MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

One-Handed Gesture Recognition Using Ultrasonic Doppler Sonar

Kaustubh Kalgaonkar, Bhiksha Raj

TR2009-014   May 2009

Abstract

This paper presents a new device based on ultrasonic sensors to recognize one-handed gestures. The device uses three ultrasonic receivers and a single transmitter. Gestures are characterized through the Doppler frequency shifts they generate in reflections of an ultrasonic tone emitted by the transmitter. We show that this setup can be used to classify simple one-handed gestures with high accuracy. The ultrasonic doppler based device is very inexpensive - 20 USD for the whole setup including the acquisition system, and computationally efficient as compared to most traditional devices (e.g. video).

ICASSP 2009

ONE-HANDED GESTURE RECOGNITION USING ULTRASONIC DOPPLER SONAR

Kaustubh Kalgaonkar

Electrical and Computer Engineering, Georgia Tech.
Atlanta, GA 30332
kaustubh@ece.gatech.edu

Bhiksha Raj

Mitsubishi Electric Research Laboratories
Cambridge, MA 02139
bhiksha@merl.com

ABSTRACT

This paper presents a new device based on ultrasonic sensors to recognize one-handed gestures. The device uses three ultrasonic receivers and a single transmitter. Gestures are characterized through the Doppler frequency shifts they generate in reflections of an ultrasonic tone emitted by the transmitter. We show that this setup can be used to classify simple one-handed gestures with high accuracy. The ultrasonic doppler based device is very inexpensive – $20 USD for the whole setup including the acquisition system, and computationally efficient as compared to most traditional devices (e.g. video).

1. INTRODUCTION

The act of gesturing is an integral part of human communication and may be used to express a variety of feelings and thoughts, from emotions as diverse as taunting, disapproval, joy and affection, to commands and invocations. In fact, gestures may be the most natural way for humans to communicate with their environment and fellow humans, next only to speech [1].

In recognition of this fact, it is now becoming increasingly common for computerized devices to use hand gestures as a major mode of interaction between the user and the system. The resounding success of the "Nintendo Wii", incorporating the Wii remote has shown that allowing users to interact with computer games using hand gestures can enhance their user experience greatly. The DiamondTouch [2] table, the Microsoft *Surface*, and the Apple *iPhone* all allow interaction with the computer through gestures, doing away with the traditional keyboard and mouse input devices. Other, more futuristic user interfaces allow a wider range of gestures.

However, for gesture-based interfaces to be effective, it is crucial for them to be able to recognize the actual gestures accurately. This is a difficult task and remains an area of active research. In order to reduce the complexity of the task, gesture-recognizing interfaces typically use a variety of simplifying assumptions. The DiamondTouch, Microsoft Surface and iPhone expect the user to touch a surface, and only make such inferences as might be inferred from the location of the touch, such as the positioning or resizing of objects on the screen. The Wii requires the user to hold a device and even so only makes the simplest inferences that might be deduced from the acceleration of a hand-held device. Other gesture recognition mechanisms that make more generic inferences may be broadly classified into 1) Mouse or Pen based input [3], 2) Methods that use data-gloves [4], and 3) Video based techniques [5, 6]. Each of these approaches has its advantages and disadvantages. Mouse/Pen based methods require the user to be in physical contact with a mouse or pen (in fact, the DiamondTouch, Surface and iPhone may all arguably be claimed to be instances of pen-based methods, where the "pen" is a hand or a finger). Data glove based methods demand that the user wear a specially manufactured glove. Although these methods are highly accurate at identifying gestures they are not truly freehand. The requirement to touch, hold or wear devices could be considered to be intrusive in some applications. Video based techniques, on the other hand, are free-hand, but are computationally very intensive.

In this paper we propose a new device, based on the Doppler effect, for the recognition of one-hand gestures. Our device consists of an Acoustic Doppler Sonar(ADS) including a single ultrasonic transmitter and three receivers. The transmitter emits an ultrasonic tone that is reflected by the moving hand. The reflected tone undergoes a Doppler frequency shift that is dependent on the current velocity of the hand. Together,they form a composite characterization of the velocity of the hand in multiple directions, as a function of time. The signals captured by the three receivers are then used to recognize the specific gesture.

The ADS apparatus can be assembled with off-the-shelf components and costs less than $20 at the time of prototype assembly (2007). The device is non-intrusive as the user need not wear/touch any special device. Computationally too, the ADS based gesture recognizer is inexpensive, requiring only simple signal processing and classification schemes. The signal from each of the sensors has low bandwidth and can be efficiently sampled and processed in realtime. We are currently able to sample the signal from one sensor through a standard sound card on a Computer; with efficient multiplexing it is possible to sample multiple sensors through a single sound card, thereby cutting the cost of expensive and complicated data acquisition units needed for traditional gesture sensors. Consequently, the ADS is significantly less expensive overall than other popular and currently available devices such as video, data glove, etc.

