signal. Last the distance and resistance of the skin at the bottom touching the receiver electrodes.

The first two sources can be considered constant or only changing slightly over time. This slight change can be compensated without much effort. This leaves us with only the distance and resistance between the bottom of the wrist and the receivers.

Gesture Wrist now uses four receiver electrodes to distinguish two hand poses. One is Point and one is Fist. The paper gives no exact numbers on recognition rate but the pictures included in the paper imply that the recognition rate is close to 100%.

The hand pose is then combined with the data of an accelerometer to create a set of gestures like making a pointing pose and then moving the arm to the right. Additionally the user can rotate its arm to give input. This allows him to first select a virtual slider or knob using the gesture set and the adjust the value of the slider by rotating its arm.

Wrist Contour

The next paper called Hand Shape Classification with a Wrist Contour Sensor: Development of a Prototype Device[5] uses photoreflectors to reconstruct the hand pose.

Photoreflectors are small sensors (around $1mm^3$) that consist of an infrared emitting LED and an infrared proximity sensor. The light emitted by the LED is sent of away from the photoreflector and reflected by a surface in front. The reflected infrared light is the sensed by the sensor and used to measure the distance between the photoreflector and the surface in front of it.

The prototype device described in the paper put 150 photoreflectors in two stripes inside a wrist band. A simple machine learning algorithm is then used to differentiate between eight hand poses. This analysis was performed offline.

Unfortunately the results from the testing phase shows us that the relatively weak Machine Learning algorithms combined with very similar looking hand poses do not yield a satisfying result. As shown in the confusion matrix4 the different hand poses are often confused with each other which leads to a mean recognition rate of only 48%.

Digits

The last paper in the series of papers related to the Data Glove is called Digits[7].

The device contains an infrared camera, an infrared laser line generator and four diffuse infrared LEDs. The laser line generator is placed below the camera shifted a bit towards the hand. It projects an infrared line onto the segment of the finger that is nearest to the palm.

The algorithm to triangulate the 3D position of the laser line relative to the camera works in multiple steps. First it illuminates the hand using the diffuse LEDs. Then it subtracts the unilluminated picture from the illuminated one to eliminate the background. In the next step it segments the fingers in the image. After that it triangulates the 3D position of the laser line on the finger by knowing the base line distance between the camera and the laser line generator and the fact that the line moves towards the palm if the fingers a bent and away

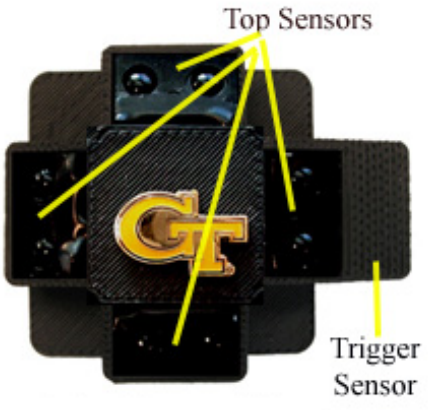

Figure 6: A picture showing the Gesture Watch device

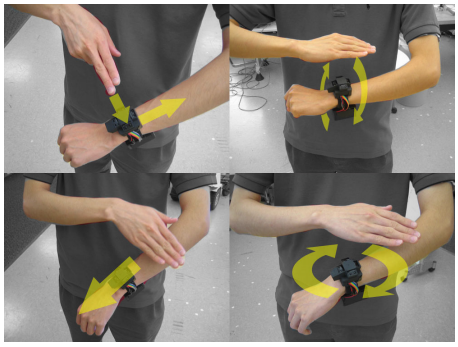

Figure 7: A set of example gestures for the Gesture Watch device

from the palm if the fingers are straightened.

This data is then fed into a forward kinematics model that fits a model of the hand as closely as possible to the calculated position data. To further improve accuracy of the 3D hand model the system can also use the 2D position of the fingertips in the image to feed a inverse kinematics model.

To get the position of the fingertips the algorithm again takes the image with a subtracted background and then searches for the brightest spots in the image which are then assumed to be the fingertips.

Digits is the only device in all analyzed papers that has an accuracy comparable to a Data Glove. In some test instances the authors even reported a higher accuracy.

IN-AIR GESTURES

Until now I have discussed devices that were trying to reconstruct the pose of the hand. One of them, Digits, even reconstruct a full 3D hand model and thus is comparable in resolution to a Data Glove.

But sometimes one does not need information about the hand pose but a simple set of motion gestures. The next two papers are hence focused on recognizing simple motion gestures that are carried out in front of the device.

Gesture Watch

Gesture Watch is a wrist watch like device that senses gestures executed above it. To recognize motion it embeds four infrared proximity sensors that are arranged in a cross shape and a fifth one oriented towards the hand.

These five sensors have a simple binary output - 1 for no occlusion, 0 for occlusion. Their data is then sent via Bluetooth to an external device that processes it. To recognize gestures the authors make use of Hidden Markov Models (HMM). HMMs are a type of machine learning algorithms that are well suited for time depended analysis.

