https://doi.org/10.2352/ISSN.2470-1173.2020.1.VDA-389

Human Computer Interface Based on Tongue and Lips Movements and its Application for Speech Therapy System

Zuzana Bílková, Adam Novozámský, Michal Bartoš, Adam Domínec, Šimon Greško, Barbara Zitová, Markéta Paroubková, Jan Flusser; The Czech Academy of Sciences, Institute of Information Theory and Automation, Pod Vodárenskou věží 4, Praha 8, 182 08, Czech Republic; Charles University in Prague, Faculty of Mathematics and Physics, Ke Karlovu 3, Praha 2, 121 16, Czech republic

Abstract

The novel human computer interface is introduced, based on tongue and lips movements and using video data from a commercially available camera. The size and direction of the movements are extracted and can be used for setting cursor actions or to other relevant activities. The movement detection is based on convolutional neural networks. The applicability of the proposed solution is shown on the ASSISLT system [1], aimed to support speech therapy for adults and children with inborn and acquired motor speech disorders. The system focuses on individual treatment using exercises that improve tongue motion and thus articulation. The system offers an adjustable set of exercises which proper performance is motivated using augmented reality. Automatic evaluation of the performance of therapeutic movements allows the therapist to objectively follow the progress of the treatment.

Introduction

The human computer interface can have various forms. Next to widespread keyboards and mouses, there are other devices using haptic inputs [2], EEG [3], eye movements [4], to name a few examples, aimed at users with motor disabilities or at situations when hands cannot be used.

Tongue movements have been exploited in the past, too, however they were detected using specialized hardware attached to the tongue [5] which can be burdensome and even non hygienic. Moreover, having the tongue movement detection tool without any necessary specialized hardware, it can be utilized in the area of speech therapy, when often at the first stages patients with speech disorder or dysarthria are asked to strengthen and improve their control over the speech producing muscles. This is realized using specialized tongue and lips exercises and the detection tool can help to motivate and to evaluate the patient performance and the progress of the therapy.

Objective

The goal of the project was to develop the human computer interface based on video data from a commercially available camera and applicable in the general public environment. The size and direction of the movements are extracted from the video data using digital image processing methods and they are then used for setting cursor actions or to other relevant activities. Linking the tongue movement and characters in a computer game will enable to control the game and thus to motivate children in the muscle strengthening.

The tongue movement detection is complicated by high variability of skin and tongue coloring and the low discriminability of these two objects of interest. Moreover, for the speech therapy application, the computational time cannot be too high to be able to efficiently evaluate patient's performance. To increase the usefulness of the proposed solution we have implemented the lips movement detection and teeth detection, too. Many exercises in the speech therapy are based on the lips and teeth thus the system will offer higher functionality to the speech therapists.

Method

In this section we first describe the detection of lips, which is performed using the results from the Dlib library, followed by tongue movements detection based on convolutional neural networks and finally the teeth detection.

In the preliminary step, a patient's mouth is located in the face image. The detection of face parts is realized by applying the modified Dlib library [6], which is based on an ensemble of regression trees that are trained to estimate the facial landmark positions directly from the pixel intensities. The Dlib library automatically detects 68 points in the face image in real time allowing detection of the mouth and the lips. The face is then normalized to a vertical position using detected points of eye corners.

Lips detection

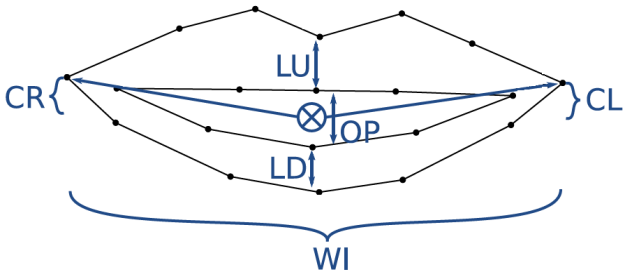


Figure 1. *Key points detected by Dlib library [6] and features proposed for analysis of lips exercises: CR/CL - vertical shifts of the mouth corners, OP - lips distance, WI - mouth width.*

