

One-shot learning HMI for people with disabilities

Hanno Homann, Cedric Rohbani, Jens Christian Will

Suggested citation:

Homann, Hanno, Cedric Rohbani, and Jens Christian Will. 2024. "One-shot learning HMI for people with disabilities." *Current Directions in Biomedical Engineering* 10 (4): 319–23. <https://doi.org/10.25968/opus-3543>.

Abstract

For people with physical disabilities, it is often desirable to regain control over their personal environment and communication tools. This paper introduces a novel Human-Machine Interface (HMI) using one-shot learning for individualized control signals without extensive training or specialized hardware. Our work suggests a modular system that utilizes common, easily accessible devices like webcams to interpret user-defined gestures and commands through a single demonstration. As a feasibility study on healthy volunteers, we investigate the control of a computer mouse by head movements only. We demonstrate the technical details of the HMI and discuss its potential applications in enhancing the autonomy and interaction capabilities of users with disabilities. By combining usercentric design principles with the advancements in one-shot learning, we aim to forge a more inclusive, accessible path forward in the development of assistive technologies.

Terms of use

CC BY 4.0

This document is made available under these conditions:
Creative Commons - CC BY - Namensnennung 4.0 International
For more information see:
<https://creativecommons.org/licenses/by/4.0/deed.de>



Hanno Homann*, Cedric Rohbani, and Jens Christian Will

One-shot learning HMI for people with disabilities

<https://doi.org/10.1515/cdbme-2024-2078>

Abstract: For people with physical disabilities, it is often desirable to regain control over their personal environment and communication tools. This paper introduces a novel Human-Machine Interface (HMI) using one-shot learning for individualized control signals without extensive training or specialized hardware. Our work suggests a modular system that utilizes common, easily accessible devices like webcams to interpret user-defined gestures and commands through a single demonstration.

As a feasibility study on healthy volunteers, we investigate the control of a computer mouse by head movements only. We demonstrate the technical details of the HMI and discuss its potential applications in enhancing the autonomy and interaction capabilities of users with disabilities. By combining user-centric design principles with the advancements in one-shot learning, we aim to forge a more inclusive, accessible path forward in the development of assistive technologies.

Keywords: One-shot learning, webcam, neural network, human-machine interface

1 Introduction

Human machine interfaces (HMIs) have revolutionized how we interact with technology. However, for individuals with disabilities, traditional interfaces can pose significant challenges. This paper presents a novel, one-shot learning HMI specifically designed to address these limitations and empower people with disabilities.

Video cameras have shown great potential for body control interfaces e.g. utilizing eye blinks for communication [1], and hand gesture control in music applications [2]. Moreover, eye tracking holds significant potential for human-machine interaction [3].

For electronic wheelchair control, cameras have also shown great potential for HMIs where a classical joystick control cannot be used. Eye-gaze control systems have been implemented [4], as well as head and gesture recognition [5] for this purpose. Even a small set of facial expressions, such as smile or eyebrow raises, has been explored [6]. Further, head and hand gesture-based approaches to control a computer have been realized [7, 8].

While these advancements are encouraging, they require specialized hardware and extensive training or are limited to a specific type of input. Aggarwal et al. [7] use a pre-trained classifier which may be limiting for persons with reduced motion capabilities. In their patient study, Esiyok et al. [8] utilized a set of 6 head poses, which we adopted in the present study. However, they did not consider an adjustment of the pose set to individual needs. This work suggests a modular HMI design that leverages readily available commodity webcams, fostering wider accessibility. Our approach prioritizes user-specific customization, allowing patients to define a minimalist input alphabet based on their capabilities. While this study focuses on the head pose, the method can be extended to other expressions like eye gaze, arm positioning, and even lip movements.

The paper introduces a key innovation: one-shot learning, making our HMI capable of learning user-defined control positions from a single demonstration. This significantly reduces setup time and avoids frustration for users, promoting a more intuitive and user-friendly experience. Similar works have shown the capability of one-shot learning in body action recognition [9] or hand gesture recognition [10].

This general architecture enables diverse applications, including ambient light control, speech synthesis interaction, and accessible gaming experiences, all facilitated through the user's chosen control positions. By combining user-centric design, one-shot learning, and readily available sensor technology, this work aims to create a more accessible and empowering HMI experience for people with disabilities.

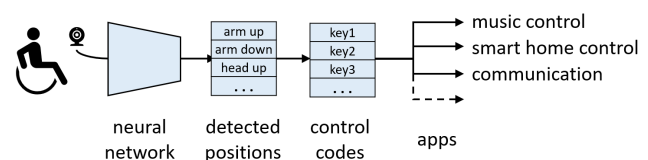


Fig. 1: Concept for a single-shot body controlled HMI: a neural network classifies input images from a webcam into body positions which can be pre-defined according to the user's capabilities. Each position is mapped to a control code for interaction with a set of applications.