Using HMMs the recognition algorithm is now able to distinguish between several different gestures. Some of them are shown in figure 7. The mean accuracy achieved is also quiet good. Rangeing from 90% while walking outside to 98% while standing inside.

This shows that the technology used is already quiet mature and could be used in a real live application. One drawback is that the user has to accept every gesture by triggering the fifth proximity sensor.

Sound Wave

The last paper discussed in this series of papers is called Sound Wave[6]. It uses the speakers and microphone embedded in most modern laptops to detect motion in front of it exploiting a physical effect called The Doppler Effect.

The Doppler Effect describes what happens if a wave emitting source is moving towards or away from a receiver. A concrete example would be a car making a constant frequency sound through honking. While the car drives towards a person the person hears a higher honk tone than the car is actually emitting. But as soon as the car has passed the pitch of the tone will drop for the person and it will receive it at a lower height than the car is actually emitting.

This effect is now used by the algorithm inside Sound Wave. It emits a constant pitch base tone from the laptops speakers. The tone has a frequency of about 18kHz which is normally not hearable by humans. If an object is then moving towards the microphone the recorded frequency shifts higher in the spectrum. On the other hand if an object moves away from the microphone the recorded frequency shifts down in the spectrum.

Additional to recognizing the speed of an object moving towards or away from the microphone, the algorithm is also able to approximate the distance or size of the object in front of it. This is because bigger or nearer objects reflect more sound waves which then in turn yields a higher amplitude of the shifted tone. Also the algorithm can detect an object that passes the speakers and can tell its direction and velocity.

This approach yields a minimal recognition rate of 86% for the double tap gesture in a quiet environment. The best recognition rate of 100% was accomplished for the two handed gesture where the user moved both hands in opposite direction. Extraordinarily this rate was achieved in a noisy environment at a cafe.

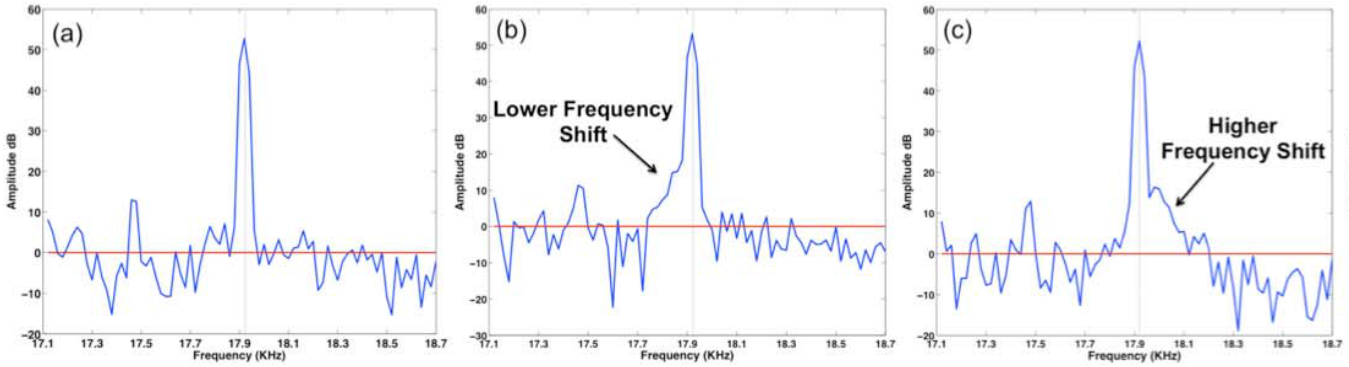

Figure 8: A picture showing the frequency shift that occures when the Doppler Effect is happening

	Mobility	Accuracy	Instrumentation	Main Application
Muscle Computer Interface	Designed for mobile use, data sent via wifi/BT	65% busy hand, no feedback, 4 fingers 91% busy hand, feedback, 3 fingers	An arm band at the upper forearm	Gesture recognition with busy hands
Gesture Wrist (Capacity sensing)	Designed for mobile use, data sent via body network	N/A	Wrist watch like utility	Hand shape recognition, authentication
Wrist Shape (Photoreflectors)	Designed for mobile use, offline processing atm.	45-48%	Wrist watch like utility	Hand shape recognition
Digits (3D reconstruction)	Designed for mobile use, data sent via wifi/BT	91%, varying from finger to finger	Small camera worn at a wrist band	Reconstructing 3D model of hand
Gesture Watch (in air over hand)	Designed for mobile use, data sent via wifi/BT	95 %	Wrist watch like utility	Simple gesture recognition using one hand
Sound Wave (in air over laptop)	Bound to Laptop	90-95%	None, using existing hardware	Add simple gesture recognition to laptop

Figure 9: Table comparing the different technologies

COMPARISON

To compare this six technologies and propose some of them for specific use cases I focus on three different aspects: Mobility, Accuracy, and Instrumentation. Figure 9 shows the comparison as a table.

