CONCERT-MASTER: SYNTHESIS OF MUSIC FOR THE BLIND USING CLASSIFICATION OF HAND GESTURES

Anirud Thyagharajan

Electrical Engineering Department,
Indian Institute of Technology Kharagpur
WB, India - 721302

SL Happy

Electrical Engineering Department,
Indian Institute of Technology Kharagpur
WB, India - 721302

ABSTRACT

Gesture detection is an active area of research in the fields of Computer Vision, Machine Learning and Artificial Intelligence. It involves classifying certain positions of the human body and using them to actuate activities. This paper proposes a gesture based music synthesis system for the blind, to enable them to generate music notes and/or select songs solely by moving their hands. The proposed system consists of a camera (preferably webcam) which detects the user's hands – contour and fingertips, and can be operated in two modes, (1) playing a specific tone of a specific octave, and (2) playing a song from a list of them. The detection of the hands are done using skin detection in the Y-Cb-Cr space followed by a contour analysis of the same. Other modes also include segmentation in the HSV space, and adaptive thresholding using histogram backprojection. The system is programmed in *Python* and has a friendly command line interface for operating various modes.

Index Terms— Gesture, Hand Detection, Music, Y-Cb-Cr

I. INTRODUCTION

The gesture is an important part of human communication, and it is used often - even unconsciously - as a means of expression and interaction with the world, having a strong impact on how humans perceive and interpret themselves [1]. There is an important distinction between gesture and movement [2]. Gesture presupposes an intention, a meaning and the movement is the physic action itself (e.g. a set of arm movements waving composes the gesture of saying goodbye). Nowadays, different methods can be used to capture human movements using, for instance, video cameras, body wearable sensors or external sensors, such as infra-red Motion Capture (MOCAP) systems. All these processes allow capturing and gathering signal data, from where relevant gesture features can be extracted for further analysis and processing (e.g. estimating amplitude, periodicity, rhythm, diversity, etc., of a gesture or movement). Such features can then be used as inputs for real-time algorithmic music composition systems, paving the way for novel expressive and artistic works, where humans and machines interact in a more semantically and artistically meaningful dialog. The motivation behind this particular system, is the possibility of analyzing human gesture at a higher level. Hence, the musical output can present more abstract and complex relations with the human gestures, rather than the direct mapping of human movements.

This paper describes the system Concert Master, a modular system that permits the capture and analysis of human gestures using non-invasive methods (uses the ordinary Laptop camera in a well lit room). The system supports multiple modes, of which we will talk about in the sections to come. The camera captures frames and sends it to a software analyser, which will make sense of the frames depending upon the operating mode. In this paper, we employ 3 techniques to detect the hand, using (1) HSV Thresholding, (2) HSV Thresholding followed by a histogram backprojection and (3) Y-Cb-Cr thresholding. After the gesture or motion of the user is detected, this analyser will trigger a corresponding musical response to the gesture made. The novelty in this system is regarding the simplicity of the hardware and the granularity of selection of music pieces.

The paper is organised as follows. Section I presents a brief introduction to the research area that the system targets and the associated proposed system. Section II presents the prior art in the area and the pathflow of methodology associated. Coming to section III, it talks about the system design and the implementation of the system including the modes of operation. Section IV conveys the preliminary results and observations, while Section V concludes with applications and future work proposed.

II. METHODOLOGY

The basis of the work is based on hand detection. Hand detection in the absence of other body parts resolves to skin detection; and hence the primal investigation in this direction occurs in image processing methods for skin detection. Primitive methods suggest the use of HSV thresholding; with the HSV space calculated as follows:

$$R' = R/255,\ G' = G/255,\ B' = B/255 \tag{1}$$

$$\begin{aligned} C_{max} &= max(R', G', B') \\ C_{min} &= min(R', G', B') \end{aligned} \tag{2}$$

$$\Delta = C_{max} - C_{min} \tag{3}$$

$$H = \begin{cases} 0 & \Delta = 0 \\ 60 * \frac{G'-B'}{\Delta} mod 6 & C_{max} = R' \\ 60 * (\frac{B'-R'}{\Delta} + 2) & C_{max} = G' \\ 60 * (\frac{R'-G'}{\Delta} + 4) & C_{max} = B' \end{cases} \tag{4}$$

$$S = \begin{cases} 0 & C_{max} = 0 \\ \frac{\Delta}{C_{max}} & C_{max} \neq 0 \end{cases} \tag{5}$$

$$V = C_{max} \tag{6}$$

Thus, using the above equations, the HSV space is generated from the RGB color space. After doing so, the image is thresholded as follows. Samples are thresholded with respect to minimum and maximum conditions: H(0, 14), S(66, 154) and V(110, 238). However, as is palpable from our experiments and previous research work, the HSV space was not resilient to changes in illumination, and hence skin detection in such cases was hampered severely.

As a result of this, some steps were taken in the experiments. The proposed system considers a 10 second initial learning phase when it requests the user to place his/her hand in certain squares of the window. The system thereby attempts to estimate the concentration of pixels in the hands, and thereby tries to learn the thresholds adaptively using a histogram backprojection method. This method can be seen in action in the above figure.

Though this method turns out to be much better than the naive HSV thresholding method, this method is also not entirely illumination invariant. [1] proposed a method of skin detection by transforming the image space to the Y-Cb-Cr space. Y-Cb-Cr is a family of color spaces used as a part of the color image pipeline in video and digital photography systems. It is used to separate out a luma signal (Y) that can be stored with high resolution or transmitted at high bandwidth, and two chroma components (Cb and Cr) that can be bandwidth-reduced, subsampled, compressed, or otherwise treated separately for improved system efficiency.