The detected lips are described using specially designed features shown in Figure 1, enabling evaluation of the exercises. They are defined as follows: vertical shifts of the mouth corners, heights of the lips, lips distance when the mouth is open and a horizontal distance of the left and right corner. For each therapeutic exercise, there is a characteristic set of features to be maximized and/or minimized, e.g. for the exercise of closed smile we eval-

uate min(CL,CR), -OP, WI. If the exercise is executed correctly, the features are maximized, i.e. both corners are supposed to be lifted, value of -OP is maximized when the mouth is closed and WI captures the width of a mouth which is larger when we smile. Having detected lips in the preliminary step, the values of these features are computed and the feedback is given to the patient and as well stored for further progress evaluation.

The detected position of the face and lips enables to apply augmented reality, when the explanation of the exercise movements are demonstrated and motivated on the PC screen, such as colorful dots at the positions where the lips should be moved. TODO obr? spis bych ho nedavala, uz tam mame Babet s motylem

Tongue detection

The tongue based exercises are using data from the tongue detection module when the segmentation of the tongue body and of its tip is realized. Our solution is based on the convolutional neural network U-net [7] which proved to be sufficiently fast and robust. We use a single neural network to output both results segmentation of the tongue and its tip to achieve higher speed.

The network is trained on data we both recorded and downloaded from the Internet. The data are of different quality and with people of different ages to ensure the robustness of our method due to the complexity and variations of individual tongues. The data were manually annotated.

The input to the network is a cropped image of the mouth normalized to 128x128 pixels. We use 1417 different frames from 70 videos of both children and adults that capture distinct position, structure and colour of the tongue. 60 videos with 1252 frames were used for training, validation was performed with 10 videos and 165 frames. It proves that 200 epochs is enough for convergence.

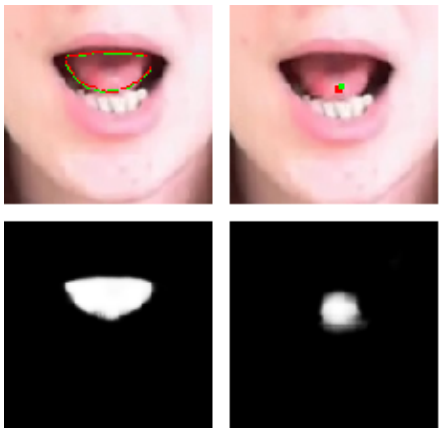


Figure 2. *Output of convolutional neural network U-net for segmentation of tongue and its tip. In the first row green contour and dot represent ground truth and red color represent predictions. Second row shows the output of U-net.*

The network thus two output frames as shown in Figure 2. In Figures 2 and 3 the green lines and dots represent the manually annotated ground truth and the red lines and dots are the predictions. The red dots are the center of the mass of a mask predicted by the network, as shown in right bottom image in Figure 2. We can see that the results correspond well to the ground truth and are robust for all the tongue positions in the mouth.

The position of the detected tongue and its shape parameters are used for the evaluation of the exercises. Moreover, the augmented reality motivation approach is applied here too, i.e. catching a butterfly with the tip of the tongue as shown in Figure 4.

Teeth detection

A new need to know the position of the teeth arose during the design of the ASSISLT system because some speech therapy exercises require e.g. either clenched teeth or teeth to be hidden behind lips. The detection of teeth is based on their segmentation and evaluation of their position and the gap in-between them.

