HandSonor: A Customizable Vision-based Control Interface for Musical Expression

Srinath Sridhar
MPI Informatik and
Universität des Saarlandes
Campus E1.4, 66123
Saarbrücken, Germany
ssridhar@mpi-inf.mpg.de

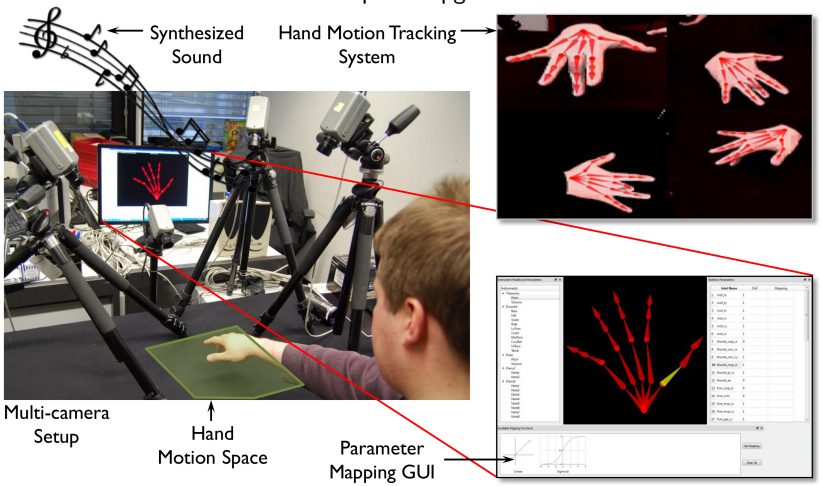

Figure 1: HandSonor provides users with a customizable vision-based control interface for musical expression. Using multiple cameras the articulated 3D pose of the hand is estimated. Users can map parametrized hand motion to instrument parameters and produce music just by moving their hand.

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

Abstract
The availability of electronic audio *synthesizers* has led to the development of many novel control interfaces for music synthesis. The importance of the human hand as a communication channel makes it a natural candidate for such a control interface. In this paper I present **HandSonor**, a novel non-contact and fully customizable control interface that uses the motion of the hand for music synthesis. HandSonor uses images from multiple cameras to track the realtime, articulated 3D motion of the hand without using markers or gloves. I frame the problem of transforming hand motion into music as a parameter mapping problem for a range of instruments. I have built a graphical user interface (GUI) to allow users to dynamically select instruments and map the corresponding parameters to the motion of the hand. I present results of hand motion tracking, parameter mapping and realtime audio synthesis which show that users can play music using HandSonor.

Author Keywords
Musical expression; control interface; hand motion capture; computer vision; music synthesis

ACM Classification Keywords
H.5.2 [User Interfaces]: Auditory (non-speech) feedback; I.4.8 [Scene Analysis]: Tracking.

Related Work – Music Control Interfaces

Theremin: A non-contact, continuous sound musical instrument that is controlled by adjusting the distance of the hands from two antennae. The theremin is suitable only for generating continuous sounds.

The reacTable [4]: A tangible tabletop musical instrument that tracks marked pucks on a table to synthesize sounds. This device is more suited for created *electronic music*.

Gesture-based Control: There are many systems that use marker-based optical tracking of the hands to control music [9, 3]. Markers restrict the free movement of the hand.

Related Work – Markerless Motion Tracking

Body Motion Tracking: There are numerous techniques for markerless tracking of the articulated 3D motion of humans. Some use multiple cameras [8] while others use just depth sensors [1, 7].

Hand Motion Tracking: There exist articulated hand motion tracking techniques that use discriminatively learnt image features [2] or image & depth data together [6]. Special depth sensors have also been developed for this purpose (e.g. Intel Gesture Camera, Leap Motion). However, none of these methods or sensors provide a *kinematic skeleton* at *interactive* rates as HandSonor does.

Introduction

Electronic audio *synthesizers* are useful tools that provide musicians and artists with a way to mimic real instruments or even create new timbres. Typically, synthesizers are controlled using conventional instrument interfaces (e.g. electric guitar). Conventional interfaces enforce a particular way of playing the instrument which has led musicians to explore new physical and software control interfaces for synthesizers.

