ETH zürich

Solfège hand sign recognition with a wearable camera

Gábor Sörös, Julia Giger, Jie Song

Institute for Pervasive Computing, ETH Zurich

[a]

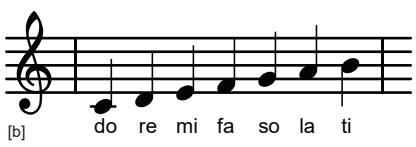
1 Abstract

We present a fast and robust method for recognizing solfège hand signs with smart glasses in egocentric perspective. Our method achieves above 95% classification rate and close to real time performance running on an unmodified Google Glass device.

2 Introduction

A widely used approach in music education is sight singing, where different hand signs are associated with different tones. Solmization is the system of assigning different syllables to each tone in a musical scale, and solfège is one form of solmization practiced in Europe and most English-speaking countries.

Solfège is based on the seven syllables *do, re, mi, fa, so, la, ti.* The associated gesture set is well defined since the XIX. century and contains a small number of gestures only, but is challenging enough to make heuristic recognition approaches fail.



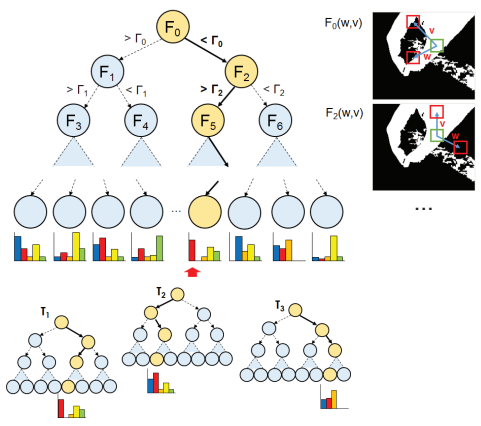
[b]

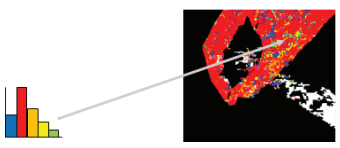

[c]

The egocentric perspective of smart glasses is advantageous for learning and practicing these hand gestures.

3 Method overview

Pixel-wise labeling of binary hand silhouette images with a randomized decision forest, similar to the method described in [1]. Advantages: very fast, only RGB camera, can be parallelized on GPU, robust to segmentation errors, can be extended to depth regression [2].



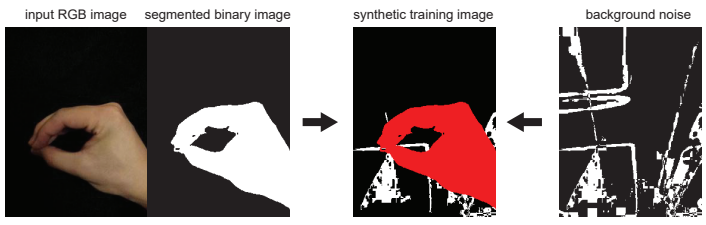


Pooling over all trees: this pixel is red

Pooling over all pixels: this gesture is red

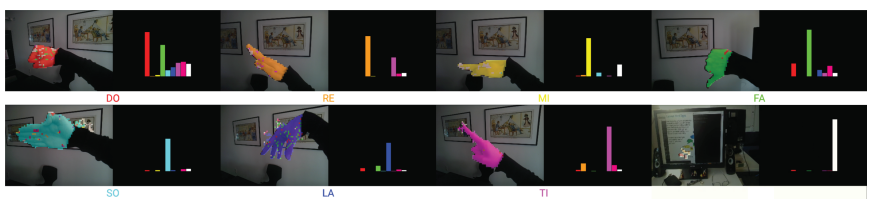
4 Training

We asked six persons to perform the gestures under natural variation and recorded short sequences of each. In total, we recorded 3700 images per gesture covering enough variation in rotation, depth and appearance. The training set also included a *no gesture* hand class where participants casually moved their hands in front of the camera. Additionally, binary noise samples are added to the clean labeled images with a *noise* label prior to training.



5 Results

Screenshots of our application running on a Google Glass are shown below. Each pixel in the image gets classified separately as one of the signs or noise and gets colored accordingly. The left hand side shows the RGB camera input and the pixel-wise classification results overlaid, while the right hand side shows the current probability distribution over all classes. The last screenshot shows how noise pixels are recognized.



To also quantitatively test the classification forest, we used a test set consisting of 2300 images per gesture. Our method classified over 95% of the images correctly.

Built on top of our recognition pipeline, we present a solfège teaching app. The application shows a sheet of music, and the user has to practice the hand signs that reproduce that piece of music. The application recognizes the gestures and gives audio and visual feedback.

6 References

[1] **Jie Song, Gábor Sörös, Fabrizio Pece, Sean Fanello, Shahram Izadi, Cem Keskin, Otmar Hilliges: In-air Gestures Around Unmodified Mobile Devices.** *Proceedings of the 27th ACM User Interface Software and Technology Symposium (UIST 2014), Honolulu, Hawaii, 2014*

[2] **Jie Song, Fabrizio Pece, Gábor Sörös, Marion Koelle, Otmar Hilliges: Joint Estimation of 3D Hand Position and Gestures from Monocular Video for Mobile Interaction.** *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2015), Seoul, South Korea, 2015*

[3] **Gábor Sörös, Julia Giger, Jie Song: Solfège hand sign recognition with a wearable camera.** *Proceedings of the First International Workshop on Egocentric Perception, Interaction, and Computing (EPIC 2016), Amsterdam, the Netherlands, 2016*

[a] drawing adapted from www.classicsforkids.com

[b] drawing adapted from Wikimedia Commons

[c] photo from www.icscamp2016.blogspot.ch