9-1-2023

PyMAiVAR: An open-source Python suit for audio-image representation in human action recognition

Muhammad B. Shaikh
Edith Cowan University

Douglas Chai
Edith Cowan University

Syed M. S. Islam
Edith Cowan University

Naveed Akhtar

10.1016/j.simpa.2023.100544
Shaikh, M. B., Chai, D., Islam, S. M. S., & Akhtar, N. (2023). PyMAiVAR: An open-source Python suit for audio-image representation in human action recognition. Software Impacts, 17, article 100544. https://doi.org/10.1016/j.simpa.2023.100544

Software Impacts 17 (2023) 100544

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

PyMAiVAR: An open-source Python suit for audio-image representation in human action recognition

Muhammad Bilal Shaikh [a,*], Douglas Chai [a], Syed Mohammed Shamsul Islam [a,b], Naveed Akhtar [c]

[a] School of Engineering, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Perth, Australia
[b] School of Science, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Perth, Australia
[c] Computer Science and Software Engineering, The University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Perth, Australia

ARTICLE INFO

Keywords:
Multimodal
Human action recognition
Image representations
Computer vision

ABSTRACT

We present PyMAiVAR, a versatile toolbox that encompasses the generation of image representations for audio data including Wave plots, Spectral Centroids, Spectral Roll Offs, Mel Frequency Cepstral Coefficients (MFCC), MFCC Feature Scaling, and Chromagrams. This wide-ranging toolkit generates rich audio-image representations, playing a pivotal role in reshaping human action recognition. By fully exploiting audio data's latent potential, PyMAiVAR stands as a significant advancement in the field. The package is implemented in Python and can be used across different operating systems.

Code metadata

Current code version	v1
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2023-226
Permanent link to reproducible capsule	https://codeocean.com/capsule/6797263/tree/v1
Legal code license	GNU General Public License (GPL)
Code versioning system used	Git
Software code languages, tools and services used	Python, Librosa, NumPy, matplotlib, ffmpeg
Compilation requirements, operating environments, and dependencies	See the README file on the git repository
If available, link to developer documentation/manual	https://github.com/mbilalshaikh/pymaivar/blob/main/README.md
Support email for questions	mbs.techy@gmail.com; mbshaikh@our.ecu.edu.au

1. Introduction

Within the field of human action recognition, audio emerges as an instrumental data modality, owing to its streamlined structure. It embodies numerous inherent physical properties that, when systematically decoded, can significantly augment the precision and efficiency of the recognition process. Despite the richness of information encapsulated within audio data, it necessitates judiciously designed feature extraction methodologies for its effective utilization.

An array of methodologies for feature extraction from audio data have been proposed in previous works. Nevertheless, the advent and subsequent advancements in deep learning paradigms offer an innovative approach. Specifically, the incorporation of visual representations of audio data, which have been influenced by the notable strides made by vision models in an array of applications such as image classification, object detection, and image segmentation. The amalgamation of auditory data and visual representations heralds a multidimensional understanding of human behavior and actions.

To enhance the multimodality aspect of action data, recent advancements involve merging data from various modalities [1], such as optical flow, RGB-difference, warped-optical flow, and more. These hybrid features can be incorporated into a range of action recognition methods, including Temporal Segment Networks (TSN) [2], Temporal Relation Networks (TRN) [3], and Temporal Shift Modules (TSM) [4]. Inspired by the success of CNN-based models in object detection and

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: mbshaikh@our.ecu.edu.au (M.B. Shaikh), d.chai@ecu.edu.au (D. Chai), syed.islam@ecu.edu.au (S.M.S. Islam), naveed.akhtar@uwa.edu.au (N. Akhtar).

https://doi.org/10.1016/j.simpa.2023.100544
Received 25 May 2023; Received in revised form 25 June 2023; Accepted 1 July 2023