The input RGB image from a camera is transformed into a color space described in [8]. This color space emphasizes the red channel, therefore a combination of the inverted transformed pic-

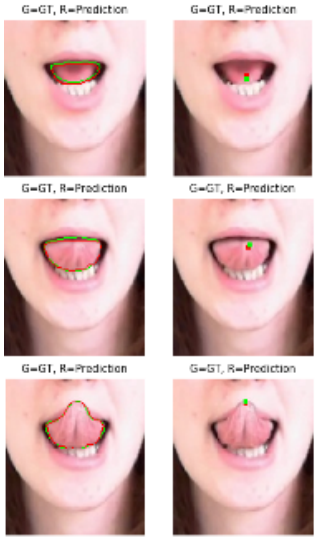


Figure 3. *Visualization of ground truth and predicted tongue contours and tongue tip.*

Figure 4. *Example of a tongue exercise motivated by augmented reality.*

ture and the green component of the original picture emphasizes the teeth.T he process is shown in Figure 5. The image is then cropped to mouth location and masked based on points detected by Dlib library. The area outside of mouth is masked to zero, as shown in Figure 6. In the next step we compute autmatically a threshold for the masked image to segment teeth.

We then compute the sum of each row of the thresholded picture and again threshold it by one third of maximum value of this function, as shown in Figure 7. This step ensures robustness of the method, e.g. small very bright reflections that occur on tongue or jowl, which could be falsely marked as teeth, are eliminated. To detect whether the teeth are closed or open we compute differences of values -1, 0 and 1 and evaluate the number of negative differences, see Figure 7.

Interface

The vector of the change of the tongue tip position described in Section Tongue detection can be used for controlling the character in a computer game. The speech therapy software offers a mode which enables to control the keyboard arrows by tongue movement. This mode can be used in any game controlled by arrows which enables strenghtening of tongue, lips and mouth muscles in an attractive way for the children patients. This human computer interface is also applicable in other areas such as a device for people with motor disabilities or quadriplegics.

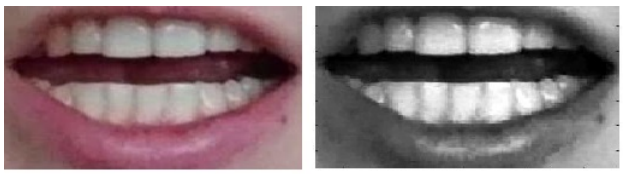


Figure 5. *Original RGB input on the left is transformed into a special color space, the combination of the inverted transformed input and the green component of the original picture is shown on the right.*

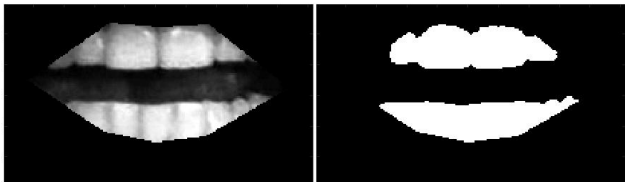


Figure 6. *The transformed image is masked to the area inside of the lips and thresholded to detect the teeth.*

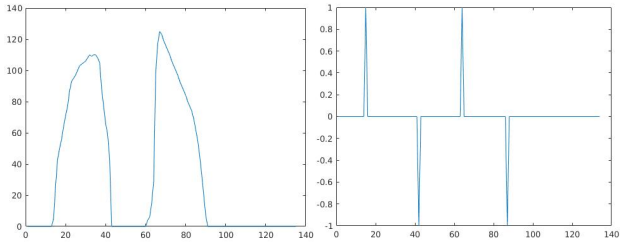


Figure 7. *Left figure shows row sums of the thresholded image of teeth and diferences are computed and ploted in the right figure.*

Results

The interface was tested separately for the tongue and lips motion detection. Testing of the features used for lips exercise evaluation was realized using manually annotated data to analyze the feature performance. The position of the lips was detected manually and the proposed features were evaluated and plotted to check if they reflect the expected value characteristics.

Figure 8. *Value of individual features in a video of performance of selected lips exercises.*

Figure 9. *Images of selected exercises - closed smile, open smile and crooked smile.*

Figure 8 shows the values of four different exercises closed smile, open smile and crooked smile with only left or right mouth corner up, shown in 9. The values of four features wide, open, left/right corner, as shown in Figure 1 are shown for each video frame. We can see that the chosen features correctly capture the desired pattern of the exercises, e.g. for the closed smile the high values for width and height of both corners and no significant change in the feature describing opening and similar expected behavior for the other exercises.