In HCI literature, many new control interfaces for musical expression including gesture-based control have been proposed [4, 9] (also see sidebar). Controlling music using articulated hand motion (gestures) is intuitive for users since almost all musical instruments are controlled by the hand. Moreover, the hand is an important communication channel with studies showing that the upper bound for the information capacity of the hand is 150 bits per second (bps) [5]. However, most current approaches provide rigid control interfaces that cannot be customized by the user. This makes them no different from conventional instrument interfaces.

In this paper I introduce **HandSonor**, a novel, fully customizable non-contact control interface for playing musical instruments using hand motion. HandSonor allows users to completely customize the correspondence between their hand motions and music synthesized. This provides users with a unique opportunity to explore new forms of musical expression.

There are numerous challenges that need to be overcome in designing interfaces that use hand motion for musical control. The first is fast and robust tracking of the hand with minimum latency. Some systems use marker-based optical tracking [3] but require the user to wear gloves that restrict the free motion of the hands. The second challenge is the design of appropriate mapping schemes between hand motion and music. The *theremin* (see sidebar) maps distances to the pitch and volume of a sound wave. The final challenge is the selection of appropriate instruments to control. HandSonor addresses many of these challenges. The key contributions and benefits of my approach are listed below.

- A markerless *hand motion tracking system* that uses images from multiple cameras to estimate the articulated 3D pose of the hand in the form of a *kinematic skeleton*. This system does not restrict the motion of the hand by enforcing users to wear gloves or makers.
- A *parameter mapping system* for creating correspondences between parameters of the *kinematic skeleton* and parameters of common musical instruments. Users can customize the control interface using a graphical user interface (GUI) to best suit their style and need.
- A realtime *music synthesis system* that can model many different kinds of musical instruments. My approach allows users to explore different musical instruments and is not restricted to any one type.

In the following sections I describe the various components of HandSonor in detail. I provide results of hand tracking, parameter mapping and music synthesis. I plan to conduct a comprehensive user evaluation of HandSonor but present results from a pilot study in this paper. In order to save space, a review of related work is provided in the sidebar on the left.

System Description

Figure 2 gives an overview of the HandSonor pipeline. HandSonor operates in two phases – offline and online.

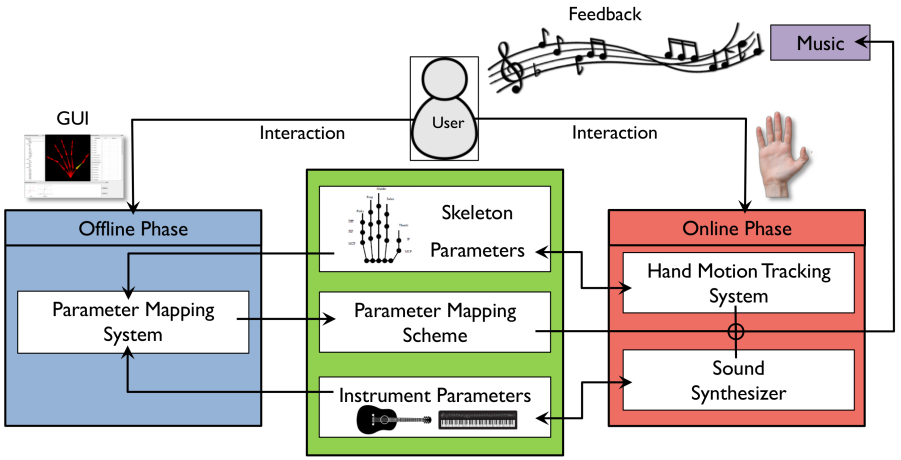

Figure 2: HandSonor Pipeline.

In the offline phase the parameter mapping system is used to create mapping schemes between the skeleton pose parameters (*kinematic chain*) and the instrument parameters. In the online phase, the defined mapping schemes are used along with realtime tracking of the articulated 3D hand pose and instrument synthesizers to generate appropriate sounds.

Hand Motion Tracking System

The hand motion tracking technique I use is based on the full body motion tracking system described by [8]. This is a model-based approach for tracking the articulated motion using multiple, calibrated and synchronized cameras. The model consists of a kinematic skeleton and the volume of the body approximated by a sum of 3D Gaussians (3D SoG).