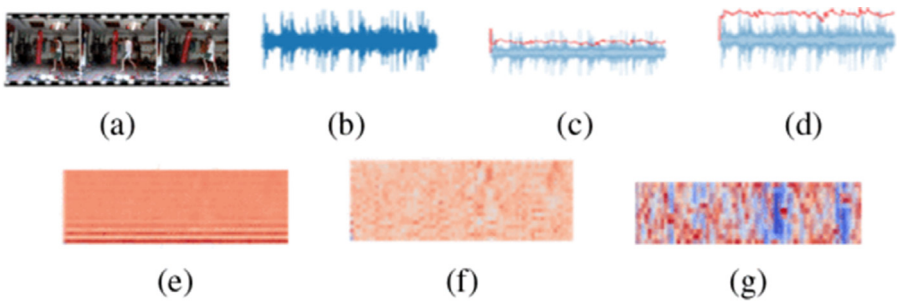

Fig. 1. Six audio-image representation are illustrated (Best viewed in color). (a) Segmented video input and six different audio-image representations of the same action: (b) Wave plot, (c) MFCC, (d) MFCC feature scaling, (e) Spectral centroids, (f) Spectral Rolloff and (g) Chromagram. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: adapted from [11]

image classification, the code explores the capacity of CNNs to comprehend intricate image-based audio representations and their potential to influence multimodal classification.

Given this backdrop, the present paper elucidates PyMAiVAR (Python Multimodal Audio-image and Video Action Recognizer), an innovative tool designed to generate comprehensive and insightful representations of audio data. This tool is geared towards the extraction and projection of distinctive audio features, offering an unparalleled depth of understanding for action recognition tasks. The salient aspect of these features lies in their capacity to seamlessly integrate with both visual and non-visual sensory data, thereby potentially amplifying the precision of action recognition endeavors.

The audio representations (see Figure 1) embedded within PyMAiVAR encompass: (1) Wave plot [5], which provides a temporal depiction of audio signal variations, (2) Spectral Centroid [6], a measure indicative of sound 'brightness', (3) Spectral Roll Off [7], offering valuable insights into the spectral shape of sound, (4) Mel Frequency Cepstral Coefficients (MFCC) [8], effectively distilling the power spectrum of the audio signal, (5) MFCC Feature Scaling [9], a process to normalize the MFCC features thereby optimizing performance, and (6) Chromagram [10], delineating the pitch distribution across the audio signal, analogous to musical notation. This suite of audio representation techniques endows PyMAiVAR with the capability to redefine paradigms in the domain of human action recognition, capitalizing on the unexplored potential of audio data.

We present the software implementation of the PyMAiVAR [12] that:

- Loads the music file given in standard audio formats and extracts the individual representation of the loaded file.
- Implements the audio-image feature and its underlying components.
- Applies standard normalizations over the extracted feature.
- and provides visualization of the extracted features for visual understanding.

These capabilities combine to render our implementation of PyMAiVAR an invaluable tool in the domain of human action recognition through audio-visual data analysis.

2. Key features and functionalities

The PyMAiVAR library, developed using the Python programming language, is openly available on GitHub under the GNU-GPLv3. Its implementation is organized into two main sub-modules, namely "core" and "utils", following best practices for maintainability. The "core" sub-module consists of essential PyMAiVAR functions, while the "utils" sub-module provides helper and wrapper functions. Detailed information about the Python functions and their descriptions can be found in Table 1. Users can refer to the GitHub repository for comprehensive documentation on the library's default parameters and API. Additionally, the default values can be customized to achieve the desired parameter effects.

3. Impact overview

Impacts of using PyMAiVAR are two-fold for action recognition. Firstly, it has demonstrated superior performance compared to the state-of-the-art model on the UCF-101 dataset when evaluated using video representation. Secondly, PyMAiVAR is unique in its ability to handle audio inputs, making it applicable to various tasks without the need for architectural changes. This work proposes a new feature representation strategy to select the most informative candidate representations for audio-visual fusion. Effective audio-image-based representations that complement video modality for better action recognition are also included. PyMAiVAR is proposed for audio-visual fusion that supports different audio-image representations and can be applied to different tasks. Finally, the work by [11] reports state-of-the-art results for action recognition on audio-visual datasets, highlighting the impact of this work in the research community. It is worth noting that PyMAiVAR is the first package which enables audio-image to video fusion-based action classification model. We have used the UCF101 data set [13] to conduct a experiment for human action recognition. Extensive experiments are conducted in the following publications against several features.

