The HealthSign project, current state and future activities

C. Constantinopoulos[1], D. Kosmopoulos[1], A. Argyros[2], I. Oikonomidis[2], V. Lampropoulou[1], K. Antzakas[1], C. Panagopoulos[3], A. Menychtas[3], and C. Theoharatos[4]

[1] University of Patras
{dkosmo,kkonstantino,v.lampropoulou,k.antzakas}@upatras.gr
[2] Foundation for Research and Technology - Hellas
{argyros,oikonom}@ics.forth.gr
[3] Bioassist S.A. {cpan,amenychtas}@bioassist.gr
[4] Irida Labs S.A. htheohar@iridalabs.gr

Abstract. This paper presents the HealthSign project, which deals with the problem of sign language recognition with focus on medical interaction scenarios. The deaf user will be able to communicate in his native sign language with a physician. The continuous signs will be translated to text and presented to the physician. Similarly, the speech will be recognized and presented as text to the deaf users. Two alternative versions of the system will be developed, one doing the recognition on a server, and another one doing the recognition on a mobile device.

Keywords: sign language recognition · computer vision · deep learning.

1 Introduction

Sign Languages (SLs) are the main means of communication for deaf people. The access to SL is essential for the fulfillment of basic Human Rights, however there is a shortage of interpreters, which undermines these rights and often endangers the lives of the deaf, especially in cases of emergency or serious health incidents.

HealthSign proposes to develop an application for the automated interpretation of the Greek Sign Language (GSL) over internet with focus on the health services, which are the most common reason to seek for an interpreter. The high demand for interpreters is often not met or requires long waiting. On the other hand, the availability of interpreters facilitates the integration of the deaf community into the society and .

Vision is probably the only sensor modality that could be of practical use because (a) only vision can capture manual and non-manual cues, which provide essential information for Sign Laguage Recognition (SLR), (b) camera-equipped hand-held devices with powerful processors are a commodity nowadays and (c) recent advances in computer vision and machine learning render mainstream visual SLR a realistic option.

The HealthSign project aims to fulfill the following innovative goals:

- Develop a database of GSL from native speakers with emphasis on health services.
- Implement an internet-based platform for synchronous communication and interpretation with health professionals.
- Develop in parallel a lightweight version which will be able to run on an embedded platform.
- Develop algorithms for recognition of SLs, using computer vision and deep learning to interpret the cues from hands and face and body.
- Implement the algorithms on embedded platforms using FPGAs.

In the long term we aim at the viability of the proposed application, which will be achieved by (a) simple off-the-shelf equipment for the users, (b) efficient implementation of the proposed algorithms, (c) simple installation, and (d) development of a business plan to facilitate the longevity of the proposed product.

The consortium is composed of (a) The Signal Processing and Telecommunications Lab of the University of Patras as expert in machine learning, which is necessary for the recognition of GSL, (b) the Computer Vision and Robotics Lab of ICS-FORTH, which will adapt their 3D hand model for tracking, (c) the Bioassist S.A., which specializes in assistive technologies and develops an internet based platform, (d) the IRIDA S.A. to develop the embedded application, and (e) the Deaf Studies Unit at the University of Patras, which will bring the users, the interpreters and the GSL experts.

2 Project Description

In the following we describe the main components of the project: (a) the detection and tracking methods we are going to use, (b) the methods for sign language modeling and recognition, (c) the architecture for SLR implemented on a server, and (d) the embedded SLR on a mobile device.

2.1 Detection and Tracking of Hand and Body Poses

We propose the adaptation and improvement of the state of the art for detection and tracking of the upper body and the hands to use in sign language recognition. This is not a trivial task, due to the high requirements for accuracy and usability. We will use sequences of hand and body poses using color cameras and depth sensors (RGBD) or alternatively stereo cameras. We will record offline predefined gestures, and we will estimate the poses using color and depth data. The poses along with the images will be used as input to the next learning step.

In our attempt to improve our method we will use higher resolution data, stereo images, as well as methods for the detection of fingertips. To this end we will use the Kinect-2 sensors which offer higher image resolution and frame rate, machine vision cameras or narrow baseline stereo cameras with high frame rate. We will also develop an initialization and reinitialization procedure using discriminative methods. Such a method can be based on neural networks for the

direct estimation of the pose. The availability of additional information such as the position of the fingertips can contribute to initialization accuracy and better tracking. The goal is to learn the function that associates the body/hand pose with color and depth just for offline pose estimation. The online association of image with pose will have the form of a probability density function.

2.2 Sign language modeling and recognition

We will do the modeling by combining (a) deep networks (b) the tracking results and (c) linguistic constraints. The deep networks provide state of the art performance and the linguistic constraints along with the tracking results are expected to drive the optimization close to optimal solutions.

An initial approach will employ available networks, which have been trained with the ImageNet data set. The output sequences will be classified with a conditional random field using the L-BFGS algorithm. We will experiment on how to approximate the Hessian matrix. Our second approach will use data from body/hand tracking without using depth. Because (a) we cannot expect the mainstream devices to have depth sensors and (b) depth is not really necessary,beyond the training stage. From tracking we will get the conditional probability of the pose given the image. The pose can be used for sign language recognition with distance matrix regression, which offers some advantages like: (a) detailed generative model which renders the system more tractable (b) insertion of linguistic constraints like the initial/final hand/body/face configurations, which may be easily integrated as Bayesian priors.