*Corresponding author: Hanno Homann, Cedric Rohbani, Jens Christian Will, Hannover University of Applied Sciences and Arts, Germany, e-mail: hanno.homann@hs-hannover.de

Open Access. © 2024 The Author(s), published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

2 Methods

Body-controlled HMI concept

This work explores a novel Human-Machine Interface (HMI) for people with disabilities, utilizing a one-shot learning approach. We propose a modular design that prioritizes flexibility and user-specific customization. The general concept is illustrated in Fig. 1. To enable a cost-effective and wide application, our investigation considers readily available commodity input devices. For this initial investigation, we specifically focus on a webcam as a visual input device, but plan to incorporate microphones as an alternative audio control in the future.

As the input alphabet, a personalized set of body positions is defined depending on the user’s capabilities. The input images are classified according to these positions by a neural network. Each body position is then mapped to a control code to interact with the user’s preferred applications. For example, the user can switch between various applications: control a music and entertainment system or the room’s ambient light. For persons with speech disorders, a speech synthesis system might be provided. For users with fine motor disorders, this input system could replace generic computer input devices such as the mouse and the keyboard.

As a minimum, four different body positions with four corresponding control codes are needed. With this minimal alphabet, a typical user-control menu can be navigated when the 4 body positions are mapped to the control codes “MENU UP”, “MENU DOWN”, “ENTER/SELECT”, and “ESCAPE/EXIT”. In addition, we require a reference position that we refer to as the “rest pose” which has no control code associated. The control set can be extended depending on the user’s capabilities and application type.

In our feasibility study, we investigated the control of a computer mouse by 6 different head positions. These positions are mapped to mouse control events as illustrated in Fig. 2.

Model architecture for one-shot training

Our work emphasizes the one-shot learning paradigm, where the system learns user-defined body positions from a single demonstration. This approach eliminates the need for repetitive training procedures, simplifying usability for individuals with disabilities. For each body position, only a single training image is required.

Another advantage of this sparse dataset is the ability to fine-tune pre-trained neural networks with minimal computational effort. This allows for training on personal devices, such

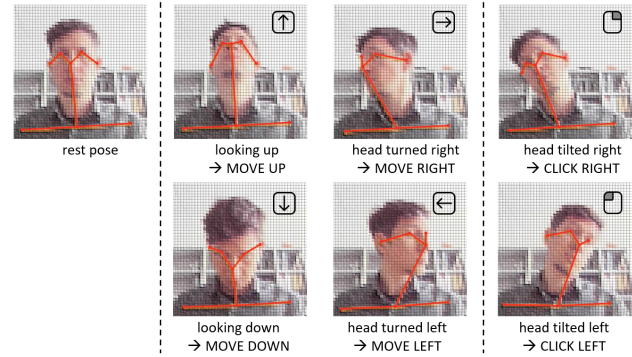


Fig. 2: Set of head positions used to control the computer mouse in this study: The reference pose (“rest pose”, on the left) has no control code associated. Turning the head up, down, left or right is mapped to a movement of the mouse pointer in the respective direction (center). Finally, a tilted head position to the left or to the right is used to trigger a mouse click event (right).

as laptops, without the need to upload training images from private environments. Especially among vulnerable groups, such as individuals with disabilities, local training helps to avoid privacy and ethical concerns. Additionally, local training can be performed without an internet connection, providing further convenience and security.

Two-stage model

To achieve training with scarce data and minimal computational resources, we propose a two-stage model as depicted in Fig. 3. As the first stage, we use a pre-trained pose model to detect body keypoints. The OpenPose model [11] allows for real-time, multi-person keypoint detection of the body, the face, or the hands. Different variants of the model allow for 2D or 3D keypoint detection of the body, the face, or the hands. As the winner of the Coco 2016 keypoints challenge it promises good results and it’s free for non-commercial use. The model has been applied in healthcare applications like fall detection of elderly persons [12] and gait analysis [13].

In this work, we use a lightweight adaptation [14] of OpenPose written in PyTorch and pre-trained on the MS-COCO 2017 keypoints dataset. This model predicts the two-dimensional image coordinates of up to 18 keypoints (ears, eyes, nose, neck, shoulders, elbows, wrists, hips, knees, and ankles).

To achieve scale invariance, we normalized the predicted keypoints by the distance of the eyes. For shift invariance, the difference between each keypoint and its neighbor keypoint is computed, starting from the outermost keypoints of the skeleton towards the center.

The resulting vector of keypoint differences is then flattened and fed into a single fully connected classification layer.

