Adapted Contrastive Predictive Coding Framework for Accessible Smart Home Control System Based on Hand Gestures Recognition

Nelly Elsayed
School of IT
University of Cincinnati
elsayeny@ucmail.uc.edu

Constantinos L. Zekios
Dep. of Elect. & Comp. Eng.
Florida International University
kzekios@fiu.edu

Navid Asadizanjani
Dep. of Elect. & Comp. Eng.
University of Florida
nasadi@ece.ufl.edu

Zag ElSayed
School of IT
University of Cincinnati
elsayezs@ucmail.uc.edu

Abstract

Smart home control systems have been widely used to control multiple smart home devices such as smart TVs, smart HVAC, and smart bulbs. While targeting the design of a smart home environment, the accessibility of the entire system is crucial to achieving a comfortable and accessible environment for all home individuals. Thus, in this paper, we are aiming to improve the daily quality of life of the elderly and Disabled individuals by improving the user experience of a vision-based accessible home control system based on an integrated self-supervised and supervised learning framework that provides a robust capability of extracting the features to enhance the overall performance via adapting the contrastive predictive coding concepts and deep learning for real-time hand gesture recognition system. The proposed framework provides an accessible smart home control system with a significant improvement in the user experience.

Home control systems have been designed to improve the quality of daily life by providing a controlling system of different devices and elements at the smart home. There are different smart home control systems that utilize sensors, microphones, cameras, and mobile devices to provide a comfortable control system. Mobile applications for home automation are widely used, and they are the most suitable for remote control of home systems while being outdoors. However, for a home environment, many users prefer device-free zones and no screen, which such a control system can not provide. Voice-activated control systems are commonly used nowadays. However, they are not suitable for the Mute and Deaf individuals. In addition, most of the voice-based systems do not consider linguistic accents and speech disorders, significantly reducing the user experience. Gestures-based home assistance control systems are convenient solutions to provide accessibility for Mute and Deaf individuals as well as for individuals who do not prefer using their voice to control their home devices (Liang 2013). However, most existing methods are limited to specific device controls using static gestures and do not consider the physical limitations of gestures that can be performed by the user, limiting the ability of motor-disabled and elderly individuals to use such systems.

Aiming to improve the quality of life, provide accessibility for wider users, and improve the user experience. In this paper, we propose a novel framework for designing a gesture-based home assistance control system based on user-centered rather than design-based gestures. In addition, we propose a novel gesture recognition approach based on self-supervised and supervised learning via integrating adapted contrastive predictive coding and convolutional neural networks to enhance the feature extraction and improve hand gesture recognition performance, improving overall user experience. The proposed approach can be utilized in various hand gesture recognition tasks and virtual reality applications.

Accessible Smart Home Control System Framework

The proposed accessible smart home control system framework is shown in Figure. 1. The framework aims to integrate the cameras in the smart home to control any smart device rather than each camera controlling only the attached device. Thus, the proposed framework employs different gestures to handle the additional operations that the smart device may have.

Frames Patching

The frame patching performs the preprocessing stage of the input. In this stage, once the camera captures a hand gesture, the frame patching will capture one frame and exclude the repetitive frames with a high structure similarity index to the captured gesture to accelerate the gesture recognition task and manipulate the gesture as a static image. Then, the frame will be divided into a grid of overlapping patches (Xia, Liao, and Yu 2009; Cho, Avidan, and Freeman 2009), where all the patches are of the exact dimensions, such as frame transformation, to adapt the contrastive predictive coding (CPC) for feature extraction from the input images.

Contrastive Predictive Coding Featue Extractor

Contrastive predictive coding (CPC) was introduced in 2018 by Oord et al. (Oord, Li, and Vinyals 2018) as a self-supervised method for representation learning in the tempo-

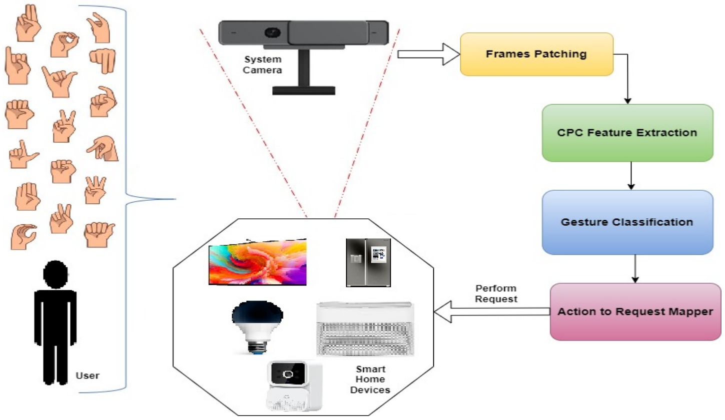

Figure 1: The proposed framework for the smart assistant system is based on hand gestures.