In order to use this method for tracking hand motion, I built a kinematic skeleton of the hand consisting of 30 joints and represented by 35 parameters, $\Theta = \{\theta_i\}$, where $0 \leq i \leq 34$ (29 joint angles, 3 global rotations and 3 global translations). Each joint angle is limited to a fixed range of angles, $\theta_i \in [l_l^i, l_h^i]$, taken from anatomical studies of the hand. The hand model also contains 49 3D Gaussians attached to the kinematic skeleton each with a fixed mean, variance and colour model. For each different hand to be tracked, a person-specific hand model needs to be constructed. One such hand model used in tracking is shown in Figure 3.

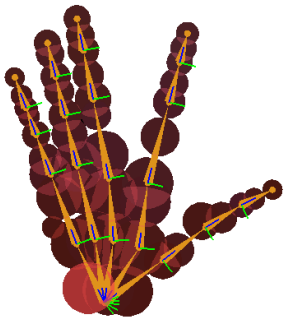

Figure 3: A hand model consisting of a kinematic chain and 3D SoG.

The specific details of the tracking procedure are beyond the scope of this paper but I provide a brief overview here [8]. The frames from each camera are converted into 2D SoG representation using a quad-tree. The parameters of the hand model are optimized so that a similarity measure between the 2D SoG images and the 3D SoG of the hand model is maximized. An objective function is defined that takes into account the similarity measure and the joint limits. A modified gradient ascent optimization of the objective function allows fast computation of the parameters.

Musical Instrument Modelling and Synthesis

Synthetic models of musical instruments are usually represented by a number of parameters, n_{inst}, which are denoted by $\Psi = \{\psi_j\}$, where $0 \leq j < n_{inst}$. Each ψ_j can take continuous or discrete values depending on the type of instrument. For synthesizing the sounds of musical instruments I use the Synthesis Toolkit (STK) [1] provides synthesizers for common musical instruments.

Parameter Mapping System

The goal of the parameter mapping system is to create a mapping scheme between hand model parameters and

[1] https://ccrma.stanford.edu/software/stk/

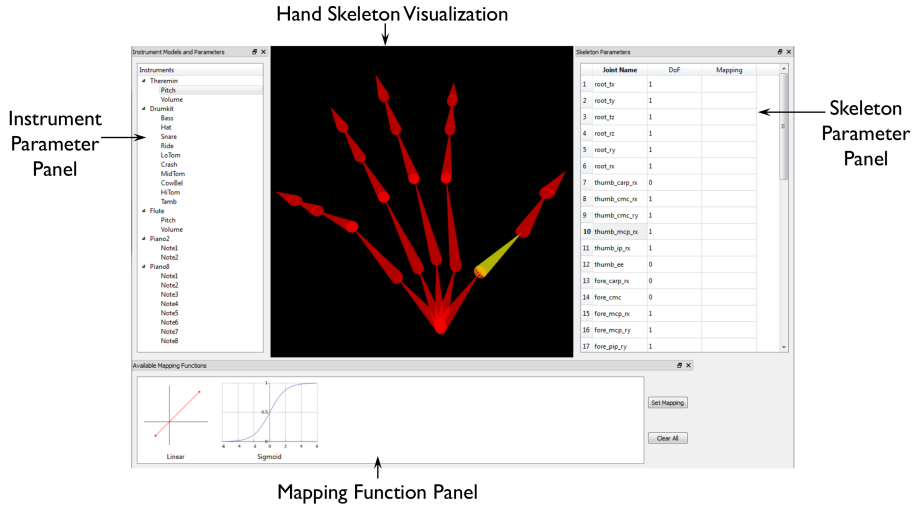

Figure 4: A Screenshot of the GUI showing different components.

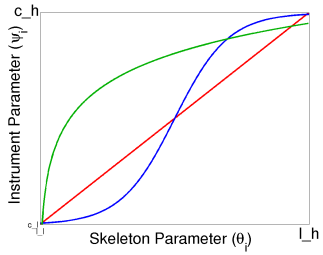

Figure 5: Mapping functions – linear (red), logarithmic (green) and logistic (blue).

instrument parameters. For the purposes of parameter mapping I classify instrument parameters into two types – continuous and discrete. Continuous instrument parameters take continuous values within a specified range of values. Discrete instrument parameters take values from a discrete set and can be boolean. Most common instruments can be represented by either continuous (*e.g.* violin), discrete (*e.g.* drum) or both continuous and discrete parameters (*e.g.* guitar).