Experiments show that even using simple signal processing and classification schemes, the ADS sensor can recognize a limited vocabulary of one-handed gestures with an accuracy of over 88%.

2. THE DEVICE AND PRINCIPLE

The Ultrasonic Doppler based device used for gesture recognition is an extension of the device suggested by Kalgaonkar and Raj [7, 8]. The device uses the Doppler effect to characterize complex movements of articulated objects such as hands or legs through the spectrum of an ultra-sound signal. A central transmitter emits an ultrasound tone that is bounced off the moving hand that makes gestures. The reflected signal is captured by three spatially separated receivers in order to characterize the motion in three dimensions. The following subsections describe the device and the principles of its operation in greater detail.

2.1. The Device

Figure 1 shows a picture of our Doppler-based gesture recognition device. It consists of a centrally placed ultrasonic transmitter (Tx), and three ultrasonic sonic receivers (L,C and R). The relative locations of the transmitter and recievers are illustrated in Figure 1, where the three receivers are located in the XY plane. For ease of description we will call the sensor labeled "**L**" as *left*, the one labeled "**C**" as *center* and the one labeled "**R**" as *right*. The transmitter is not coplanar to the receivers, it is displaced along the Z-axis and couple of centimeters behind the L, C and R plane. The transmitter lies in-line with the 'orthocenter' of the triangle formed by the three sensors. The configuration of the transmitters and the receiver was specifically chosen to improve the discriminative ability of the device (see Section 5 for details).

The transmitter is connected to a 40 kHz oscillator through a power amplifier. The power amplifier is used to control the range of the device. Long-range sensors can be used by people with disabilities to efficiently control devices and application around the house.

The ultrasonic transmitter emits a 40 kHz tone and all the receivers are tuned to receive a 40 kHz signal and have a 3db bandwidth of about 4 kHz. Both the emitter and receiver have a diameter that is approximately equal to the wavelength of the emitted 40kHz tone, and thus have a beamwidth of about 60^o, making them quite directional. The high-frequency transmitter and receiver both cost less than $1 USD.

The signals that are captured by the receivers are centered at 40 kHz and have frequency shifts that is characteristic of the movement of the gesturing hand. This bandwidth of the received signal is typically considerably less than 4 kHz. The received signals are digitized by sampling. Since the receivers are highly tuned, the principle of band-pass sampling may be applied, and the received signal need not be sampled at more than 16 kHz (although in our experiments we have sampled the signal at 96 kHz and decimated them algorithmically).

All gestures to be recognized are performed in front of the setup. The range of the device depends on the power of the transmitted signal which can be adjusted to avoid capturing random movements in the field of the sensor.

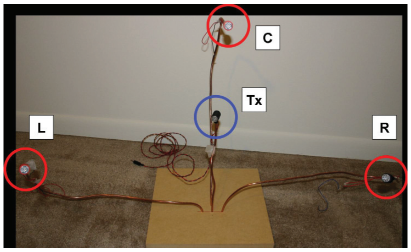


Fig. 1. Prototype device. Transmitter has been highlighted by a blue circle and the receivers is highlighted by the red circle.

2.2. Principle of Operation

The ADS operates on the Doppler's effect, whereby the frequency perceived by a listener who is in motion relative to the signal emitter is different from that emitted by the source. Specifically if the source emits a frequency f that is reflected by an object moving with velocity v with respect to the transmitter, then the reflected signal sensed at the emitter $\hat{f} = (v_s + v)(v_s - v)^{-1} f$ were v_s is the velocity of the sound in the medium. If the signal is reflected by multiple objects moving at different velocities then multiple frequencies will be sensed at the receiver.

The gesturing hand in this case can be modeled as an articulated object of multiple components moving at different velocities. When the hand moves the articulators including but not limited to the palm, wrist, digits etc. move with velocities that depend on the gesture and the subject.