ral data domain. Utilizing the transformed input image via overlapped patching (Henaff 2020; Schmarje et al. 2021), we adapted the CPC to extract the hand gesture features from the static image (Wang et al. 2022). The CPC learns the input data representation via training a neural network to distinguish between the negative and positive data samples such that the model can capture meaningful features.

Gesture Classification Model

A classification model will be trained as a supervised system on the extracted features by the CPC model. Thus, the training process is accelerated compared to the original gesture images. We used the three fully connected neural network layers to perform the classification task (Goodfellow, Bengio, and Courville 2016). The number of classification classes will be based on the number of gestures that will be used to control different devices at the home control system and the number of control operations that the user needs to have for each device. Thus, the total number of required hand gestures for the framework can be calculated by: $\#gestues = \sum_{D=1}^{n} \#operation_D$, where n is the number of devices, $operation_D$ is the number of operations (tasks) that can be performed on device D.

Action to Request Mapper

Our proposed framework is a user-defined gestures-focused system where the system is trained in multiple hand gestures and allows the user to select which gestures of the trained gestures list they prefer to assign to each of the devices and corresponding operations to build up the request mapping. Thus, the initial setup manual considers the default gestures setup by our framework that the user can override to build their preferred manual of hand gestures. Such an option significantly improved the user experience and provided the option to select comfortable and convenient gestures for the Disabled and Elderly. Thus, our proposed approach significantly enhances the accessibility of the gesture-based home control system.

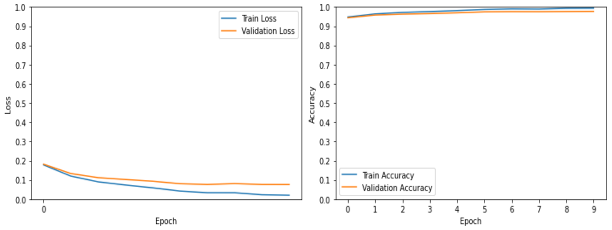

Figure 2: The proposed framework training versus validation loss (left) and accuracy (right).

Results

The initial experiments were performed on ten different gestures for classification tasks. The training versus validation loss and accuracy diagrams are shown in Figure 2. The proposed framework was capable of achieving 97.48% accuracy.

Conclusion

Creating a universal home control system is a challenging task. Speech-based systems nowadays do not handle multiple accents, and speech disorders cause incorrect interpretation of action. In addition, the Mute and Deaf populations cannot use such systems. Thus, vision-based systems can be helpful in such cases. However, not all the gestures are able to be performed by Elderly and Disabled individuals. Thus, in our proposed framework, we aim to propose user-defined gestures to control the house to help disabled individuals with limited motor ability to use simpler and suitable gestures for an accessible home control system.

Acknowledgments

This material is based upon work supported by the College of Education, Criminal Justice, Human Services and Information Technology, at the University of Cincinnati under FY 24 CECH Diversity Research Grant funding.

References

Cho, T. S.; Avidan, S.; and Freeman, W. T. 2009. The patch transform. *IEEE transactions on pattern analysis and machine intelligence* 32(8):1489–1501.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep Learning*. MIT Press.

Henaff, O. 2020. Data-efficient image recognition with contrastive predictive coding. In *International conference on machine learning*, 4182–4192. PMLR.

Liang, S.-F. M. 2013. Control with hand gestures in home environment: A review. In *Proceedings of the Institute of Industrial Engineers Asian Conference 2013*, 837–843. Springer.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Schmarje, L.; Santarossa, M.; Schröder, S.-M.; and Koch, R. 2021. A survey on semi-, self-and unsupervised learning for image classification. *IEEE Access* 9:82146–82168.

Wang, X.; Yang, S.; Zhang, J.; Wang, M.; Zhang, J.; Yang, W.; Huang, J.; and Han, X. 2022. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis* 81:102559.

Xia, T.; Liao, B.; and Yu, Y. 2009. Patch-based image vectorization with automatic curvilinear feature alignment. *ACM Transactions on Graphics (TOG)* 28(5):1–10.