Given the scarce amount of training data, this minimalist structure is designed to avoid overfitting. For keypoints that are hidden in the image or not detected by the keypoint extraction network for some other reason, the respective input is forced to zero. Finally, a softmax activation function is employed to predict the probabilities of the patient-defined body positions.

Training setup

For one-shot training, the model parameters of the pose model are kept fixed. Only the fully connected layer is optimized using the Adam optimizer in PyTorch. The training process runs for 200 epochs with a learning rate of 0.1. A categorical cross-entropy loss is used to maximize the probability of the true class while ignoring the probabilities of the remaining classes.

Despite this restriction, the number of trainable network parameters slightly exceeds the number of data points by N_{bp} , where N_{bp} is the number of pre-defined body positions when using a single input image per body position. Hence, a small risk of overfitting remains. Hidden or undetected keypoints further aggravate this problem. To resolve the issue, we applied a small weight-decay (10^{-4}) to settle unused weights to zero.

Feasibility study

To evaluate the feasibility of the concept, we focused on controlling a computer mouse using head gestures. We defined six commands based on the body positions illustrated in Fig. 2. Turning the head up, down, left, or right moves the mouse in the corresponding direction. Additionally, tilting the head left or right triggers a click of the left or right mouse button, respectively.

Seven healthy volunteers participated in the study (6 male, 1 female; age 28–69 years, mean 46.7 years). Each person was sitting normally in front of a notebook (Dell Precision 5560). For each volunteer, we captured a single image at each of pre-defined head position using the notebook’s built-in RGB camera (1280 x 720 pixels, 30 fps). For real-time inference, we used the notebook’s built-in GPU (NVIDIA T1200). Training

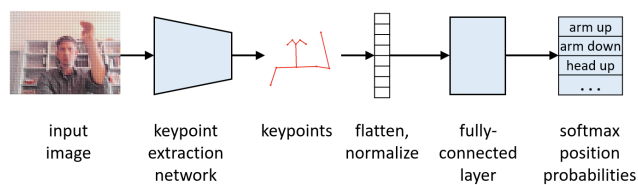


Fig. 3: Neural-network for one-shot classification: A pre-trained convolutional neural network is used to extract the body keypoints. After normalization, a single fully connected layer determines the probabilities of all pre-defined classes.

was conducted using only images of a single volunteer at a time, while the data from the remaining six volunteers were reserved for validation purposes. This way a personalized

3 Results

In the volunteer study, a training accuracy of 100% for all volunteers and an average validation accuracy of 74.3% was achieved. Fig. 4 shows the corresponding confusion matrix of the validation cases, revealing some misclassifications between the reference pose and the downward-looking head pose as well as for the tilted head positions. Such errors are to be expected as the head positions effectively have continuous transitions rather than distinct classes.

While user experience was not systematically evaluated in this initial study, preliminary feedback of our healthy participants indicates that the mouse pointer followed the head directions reasonably well. The direct feedback of the visually observed mouse pointer probably allowed for targeted actions. Conversely, triggering mouse click events repeatedly failed due to the misclassifications described above.

The training was performed on the CPU and finished within a few seconds. At inference, the Python application constantly achieves 30 FPS when running the notebook’s built-in GPU. A typical inference rate of 1.7 FPS was achieved when using only the CPU (8 cores).

4 Discussion

Our results demonstrate the feasibility of a body control interface using a standard webcam with training from single person-specific images per control position. The cross-user validation demonstrated some degree of robustness of the approach. However, user independence is not the original intent of our single-shot method. Personalized training is suggested here, as for persons with disabilities the expressed body positions may differ significantly from reference positions of average persons.

Turning the head toward the major four directions (up, down, left, right) was shown to be practicable as a control input. This could be particularly advantageous for patients with quadriplegia, as it may enable them to regain autonomy in controlling computers, entertainment systems, and smart home appliances.

The visual feedback of the mouse pointer was important for reliable control as the boundaries between the head positions are not distinct but continuous. For that reason, the use of the left/right head tilt as two additional control positions

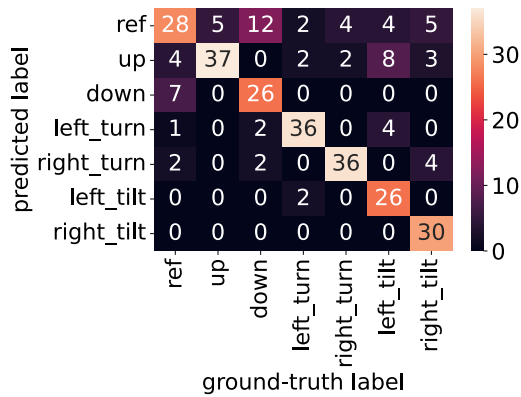


Fig. 4: Confusion matrix of the head position prediction: One volunteer’s true reference position was classified as looking down, the reverse case was observed two times. The tilted and the turned head positions were also confused.

cannot be recommended. Instead, the framework should be extended by further input means depending on the user’s capabilities. This includes classification of the facial expression, such as opening and closing the mouth and the eyes as well as speech commands to trigger control events. Visual or acoustic feedback could increase confidence in the system.