The ultrasonic signal reflected off the hand of the subject has multiple frequencies each associated with one of the moving components. This reflected signal can be mathematically modeled as

$$d(t) = \sum_{i=1}^{N} a_i(t) cos(2\pi f_i(t) + \phi_i) + \Upsilon \tag{1}$$

where f_i is the frequency of the reflected signal from the i^{th} articulator, which is dependent on v_i velocity of the component (direction of motion and speed). f_c is the transmitted ultrasonic frequency(40 kHz). $a_i(t)$ is a time-varying reflection coefficient that is related to the distance of the articulator from the sensor. ϕ_i is an articulator-specific phase correction term. The term within the summation in Equation 1 represents the sum of a number of frequency modulated signals, where the modulating signals $f_i(t)$ are the velocity functions of the articulators. We do not, however, attempt to resolve the individual velocity functions via demodulation. The quantity Υ captures the background reflections which are constant to a given device environment. Figure 2 shows a typical Doppler signal acquired by the three ultrasonic receivers. Due to the narrow beamwidth of the ultrasonic receivers the three sensors capture distinct signal.

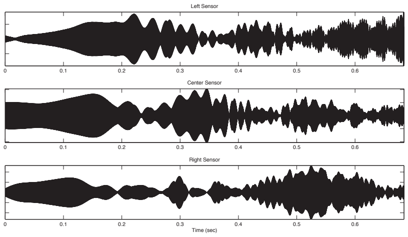


Fig. 2. Doppler Signal clockwise Gesture.

The functions $f_i(t)$ in $d(t)$ are characteristic of the velocities of the various parts of the hand for a given gesture. Consequently $f_i(t)$, and thereby the spectral composition of $d(t)$ are characteristic of the specific gesture.

3. SIGNAL PROCESSING

Three separate signals are captured by the three Doppler sensors. All signals are originally sampled at 96 kHz. Since the ultrasonic sensor is highly frequency selective, the effective 3 dB bandwidth of the Doppler signal is less than 4 kHz, centered at 40 kHz and are attenuated by over 12 dB at 40 kHz ± 4 kHz. It is to be noted that the frequency shifts due to gestures do not usually vary outside this range anyway. We therefore heterodyne the signal from the Doppler channel down by 36 kHz so that the signal is now centered at 4 kHz. The signal is then decimated and sampled at 16 kHz for further processing.

3.1. Feature Extraction

The actions that constitute a gesture are fast. The Doppler signal in this case in not as slow varying as seen by Kalgaonkar and Raj[7].To capture the frequency characteristics of the Doppler signal we therefore segment it into relatively small analysis frames of 32 ms. Adjacent frames overlap by 50%. Each frame is Hamming windowed and a 512-point Fourier transform performed on it to obtain a 257-point power spectral vector. The power spectrum is logarithmically compressed and a Discrete Cosine Transform (DCT) is applied to it. The first 40 DCT coefficents are retained to obtain a 40-dimensional cepstral vector.

40 cepstral coefficients are calculated for the data from each of the sensors. The data from all the three sensors($\mathbf{v}_L, \mathbf{v}_C, \mathbf{v}_R \in \mathbb{R}^{40\times1}$) are then combined to form a larger feature vector $\mathbf{v} = [\mathbf{v}_L^T, \mathbf{v}_C^T, \mathbf{v}_R^T]^T$, $\mathbf{v} \in \mathbb{R}^{120\times1}$.

The signals captured by the three receivers are highly correlated, and consequently the cepstral features computed from them are correlated as well. We therefore decorrelate $\mathbf{v}$ using PCA and further reduce the dimension of the concatenated feature vector to 60 coefficients.

4. CLASSIFIER

In this experiment we wanted to validate if the current setup and the ultrasonic sensor could be used to infer gestures so we used a simple Bayesian formulation for gesture identification. We model the distribution of the feature vectors obtained from the Doppler signal for any gesture g by a Gaussian Mixture Model (GMM):

$$P(\mathbf{v}|g) = \sum_i c_{g,i}\mathcal{N}(\mathbf{v}; \mu_{g,i}, \sigma_{g,i}) \qquad (2)$$

where $\mathbf{v}$ represents a feature vector, $P(\mathbf{v}|g)$ represents the distribution of feature vectors for gesture g, $\mathcal{N}(\mathbf{v}; \mu, \sigma)$ represents the value of a multivariate Gaussian with mean μ and variance σ at a point $\mathbf{v}$, and $\mu_{g,i}$, $\sigma_{g,i}$ and $c_{g,i}$ represent the mean, variance and mixture weight respectively of the i^{th} Gaussian in the distribution for a gesture g. This model ignores any temporal dependencies between the vectors and models them as iid.