1. M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "MAiVAR: Multimodal Audio-Image and Video Action Recognizer", in *Proceedings of International Conference on Visual Communications and Image Processing (VCIP)*, Suzhou, China: IEEE, 2022, pp. 1–5. doi: 10.1109/VCIP56404.2022.10008833 [11]
2. M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "Spectral Centroid Images for Multi-class Human Action Analysis: A Benchmark Dataset". Mendeley Data, 2023. doi: 10.17632/yfvv3crnpy.1. [14]
3. M. B. Shaikh, D. Chai, S. M. S. Islam, and N. Akhtar, "Chroma-Actions Dataset - CAD". Mendeley Data, 2023. doi: 10.17632/r4r4m2vjvh.2. [15]

4. Conclusion

We have presented PyMAiVAR Python library, which introduces the PyMAiVAR feature specifically tailored for human action recognition, drawing inspiration from the chromagram. This library follows Python standards and includes utility functions for added convenience. It is freely licensed, enabling researchers focused on multimodal human action recognition to actively participate in action feature representation studies. We have successfully applied the PyMAiVAR technique to video data, specifically for generating spectral centroid representations. Other researchers and developers can utilize the PyMAiVAR Python library to build their own applications and contribute to its further enhancement. This can involve incorporating smoothing techniques, such as normalization, to improve tolerance towards noise and loudness variations. Additionally, new features can be added to expand

Table 1
Description of API for the PyMAiVAR library. Note: Parameter 'audio' is the string path to the audio file.

Function	Description
core submodule	
get_specfilename(name, folder)	This function formats the name argument, concatenates it with folder, and appends a ".png" extension to create a path where the result image will be saved.
normalize(x, axis=0)	This function returns a normalized version of x along the specified axis using sklearn's minmax_scale() function.
gen_sc(audio)	This function generates a Spectral Centroid visualization. It computes its spectral centroid and plots it on a waveform. The plot is then saved as a PNG file.
gen_mfcc(audio)	This function generates Mel-frequency cepstral coefficients (MFCCs) visualization. It computes its MFCCs and plots them. The plot is then saved as a PNG file.
gen_waveplot(audio)	This function generates a waveplot for an audio file. It computes its spectral centroid and plots a waveplot. The waveplot is then saved as a PNG file.
gen_spec1(audio)	This function generates a spectrogram of an audio file using the linear frequency scale. It computes its spectrogram and plots it. The plot is then saved as a PNG file.
gen_spec2(audio)	This function generates a spectrogram of an audio file using the logarithmic frequency scale. It computes its spectrogram and plots it. The plot is then saved as a PNG file.
gen_specrf(audio)	This function generates a visualization of the spectral rolloff. It computes its spectral rolloff and plots it on a waveform. The plot is then saved as a PNG file.
gen_mfccs(audio)	This function generates a visualization of the scaled MFCCs. It computes its MFCCs, scales them, and plots them. The plot is then saved as a PNG file.
gen_chrom(audio)	This function generates a Chromagram. It computes its chromagram and plots it. The plot is then saved as a PNG file.

the library's capabilities. Looking towards future development, the PyMAiVAR feature, and toolbox can be updated to perform tasks such as distance classification and fusion with non-sensory data within large action collections. Furthermore, there is potential to incorporate the library with action recognition functionality, enhancing its capabilities for spectral centroid and multimodal human recognition.