Mapping Continuous Instrument Parameters
Continuous instrument parameters, ψ_j, take continuous values over a fixed range, $[c_l^j, c_h^j]$. This allows for a natural mapping to be defined with the hand skeleton parameters which are continuous too. Formally, the mapping can be written as,

$$\psi_j^k = f(\theta_i) \ \forall \ k, j, \tag{1}$$

where $f(.)$ denotes the mapping function, k is the instrument index, ψ_j^k is the j-th parameter of the k-th instrument. Figure 5 shows some of the mapping functions used in my system. Different mapping functions are needed because ψ_j^k for real instruments are distributed unevenly. Some parameters like volume can be linearly controlled while others like pitch work well with a sigmoid mapping function. Notes in different octaves can be mapped to a logarithmic function corresponding to the logarithmic increase in note frequencies.

Mapping Discrete Instrument Parameters
One possible way to represent discrete instrument parameters, ψ_j, is to let them take discrete values such as musical notes. For instance, a piano with n_{notes} number of notes can be modelled using just one parameter ($n_{inst} = 1$) that takes n_{notes} discrete values. However, the mapping scheme becomes much easier if each note were to be represented by a boolean parameter. In this case a piano would be modelled by $n_{inst} = n_{notes}$ parameters. The mapping can then be written as,

$$\psi_j^k = \mathbb{1}_T(\Theta_s) \ \forall \ k, j, \tag{2}$$

where $\mathbb{1}_T$ is the indicator function activated when Θ_s, a subset of Θ, lies within a user defined activation region T. It is important to note here that the mapping is no longer one-to-one but maps one or more skeleton parameters to each instrument parameter. All examples shown in this paper use only translational parameters of the skeleton for mapping with discrete instrument parameters.

GUI for Mapping
In order for users to easily create new mapping schemes I provide a GUI. Figure 4 shows a screenshot of it being used. It consists of a central display that shows a 3D model of the hand skeleton. The right pane on the window

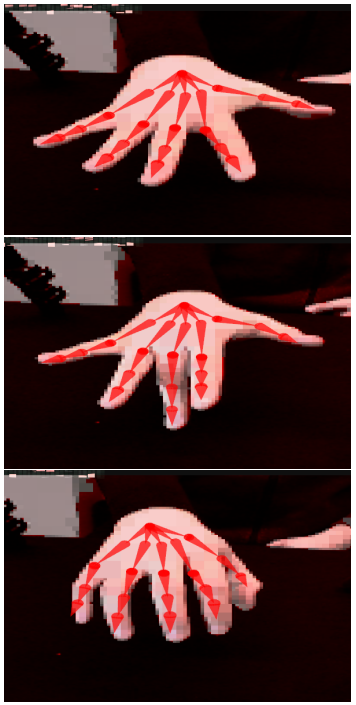

Figure 6: Three frames from a sequence of hand motion being tracked.

lists all the parameters of the skeleton model along with a descriptive name. When a particular parameter is chosen from the list the corresponding joint in the skeleton is highlighted. The left pane lists available instruments and their parameters. Depending on the type of instrument parameter chosen, the bottom panel displays either mapping functions available to the user (continuous) or an option to specify an activation region (discrete).

Technical Assessment

My experimental setup is shown in Figure 1. I used four cameras for tracking hand motion. Figure 6 shows three frames from a video sequence of the hand being tracked. The tracker is robust to translations and articulations but lighting conditions and background clutter do affect the performance of the tracker since it uses colour information. In my setup I used a black background to improve the robustness of tracking results. The hand motion tracking system runs at an interactive rate of 17 frames per second (fps) on a desktop computer with 8GB RAM and an Intel Core i3 CPU. The audio synthesizer based on STK takes a negligible amount of time. I measured the latency from the onset of the user's motion to the corresponding audio response to be between 30-60ms. Since the latency in my system is primarily due to the hand tracking system I intend to use GPU acceleration to reduce the computation time and consequently the latency. I have setup a dedicated webpage[2] that contains videos of users playing using HandSonor.

As an example for users I created a few sample parameter mapping schemes as shown in Table 1.