An important role of the features is to be able to separate the individual speech therapy exercises. To demonstrate the separability of the four exercises, we plot pairs of features to show the different values for different exercises. Figure 10 shows the separability of the exercises if we compare the position of left and right corners. We can see that the values are high for closed and open smile, low for neutral position and when crooked smile is performed, only one corner has higher values. Similar behaviour of features can be observed in Figure 11 where the relation between smile and width of the mouth is shown. Figure 12 showing features of width and opennes indicates that the exercise of open smile can be easily separated from the rest of the exercises using these two features.

The evaluation of tongue detection is briefly described in the Section Tongue detection and it is still in progress, as well as the

evaluation of detection of teeth.

Novelty

The presented method for the human computer interface, based on CNNs, enables automatic evaluation of speech therapy exercises. The method is based on the detection of the lips, tongue movements and teeht. The speech therapy system can offer an adjustable set of exercises recommended by a therapist, motivation by using augmented reality, and evaluation of the performance of therapeutic movements. Linking the tongue movement and characters in a computer game will motivate children. Besides, the proposed human computer interface is also applicable in other areas, such as controlling various devices by users with motor disabilities (quadriplegics) or in situations when hands cannot be used.

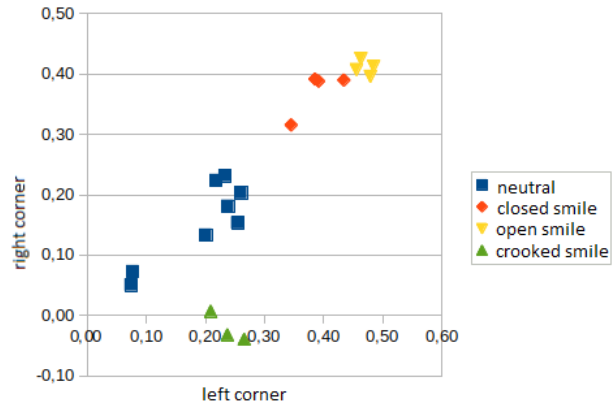


Figure 10. *Values of features representing left and right corner.*

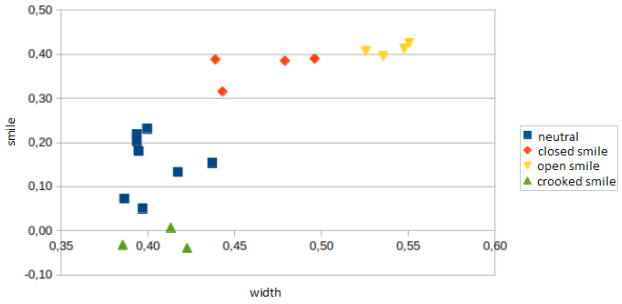


Figure 11. *Values of features representing width and smile.*

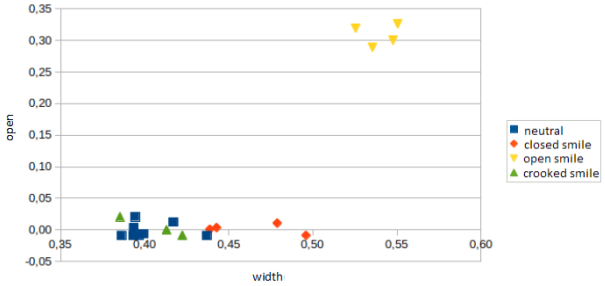


Figure 12. *Values of features representing width and opennes of the mouth.*

Acknowledgments

This work has been supported by the *Praemium Academiae*, by TACR, grant No. TJ01000181, and the grant SVV–2017–260452.

References

[1] Zuzana Bílková, et al., Automatic Evaluation of Speech Therapy Exercises Based on Image Data, Image Analysis and Recognition : 16th International Conference, ICIAR, 397-404, (2019).