CRediT authorship contribution statement

Muhammad Bilal Shaikh: Conceptualization, Methodology, Writing – original draft. **Douglas Chai:** Data curation, Writing, Supervision, Funding acquisition. **Syed Mohammed Shamsul Islam:** Review editing, Supervision, Validation. **Naveed Akhtar:** Review editing, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is jointly supported by Edith Cowan University (ECU), Australia and Higher Education Commission (HEC) of Pakistan under Project #PM/HRDI-UESTPs/UETs-I/Phase-1/Batch-VI/2018. Dr. Akhtar is a recipient of Office of National Intelligence National Intelligence Postdoctoral Grant # NIPG-2021–001 funded by the Australian Government.

References

[1] M.B. Shaikh, D. Chai, RGB-D data-based action recognition: A review, Sensors 21 (12) (2021) 4246, http://dx.doi.org/10.3390/s21124246.
[2] L. Wang, et al., Temporal segment networks: towards good practices for deep action recognition, in: Proceedings of The European Conf. on Comput. Vis, (ECCV), Springer, Amsterdam, Netherlands, 2016, pp. 20–36.
[3] B. Zhou, A. Andonian, A. Torralba, Temporal relational reasoning in videos, in: Proceedings of The European Conf. on Comput. Vis, (ECCV), Springer, Munich, Germany, 2018, pp. 831–846.
[4] J. Lin, C. Gan, S. Han, Tsm: temporal shift module for efficient video understanding, in: Proceedings of the ICCV, IEEE, Seoul, South Korea, 2019, pp. 7083–7093.
[5] S. Shilaskar, S. Bhatlawande, A. Vaishale, P. Duddalwar, A. Ingale, An expert system for identification of domestic emergency based on normal and abnormal sound, in: presented at the 2023 Somaiya International Conference on Technology and Information Management, (SICTIM), IEEE, 2023, pp. 100–105.
[6] E. Resendiz, N. Ahuja, A unified model for activity recognition from video sequences, in: presented at the 2008 19th International Conference on Pattern Recognition, IEEE, 2008, pp. 1–4.
[7] F. Hajjej, et al., Deep human motion detection and multi-features analysis for smart healthcare learning tools, IEEE Access 10 (2022) 116527–116539.
[8] D. Oneata, J. Verbeek, C. Schmid, Action and event recognition with fisher vectors on a compact feature set, in: presented at the Proceedings of the IEEE international conference on computer vision, 2013, pp. 1817–1824.
[9] Y. Jung, Y. Kim, H. Lim, H. Kim, Linear-scale filterbank for deep neural network-based voice activity detection, in: presented at the 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, (O-COCOSDA), IEEE, 2017, pp. 1–5.
[10] W. Wang, F. Seraj, N. Meratnia, P.J. Havinga, Privacy-aware environmental sound classification for indoor human activity recognition, in: presented at the Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, 2019, pp. 36–44.
[11] M.B. Shaikh, D. Chai, S.M.S. Islam, N. Akhtar, Maivar: multimodal audio-image and video action recognizer, in: Proceedings of International Conference on Visual Communications and Image Processing, (VCIP), IEEE, Suzhou, China, 2022, pp. 1–5, http://dx.doi.org/10.1109/VCIP56404.2022.10008833.
[12] M.B. Shaikh, Pymaivar, 2023, Accessed: May 22, 2023. [Online]. Available: https://github.com/mbilalshaikh/pymaivar.
[13] K. Soomro, A.R. Zamir, M. Shah, Ucf101: a dataset of 101 human actions classes from videos in the wild, 2012, http://dx.doi.org/10.48550/arXiv.1212.0402, ArXiv Prepr.
[14] M.B. Shaikh, D. Chai, S.M.S. Islam, N. Akhtar, Spectral centroid images for multi-class human action analysis: a benchmark dataset, 2023, http://dx.doi.org/10.17632/yfvv3crnpy.1.
[15] M.B. Shaikh, D. Chai, S.M.S. Islam, N. Akhtar, Chroma-actions dataset - cad, 2023, http://dx.doi.org/10.17632/r4r4m2vjvh.2.