[2] www.mpi-inf.mpg.de/~ssridhar/handsonor/

Instrument	Parameter Name	Skeleton Parameter	Mapping Function
Theremin	Pitch Volume	root_tx root_tz	Logarithmic Linear
Piano	Notes 1-8 in Octave-0	fore_ee_tx, fore_ee_ty and fore_ee_tz	User define activation region
Drumkit2	Cymbal & Bass	fore_ee_tx, fore_ee_ty and fore_ee_tz	User define activation region

Table 1: Sample mapping schemes. The names of skeleton parameters are abbreviated based on their finger and joint anatomy.

Pilot Study

In order to understand whether HandSonor meets its objective of letting users customize their music creation experience I intend to conduct a comprehensive user evaluation. Currently, I have conducted a pilot study with 4 subjects to understand user perception about HandSonor. All subjects were right-handed and three of them had previous experience playing at least one musical instrument. The study was conducted in two sessions. A user-specific right hand model was also created for all of them in a manual process.

Session 1 (Playing Task): The goal of the first session was to rate whether users could play music using pre-existing mapping schemes from Table 1. For each mapping scheme, users were given the task of playing *Alle Meine Entchen*, a common German children's song. At the end of this session users filled-in a questionnaire with Likert items and open ended questions. I found that users generally preferred to play instruments with discrete parameters than those with continuous parameters.

Session 2 (Exploration Task): The goal of the second

User	Mapping
User 1	Flute, theremin and drumkit
User 2	Piano and Theremin
User 3	Flute and drumkit
User 4	Piano with new activation regions

Table 2: New instrument combinations created by users. The corresponding skeleton parameters are not shown.

session was to evaluate whether users were able to use my parameter mapping system for creating new mapping schemes. I provided users with a fixed amount of time (30 mins) to come up with a new mapping scheme and play some music using this. Similar to the first session, users filled-in a questionnaire. All users successfully created new mapping schemes consisting of multiple instruments but only one was able to play music with it in the given time. Table 2 shows sample multi-instrument combination mappings that users created using HandSonor.

Conclusion and Future Work

In this paper I presented HandSonor, a new customizable non-contact control interface for musical expression. Results from hand motion tracking show that it meets the requirements for music synthesis in terms of robustness and latency. The pilot study indicates that users like the general idea of using hand motion for synthesizing music. They also found the parameter mapping system intuitive and were able to create new mapping schemes.

There still remain challenges that need to be overcome to improve HandSonor. By combining depth information with images, the hand motion tracking system can be made faster, more robust to lighting conditions and background clutter. As a result of the pilot study, I gained valuable insights from users on improvements that could be made to the parameter mapping system and GUI. In the audio synthesis system more instrument models could be incorporated. Finally, I would also like to conduct a thorough user evaluation of HandSonor.

Acknowledgements

I would like to thank my supervisors Prof. Dr. Christian Theobalt and Dr. Antti Oulasvirta for their guidance. I would also like to thank Anna Feit and Thomas Helten.

References

[1] Baak, A., Muller, M., Bharaj, G., Seidel, H.-P., and Theobalt, C. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *IEEE ICCV 2011* (Nov. 2011), 1092 –1099.

[2] Ballan, L., Taneja, A., Gall, J., Van Gool, L., and Pollefeys, M. Motion capture of hands in action using discriminative salient points. In *ECCV 2012*, vol. 7577. 2012, 640–653.

[3] Dobrian, C., and Bevilacqua, F. Gestural control of music: using the vicon 8 motion capture system. NIME '03 (2003), 161–163.

[4] Jordà, S., Geiger, G., Alonso, M., and Kaltenbrunner, M. The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proceedings of Tangible and embedded interaction 2007*, TEI '07 (2007), 139146.

[5] Mao, Z.-H., Lee, H.-N., Sclabassi, R., and Sun, M. Information capacity of the thumb and the index finger in communication. 1535–1545.

[6] Oikonomidis, I., Kyriazis, N., and Argyros, A. Efficient model-based 3D tracking of hand articulations using kinect. British Machine Vision Association (2011), 101.1–101.11.

[7] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. Real-time human pose recognition in parts from single depth images. In *IEEE CVPR 2011* (June 2011), 1297–1304.

[8] Stoll, C., Hasler, N., Gall, J., Seidel, H., and Theobalt, C. Fast articulated motion tracking using a sums of gaussians body model. In *IEEE ICCV 2011* (Nov. 2011), 951 –958.

[9] Wanderley, M., and Depalle, P. Gestural control of sound synthesis. 632 – 644.