Once the parameters of the Gaussian mixture models for all eight gestures are learned, subsequent recordings are classified using a simple Bayesian classifer. Let $\mathbf{v}$ represent the set of combined feature vectors obtained from a Doppler recording of a gesture. The gesture is recognized as a $\widehat{g}$ according to the rule:

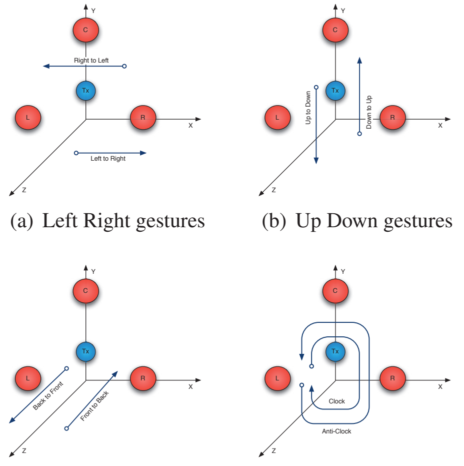
$$\widehat{g} = \underset{g}{\operatorname{argmax}}\, P(g) \prod_{\mathbf{v}\in\mathbf{V}} P(\mathbf{v}|g) \qquad (3)$$

where $P(g)$ represents the *a priori* probability of gesture g. Typically, $P(g)$ is assumed to be uniform across all the classes (gestures), since it may not be reasonable to make any assumptions about the gesture *a priori*.

5. GESTURES

To evaluate the ability of the ADS to recognize one handed gestures we chose eight distinct gestures that could be made with one hand. Figure 3 shows the actions that constitute the gestures. These gestures are performed within the range of the device. The orientation of the fingers and palm has no bearing on recognition or the meaning of the gesture.

Some of the gestures are self-explanatory, but for the sake of completeness we present a short explanation of all the gestures.



(a) Left Right gestures (b) Up Down gestures

(c) Front Back gestures (d) Clockwise / Anti-Clockwise

Fig. 3. Action Constituting a Gestures

1. **Left to Right (L2R)**: As indicated in Figure 3(a), this gesture is the movement of the hand from sensor L to sensor R.
2. **Right to Left (R2L)**: As indicated in Figure 3(a), this gesture is the movement of the hand from sensor R to sensor L.
3. **Up to Down (U2D)**: As indicated in Figure 3(b), this gesture is the movement of the hand from base (line connecting sensors L and R) towards sensor C.
4. **Up to Down (D2U)**: As indicated in Figure 3(b), this gesture is the movement of the hand from sensor C towards the base.
5. **Back to Front (B2F)**: As indicated in Figure 3(c), this gesture is the movement of the hand from subjects body towards the plane of the sensors/device.
6. **Back to Front (F2B)**: As indicated in Figure 3(c), this gesture is the movement of the hand from the sensors/device towards subjects body.
7. **Clockwise (CG)**: As indicated in Figure 3(d), this gesture is the movement of the hand in a clock wise direction.
8. **Anti-clockwise (AC)**: As indicated in Figure 3(d), this gesture is the movement of the hand in an anti-clockwise direction.

We specifically chose these eight gestures, to accentuate, the discriminative power of the device, just as an example, the clock-wise 3(d) hand movement might be misinterpreted as left-to-right 3(a) depending the trajectory taken by the hand.

Readers should also note that the configuration of the transmitter and the receiver plays an important role in the operation of the device. Gestures are inherently confusable; for instance, the L2R, R2L, U2D and D2U gestures are the part of the clockwise and anti-clockwise gestures. The distinction between these gestures would frequently not be apparent using only two sensors, regardless of their arrangement. It is to overcome this difficulty that we have three receivers that capture and encode the direction information of the hand more finely. For instance, one of the main difference between the L2R and clockwise gesture will be the signal gathered by the sensor C. The L2R gesture takes place in XZ plane with a constant Y value which will not be the case with the clockwise gesture. This motion along the Y axis is recorded by the C sensor.

The other key challenge in recognizing gestures is the inherent variability in performing the gestures. Each gesture has three stages *the start*, *the stroke* and *the end*. Gestures start and end at a resting position each individual may have his or her own start and end points. Each user also has a unique style and speed of performing the gesture. All these factors add a lot of variability in the gathered data. Gesture time is defined as the spent in preforming a single stroke. The boxplots shown in Figure 4 summarizes the gesture time information obtained from out dataset.

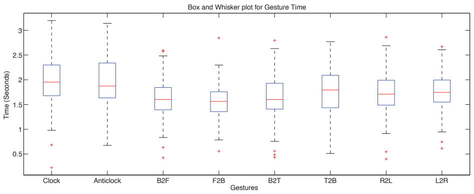


Fig. 4. Summary of Time per Gesture (Box and Whiskers Plot)

6. EXPERIMENTS AND RESULTS

We collected data for all the eight gestures from ten users. The group of subjects were both left and right handed males and females. Right and Left handed individuals have different starting stances. This was one of the primary reasons to collect data from both types of users – to ensure that the gesture recognizer worked irrespective of the users preferred handedness. The device was placed on a table in a room. Data collection process was monitored using a webcam.