5 Conclusions

This work presented a novel, one-shot learning Human-Machine Interface (HMI) designed specifically for people with disabilities. The proposed concept leverages readily available webcams and prioritizes user-specific customization through a minimalist input alphabet based on individual capabilities.

As an initial use case, we focused on controlling computer mouse operations through head pose. Although the overall results are promising, we identified some limitations, particularly with the misclassification of head tilts, which are not consistently reliable for determining mouse-click events.

Compared to reference works [7, 8] using head and hand gestures in HMIs for persons with physical disabilities, our approach promises a seamless adaption to individually adapted body poses through one-shot learning from a single demonstration.

In future work, we plan to incorporate additional modalities like facial expressions (eye blinks, mouth movements) and potentially eye gaze tracking as well as speech signals to enrich the available control options. User studies with individuals with disabilities will be crucial to assess the effectiveness and usability of the proposed HMI system in real-world scenarios.

A user feedback survey will be needed to provide a quantitative evaluation.

By addressing these aspects, we aim to refine the one-shot learning HMI into a robust and user-friendly assistive technology that empowers people with disabilities to interact more effectively with their surroundings.

Author Statement

Research funding: The author state no funding involved. Conflict of interest: Authors state no conflict of interest. Informed consent: Informed consent has been obtained from all individuals included in this study. Ethical approval: The research related to human use complies with all the relevant national regulations, institutional policies and was performed in accordance with the tenets of the Helsinki Declaration.

References

- [1] MN Mamatha and S Ramachandran. Automatic eyewinks interpretation system using face orientation recognition for human-machine interface. *IJCSNS International Journal of Computer Science and Network Security*, 9(5):155–163, 2009.
- [2] Chris Kiefer, Nick Collins, and Geraldine Fitzpatrick. Phalanger: Controlling music software with hand movement using a computer vision and machine learning approach. In *NIME*, pages 246–249, 2009.
- [3] Christof Lutteroth, Moiz Penkar, and Gerald Weber. Gaze vs. mouse: A fast and accurate gaze-only click alternative. In *Proceedings of the 28th annual ACM symposium on user interface software & technology*, pages 385–394, 2015.
- [4] Mohamad A Eid, Nikolas Giakoumidis, and Abdulmotaleb El Saddik. A novel eye-gaze-controlled wheelchair system for navigating unknown environments: case study with a person with als. *IEEE Access*, 4:558–573, 2016.
- [5] P Jia and H Hu. Active shape model-based user identification for an intelligent wheelchair. *International Journal of Advanced Mechatronic Systems*, 1(4):299–307, 2009.
- [6] Yassine Rabhi, Makrem Mrabet, and Farhat Fnaiech. A facial expression controlled wheelchair for people with disabilities. *Computer methods and programs in biomedicine*, 165:89–105, 2018.
- [7] Suhaani Aggarwal, Austin Paul, Tejashree Bhangale, Sameer Bharambe, and Jyoti More. Gesture-based computer control. In *2023 6th International Conference on Advances in Science and Technology (ICAST)*, pages 470–475, 2023.
- [8] C. Esiyok, A. Askin, and A. Tosun et al. Novel hands-free interaction techniques based on the software switch approach for computer access with head movements. *Univ Access Inf Soc*, 20(3):617–631, 2021.
- [9] Raphael Memmesheimer, Simon Häring, Nick Theisen, and Dietrich Paulus. Skeleton-dml: Deep metric learning for skeleton-based one-shot action recognition, 2021.

- [10] L. Li, S. Qin, Z. Lu, and et al. Real-time one-shot learning gesture recognition based on lightweight 3d inception-resnet with separable convolutions. *Pattern Anal Applic*, 24:1173–1192, 2021.
- [11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.
- [12] Zhanyuan Huang, Yang Liu, Yajun Fang, and Berthold KP Horn. Video-based fall detection for seniors with human pose estimation. In *2018 4th international conference on Universal Village (UV)*, pages 1–4. IEEE, 2018.
- [13] Aditya Viswakumar, Venkateswaran Rajagopalan, Tathagata Ray, and Chandu Parimi. Human gait analysis using openpose. In *2019 fifth international conference on image information processing (ICIIP)*, pages 310–314. IEEE, 2019.
- [14] Daniil Osokin. Real-time 2d multi-person pose estimation on cpu: Lightweight openpose. In *arXiv preprint arXiv:1811.12004*, 2018.