[2] Arthur E. Quaid III, System and method for using a haptic device as an input device, U.S. Patent No. 8,095,200, (2012).

[3] Christoph Guger, et al.,Prosthetic control by an EEG-based brain-computer interface (BCI), Proc. aaate 5th european conference for the advancement of assistive technology, pg. 3-6, (1999).

[4] Yoon C. Seok, Input device using eye-tracking, U.S. Patent Application 15/357,184, (2017).

[5] Takuma Hashimoto, et al., TongueInput: Input Method by Tongue Gestures Using Optical Sensors Embedded in Mouthpiece, U2018 57th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), 1219-1224, (2018).

[6] Davis E. King, Dlib-ml: A Machine Learning Toolkit, Journal of Machine Learning Research, 10, 1755-1758, (2009).

[7] Olaf Ronneberger, Phillip Fischer, Thomas Brox, U-net: Convolutional networks for biomedical image segmentation, International Conference on Medical image computing and computer-assisted intervention, pg. 234-241, (2015).

[8] Adam Novozámský, et al., Automatic blood detection in capsule endoscopy video., Journal of biomedical optics, 21(12), (2016).

Author Biography

Zuzana Bílková is a PhD. student at Charles University in Prague and she works in the Czech Academy of Sciences. She is oriented on applications of deep neural networks in image processing, especially in the area of medical imaging. She is a holder of two grants, one from Charles University on convolutional neural networks and the second from the Technology Agency of the Czech Republic, which is meant for excellent young researchers.

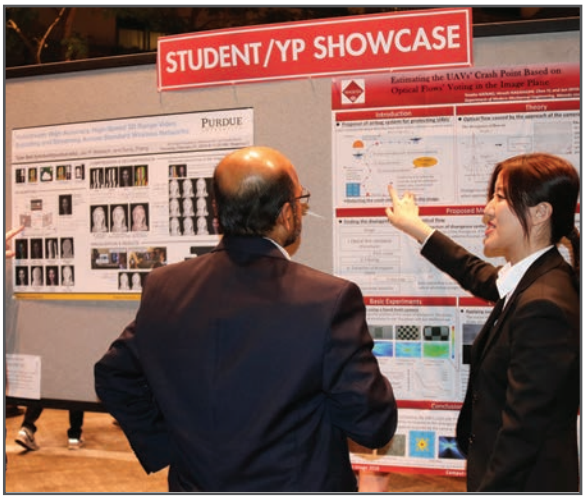
JOIN US AT THE NEXT EI!

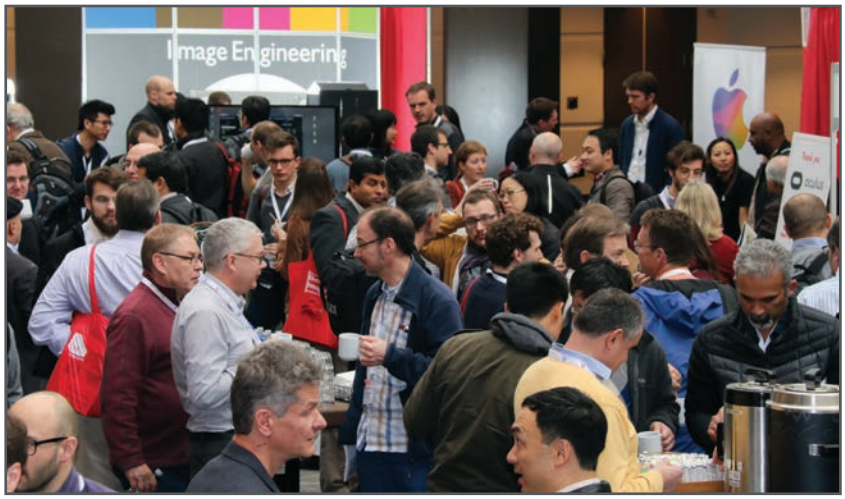
IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!





• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

IS&T
imaging.org