2.3 SLR solution implemented on dedicated servers

The proposed platform will integrate the software tools to be developed as described in the previous sections (Fig. 1). It will also include a user interface (a) for the deaf user, to display the physician and their responses as text (using a commercial speech recognition software) and (b) for the physician to display the deaf user and the related signs translated to text, or as voice using a text-to-speech software.

The internet communication platform will support the transmission of video via WebRTC on android devices. It will support the interface parametrization, as well as a lot of web services. It has been tested as a Bioassist product (Heart Around). It will be modified to support the aforementioned functionalities. The goal is to perform the whole cycle (recording-interpretation-presentation) in less than 5 seconds.

2.4 Embedded SLR solution implemented on mobile devices

We propose the development of SLR on platforms like Qualcomm Snapdragon or NVidia GPU for local real time processing on mobile devices. The implementation of deep learning networks on embedded systems faces a lot of challenges.

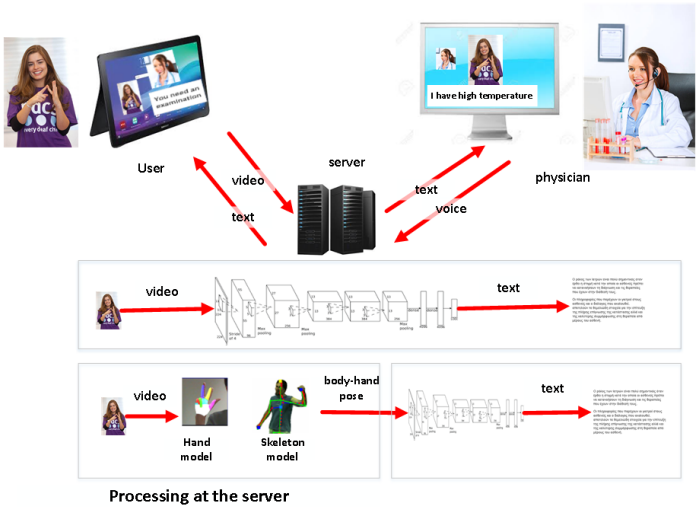

Fig. 1. The HealthSign architecture for mobile devices with the processing on the server. The processing consists in the direct interpretation of the video which is transmitted from the user device. It is then transformed to text directly or after the extraction of the hand/body pose. The physician's speech is translated to text on the server side using a speech recognition software and then transmitted to the deaf user.

These networks are quiet complex and require a lot of processing power for real time tasks. There are two main approaches (a) efficient programming model and (b) simplified networks.

In that scenario the video frames are not transmitted to the server, but are processed locally. The sign language is interpreted to text and the text is sent to the server where it becomes transformed to speech by a text-to-speech tool. The audio is sent to the physician. The physician talks to the patient and the speech data are transmitted to the server, where they are transformed to text. The text is sent to the deaf's clinet device and appear on its screen.

3 Current Status

At this early stage of the project, we have made progress in two fronts: the collection of data and the hand pose estimation. As already mentioned, we focus in the communication between the physician and the patient. To this end, we have narrowed the scope of data collection in the case of distressed patients visiting a psychiatrist. We have collected scripts, which contain spontaneous dialogue, and we further simplified and transcribed them using the proper annotation of the GSL. We currently record videos of the transcribed scripts, with the help of SL interpreters and native speakers. Besides that, we have recorded a set of images depicting the nineteen most important hand shapes of the GSL in nine common hand orientations. As an example, see Fig. 2.

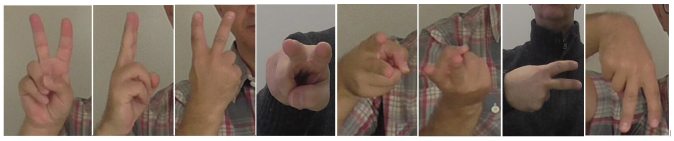

Fig. 2. The "V" handshape in nine common orientations.

Regarding hand pose estimation, our inital attempts to estimate the geometry of the hand pose yielded promising results. Using an extension of our puiblished method, we managed to recognise and track challenging handshapes. A couple of the achieved results are depicted in Fig. 3.

Fig. 3. Two examples of the estimated hand geometry.

4 Conclusions

We have introduced the basic concepts behind the HealthSign project. We have presented the basic elements of the proposed architecture and the principles of their implementation. We presented the two alternative architectures that we are going to develop, the first doing the processing on the server and the other one doing the processing on the mobile device. In the near future we are going to begin the implementation and experiments with real data, after concluding the ongoing video recordings. There are a lot of challenges to deal with, the most obvious being the modeling of the signs in a continuous form and the real time operation.

The interested reader can be informed about new developments via the dedicated project web-page (http://xanthippi.ceid.upatras.gr/HealthSign).

Acknowledgement Co-financed by the Greek Secretariat for Research and Technology and the EU, Project HealthSign: Analysis of Sign Language on mobile devices with focus on health services T1EΔK-01299 within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020.