The components can be calculated as follows. As in 1, the analog R', G' and B' components are calculated.

$$\begin{aligned} Y &= 16 + (65.481R' + 128.553G' + 24.966B') \\ C_B &= 128 + (-37.797R' - 74.203G' + 112B') \\ C_R &= 128 + (112R - 93.786G' - 18.214B') \end{aligned}$$

The above Y, C_B and C_R values are measured in the 8 bit format. Thus, [1] noted that thresholding in this space yielded even better results, as it turned out to be illumination invariant to a good extent, discounting any drastic changes (either very dark or very light). However, due to thresholding with Chroma components, the detection could be marred by having objects of similar color, and hence this method fails in such scenarios.

The proposed system has options for all the three methods listed above, and hence can be switched based on the application/environment conditions. The software section will discuss more about switching between these modes.

III. SYSTEM DESIGN

The system was built on a x64 machine running Ubuntu 14.04 in *Python*. The hardware module is just a webcam, which in this case was the default laptop webcamera. The software module had the following dependencies:

- Python 2.7
- OpenCV (2.4.x)
- Pyknon [1]
- Numpy

The software was developed with a user friendly interface, allowing the users to provide options and arguments. The list of options include:

- `--deb`: Used for employing the Debugging mode.

[1] https://github.com/kroger/pyknon

- `--num`: Used for specifying the number of hands.
- `--fre`: Used along with the above option for free music synthesis.
- `--ges`: Used for a 4-class gesture for music selection, disjoint with the above 2 methods.

III-A. Feature Processing

As mentioned earlier, the system accepts input from a computer webcamera. The following points illustrate the processing associated:

- *Free hand movement*: Under this feature, the RGB frame is converted to a Y-Cb-Cr frame. The resultant binarized, thresholded image is sent for a contour analysis, following which the 2 largest contours are identified. Now, depending upon the user's choice, the program detects one or two of the user's hands and computes a unique identifier for music synthesis and pipes it to the music synthesis process. To guide the user, the program also draws a grid showing the pitch and octave variations in the x and y directions respectively. This mode computes the centroids of the contours and sends it to the music synthesis process.

- *Gesture selection*: Under this feature, the RGB frame is converted to a Y-Cb-Cr frame. The resultant binarized, thresholded image is sent for a contour analysis, following which the largest contour is identified. After this, a convex hull is fitted over the identified contour, along with the convexity defects. Thus, this gives us the information about the vertices of the convex polygon fitted, and the defects, which are the points on the contour which lie farthest from the fitted polygon. The main idea of this is to find the junctions between the fingers, with which multiple gestures can be idenitified and classified.
  However, the convexity defects contain extraneous points than the desired junctions, and hence need filtering, based on the distance from the convex hull and the defect, and the angle subtended between two consecutive hull vertices at the defect.

IV. MAJOR HEADINGS

Major headings, for example, "1. Introduction", should appear in all capital letters, bold face if possible, centered in the column, with one blank line before, and one blank line after. Use a period (".") after the heading number, not a colon.

V. PRINTING YOUR PAPER

Print your properly formatted text on high-quality, 8.5 x 11-inch white printer paper. A4 paper is also acceptable, but please leave the extra 0.5 inch (12 mm) empty at the BOTTOM of the page and follow the top and left margins as specified. If the last page of your paper is only partially filled, arrange the columns so that they are evenly balanced if possible, rather than having one long column.

In LaTeX, to start a new column (but not a new page) and help balance the last-page column lengths, you can use the command "\pagebreak" as demonstrated on this page (see the LaTeX source below).

VI. PAGE NUMBERING

Please do **not** paginate your paper. Page numbers, session numbers, and conference identification will be inserted when the paper is included in the proceedings.

VII. ILLUSTRATIONS, GRAPHS, AND PHOTOGRAPHS

Illustrations must appear within the designated margins. They may span the two columns. If possible, position illustrations at the top of columns, rather than in the middle or at the bottom. Caption and number every illustration. All halftone illustrations must be clear black and white prints. Colors may be used, but they should be selected so as to be readable when printed on a black-only printer.

Since there are many ways, often incompatible, of including images (e.g., with experimental results) in a LaTeX document, below is an example of how to do this [1].

VIII. FOOTNOTES

Use footnotes sparingly (or not at all!) and place them at the bottom of the column on the page on which they are referenced. Use Times 9-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

(a) Result 1

(b) Results 3 (c) Result 4

Fig. 1. Example of placing a figure with experimental results.

IX. COPYRIGHT FORMS

You must submit your fully completed, signed IEEE electronic copyright release form when you submit your paper. We **must** have this form before your paper can be published in the proceedings.

X. RELATION TO PRIOR WORK

The text of the paper should contain discussions on how the paper's contributions are related to prior work in the field. It is important to put new work in context, to give credit to foundational work, and to provide details associated with the previous work that have appeared in the literature. This discussion may be a separate, numbered Section or it may appear elsewhere in the body of the manuscript, but it must be present.

You should differentiate what is new and how your work expands on or takes a different path from the prior studies. An example might read something to the effect: "The work presented here has focused on the formulation of the ABC algorithm, which takes advantage of non-uniform time-frequency domain analysis of data. The work by Smith and Cohen considers only fixed time-domain analysis and the work by Jones et al takes a different approach based on fixed frequency partitioning. While the present study is related to recent approaches in time-frequency analysis [3-5], it capitalizes on a new feature space, which was not considered in these earlier studies."

XI. REFERENCES

List and number all bibliographical references at the end of the paper. The references can be numbered in alphabetic order or in order of appearance in the document. When referring to them in the text, type the corresponding reference number in square brackets as shown at the end of this sentence [1]. An additional final page (the fifth page, in most cases) is allowed, but must contain only references to the prior literature.

XII. REFERENCES

[1] Douglas Chai and King N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 4, pp. 551–564, 1999.