Subjects were instructed at the start of the experiment on how to perform the gesture in front of the device. Each subject stood in front of the device and performed all the gestures. They were not interrupted during the experiments, neither were any attempts made to change their actions. They were allowed all the freedom to record the data; this was done to maintain diversity in recorded data.

Each subject stood facing the device and recorded ten instances of each gesture. We had 100 instances for each gesture. 60% of the data for each gesture(60 instances) was used for training and the rest of the data was reserved for testing. The data was mean normalized before training the models.

Eight user independent Gaussian Mixture Models were trained one for each gestures. We found that 20 Gaussians in the mixture were ideal and provided the best results. We found the classification accuracy for our current simple setup to be **88.42%**. Figure 5 shows the misclassification rates and trends for different gestures. The figure does not show the data for cases of correct classification.

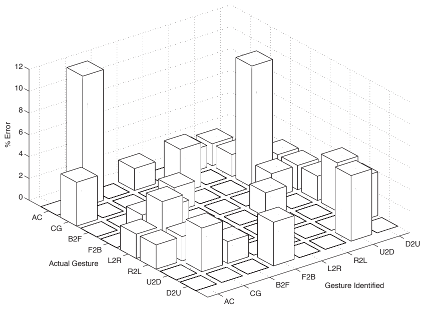


Fig. 5. Error Rate and Distribution for different Gestures

7. DISCUSSION AND FUTURE WORK

In this paper we have demonstrated that a cheap ultrasonic sensor could be used for gesture identification. The information gathered by the sensor is fundamentally different from that obtained from videos, which have been traditionally used for gesture recognition. Thus, it may also be possible to combine the two modalities to achieve significantly better performance than either could deliver by itself.

We believe that the performance reported in Section 6 can be significantly improved by a variety of methods. Currently we have not exploited the temporal characteristics present in the data – feature vectors computed from the data are assumed to be *iid*. Further, besides simple decorrelation of feature vectors, we have made no explicit use of the relationship between the measurements of the sensors. Both of these could be improved on through use of more detailed statistical models such as HMMs or more explicit Bayesian network models. It is our belief that proper modelling of this information will enable us to better the performance of the gesture recognizer.

Many of the errors currently noted, such as the fact that back-to-front(B2F) motion is frequently confused with up-to-down(U2D) movement, can be avoided by improved arrangement of sensors. The optimal arrangement of sensors remains, to our mind, an open problem that could be addressed to great benefit. Furthermore, we believe that greater resolution and accuracy may be achieved using more than 3 sensors, as some of the deficiencies introduced by suboptimal placement of sensors could be voided this way.

Also, our classifiers are currently trained purely to maximize likelihood. We believe that significantly better performance may be achieved through discriminative training and feature extraction techniques.

Finally, we have thus far restricted ourselves to one-handed gestures. We are also looking into sensor setups that will be able to accommodate more complex one-hand and two-hand gestures.

8. REFERENCES

[1] David McNeill, *Gesture and Thought*, University Of Chicago Press, 2005.

[2] P.H Dietz and D.L. Leigh, “Diamondtouch: A multi-user touch technology,” *ACM UIST*, pp. 219–226, 2001.

[3] W.V. Citrin and M.D Gross, “Distributed architecture for pen-based input and diagram recognition,” *AVI Conference on Advanced Visual Interfaces*, pp. 132–140, 1996.

[4] A. Wexelblat, “An approach to natural gesture in virtual environments,” *ACM TOCHI*, vol. 2, no. 3, pp. 179–200, 1995.

[5] A. Utsumi, T. Miyasato, F. Kishino, and R. Nakatsu, “Hand gesture recognition system using multiple cameras,” *Hand Gesture Recognition System Using Multiple Cameras*, 1996.

[6] M. Krueger, T. Gionfriddo, and K Hinrichsen, “Videoplace - an artificial reality,” *ACM CHI*, pp. 35–40, 1985.

[7] K. Kalgaonkar, Rongquiang Hu, and B. Raj, “Ultrasonic doppler sensor for voice activity detection,” *Signal Processing Letters, IEEE*, vol. 14, no. 10, pp. 754–757, Oct. 2007.

[8] K. Kalgaonkar and B. Raj, “Acoustic doppler sonar for gait recogination,” *IEEE AVSS 2007.*, pp. 27–32, 5-7 Sept. 2007.