Mobility

All of the evaluated technologies except Sound Wave are designed for a complete mobile setting. This means that they are all embedded inside a accessory like a wrist watch or an arm band. It is to note that the Digits device is rather large and clumsy and thus is obtrusive compared to the other devices.

For a completely portable gesture interface, for example to control a music player while jogging the Muscle Computer Interface, Gesture Wrist, and Gesture Watch devices could be considered. They all offer a similar amount of gestures that are all recognized quiet accurately.

If in addition the user is holding an object like a coffee mug or a heavy bag then only the Muscle Computer Interface can be considered a solution.

Accuracy

Comparing the accuracy of the different technologies turns out to be rather hard because they offer different levels of hand pose resolution. Thus we will now discuss every technology separate except if two different approaches were taken to achieve the same type of reconstruction.

The Muscle Computer Interface has a high accuracy of over 90% if it is only required to recognize a small number of gestures and allows the user to accept a recognized gesture.

Gesture Wrist and hand pose reconstruction using photoreflectors use the same approach - inferring hand pose by wrist shape - but they have a different number of poses. Gesture Wrist only distinguishes between two poses, point and fist but then enhance this information with the data of an accelerometer to create a set of richer gestures. This seems to be the right approach because trying to infer more poses, like the other paper tries to do, has shown to be inaccurate.

Digits on the other hand does an amazing job recreating the full 3D hand model using only an infrared camera. It achieves very high accuracy and can be seen as a good replacement technology for Data Gloves. Of course also Digits has limitations. One major is self occlusion of the fingers where one finger hides the other one.

Finally both Gesture Watch and Sound Wave have shown to have high accuracy even in bright or noisy environments. Their development seems to be mature and their technology could very much be implemented in a commercial product.

Instrumentation

Hand pose reconstruction using photoreflectors, Gesture Wrist, and Gesture Watch all fit inside an accessory like a wrist watch. The Muscle Computer Interface fits inside an arm band. Digits on the other hand is quit obtrusive as the camera has some size and also has to have some distance from the laser line generator. Sound Wave reuses existing hardware and thus does not require any instrumentation.

CONCLUSION

In conclusion we can say that there exist quiet some different approaches and technologies for hand pose/model reconstruction. Furthermore can we use simple algorithms to enhance existing laptops to also recognize a set of simple gestures. The overall recognition rate ranks from unusable to almost perfect but depends heavily of the environmental constraints and the number of the recognized gestures. To develop a commercial product one would definitly need to combine hand pose reconstruction with additional input for example from an inertial measurement unit. On the other hand in-air gestures through Sound Wave or the Gesture Watch could be used as a stand alone solution to implement a simple set of gestures.

Further work will mainly be invested into reducing the size of the sensors and embedding the technology into day to day objects and accessories. This will finally lead to devices that dont need an artificial user input interface but can be controlled through natual motion and gestures.

REFERENCES

1. A history of Videoplace by Myron Krueger. http://jtnimoy.net/itp/newmediahistory/videoplace/.

2. Minority Report, directed by Steven Spielberg, published in 2002. http://www.imdb.com/title/tt0181689/.

3. Overview about the G-Speak system. http://www.oblong.com/g-speak/.

4. Thomas Baudel and Michel Beaudouin-Lafon. Charade: remote control of objects using free-hand gestures. *Commun. ACM*, 36(7):28–35, July 1993.

5. Rui Fukui, Masahiko Watanabe, Tomoaki Gyota, Masamichi Shimosaka, and Tomomasa Sato. Hand shape classification with a wrist contour sensor: development of a prototype device. In *Proceedings of the 13th international conference on Ubiquitous computing*, UbiComp '11, pages 311–314, New York, NY, USA, 2011. ACM.

6. Sidhant Gupta, Daniel Morris, Shwetak Patel, and Desney Tan. Soundwave: using the doppler effect to sense gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1911–1914, New York, NY, USA, 2012. ACM.

7. David Kim, Otmar Hilliges, Shahram Izadi, Alex D. Butler, Jiawen Chen, Iason Oikonomidis, and Patrick Olivier. Digits: freehand 3d interactions anywhere using a wrist-worn gloveless sensor. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, UIST '12, pages 167–176, New York, NY, USA, 2012. ACM.

8. Jungsoo Kim, Jiasheng He, Kent Lyons, and Thad Starner. The gesture watch: A wireless contact-free gesture based wrist interface. In *Wearable Computers, 2007 11th IEEE International Symposium on*, pages 15–22. IEEE, 2007.

9. Jun Rekimoto. Gesturewrist and gesturepad: Unobtrusive wearable interaction devices. In *Wearable Computers, 2001. Proceedings. Fifth International Symposium on*, pages 21–27. IEEE, 2001.

10. T. Scott Saponas, Desney S. Tan, Dan Morris, Ravin Balakrishnan, Jim Turner, and James A. Landay. Enabling always-available input with muscle-computer interfaces. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, UIST '09, pages 167–176, New York, NY, USA, 2009. ACM.