data

MDPI

Article

MHAiR: A Dataset of Audio-Image Representations for Multimodal Human Actions

Muhammad Bilal Shaikh [1,*], Douglas Chai [1], Syed Mohammed Shamsul Islam [2] and Naveed Akhtar [3]

[1] School of Engineering, Edith Cowan University, 270 Joondalup Drive, Joondalup, Perth, WA 6027, Australia; d.chai@ecu.edu.au
[2] School of Science, Edith Cowan University, 270 Joondalup Drive, Joondalup, Perth, WA 6027, Australia; syed.islam@ecu.edu.au
[3] School of Computing and Information Systems, The University of Melbourne, Melbourne Connect, 700 Swanston Street, Carlton, WA 3053, Australia; naveed.akhtar1@unimelb.edu.au
* Correspondence: mbshaikh@our.ecu.edu.au

Abstract: Audio-image representations for a multimodal human action (MHAiR) dataset contains six different image representations of the audio signals that capture the temporal dynamics of the actions in a very compact and informative way. The dataset was extracted from the audio recordings which were captured from an existing video dataset, i.e., UCF101. Each data sample captured a duration of approximately 10 s long, and the overall dataset was split into 4893 training samples and 1944 testing samples. The resulting feature sequences were then converted into images, which can be used for human action recognition and other related tasks. These images can be used as a benchmark dataset for evaluating the performance of machine learning models for human action recognition and related tasks. These audio-image representations could be suitable for a wide range of applications, such as surveillance, healthcare monitoring, and robotics. The dataset can also be used for transfer learning, where pre-trained models can be fine-tuned on a specific task using specific audio images. Thus, this dataset can facilitate the development of new techniques and approaches for improving the accuracy of human action-related tasks and also serve as a standard benchmark for testing the performance of different machine learning models and algorithms.

Keywords: human action recognition; image representations; multimodal dataset; computer vision

Citation: Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. MHAiR: A Dataset of Audio-Image Representations for Multimodal Human Actions. *Data* **2024**, *9*, 21. https://doi.org/10.3390/data9020021

Academic Editor: Giuseppe Ciaburro

Received: 1 August 2023
Revised: 30 October 2023
Accepted: 21 November 2023
Published: 25 January 2024

1. Introduction

The recent progress in deep learning architectures, coupled with enhancements in Graphics Processing Unit (GPU) hardware and software stacks, has significantly empowered the handling of computationally demanding tasks, including Multimodal Human Action Recognition (MHAR). Analyzing human activities in a multimodal information context is a challenging endeavor that necessitates substantial computational resources [1]. This has emerged as a prominent research issue in the field of computer vision. Human Action Recognition (HAR) involves the process of categorizing human actions depicted in a sequence of images, essentially entailing the classification of objectives pursued by individuals across a series of image frames.

Video modality inherently holds spatial information, which lends itself well to Convolutional Neural Network (CNN)-based classification architectures. In the pursuit of more effectively encompassing the multimodal facets of action data, a contemporary approach involves the integration of data from various modalities, including optical flow, RGB difference, and warped optical flow. Audio is a lightweight signal in comparison to video data. However, image-based representations are optimal for vision models in machine learning, specifically for convolution neural network-based vision models. Further, features from spectral centroid-based representations are visually favorable when

compared to convolution-based methods. Spectral Centroids provide a compact and informative representation of the audio signal that captures the discriminative features and temporal dynamics of human actions. Therefore, the dataset described in this manuscript was generated during the process of screening diverse image-based representations for action sequences for multimodal fusion with video data. This dataset can thus be used to analyze critical features from the action sequences in the image form. This dataset extends our previous publication [2] which outperforms state-of-the-art methods producing an accuracy of 91.2% by focusing on multimodal representations of action sequences to present critical features in audio from different perspectives, as captured from each action sample. These datasets were also used as a pre-requisite requirement in developing an intelligent multimodal action recognition system for classifying actions using deep learning algorithms based on acoustic and video modality. To the best of our knowledge, MHAiR is the first audio-image representation dataset for multimodal human recognition that uses image-based representations of audio to leverage CNN and transformer-based architectures for improving action recognition. The key contributions of our work can be summarized as follows:

- We introduce (Multimodal Audio-image Representations), MHAiR, a new multimodal lightweight dataset.
- We build a new feature representation strategy to select the most informative candidate representations for audio-visual fusion.
- We achieve state-of-the-art or competitive results on standard public benchmarks, validating the generalizability of our proposed approach through extensive evaluation.

Value of Data

There are several ways in which this dataset can be valuable compared to the original dataset and in serving other novel use cases. The distinguished characteristics of this dataset are the following:

- It provides a significant reduction in dimensionality. The spectral centroid images represent the frequency content of the audio signal over time, which is a lower-dimensional representation of the original video dataset. This can make it easier and faster to process the data and extract meaningful features.
- It is robust against visual changes. The spectral centroid images are based on the audio signal, which is less affected by visual changes such as changes in lighting conditions or camera angles. This makes the dataset more robust to visual changes and can improve the accuracy of human action analysis.
- It offers standardization as spectral centroid images can be standardized to a fixed size and format, which can make it easier to compare and combine data from diverse sources. This can be useful for tasks such as cross-dataset validation and transfer learning. Hence, this dataset can serve as a standard benchmark for evaluating performance of different machine learning algorithms for human action analysis based on audio signals.
- It is suitable for privacy-oriented applications such as surveillance or healthcare monitoring, which may require analysis of human actions without capturing original visual information. Spectral centroid images provide a privacy-preserving alternative that can still enable effective analysis in applications where audio can be fused and aligned with non-visual sensory datasets such as HH105 and HH125[1] .
- Dataset versatility can facilitate the exploration of different approaches and the development of newer techniques for various applications and an extension of the existing ones.
- Audio images, derived from sound data, when fused with visual data can enhance interpretation, improve noise reduction, augment AR/VR experiences, refine content-based multimedia retrieval, and assist in healthcare applications like telemedicine. How-

[1] https://casas.wsu.edu/datasets/ (accessed on 19 January 2024)

ever, effective fusion requires advanced algorithms and careful attention to challenges such as data alignment, synchronization, and fusion model selection.

The structure of this paper is organized as follows. Section 2 discusses related works. Section 3 describes the key characteristics of the dataset. Section 4 elaborates on the process of extraction of distinct modalities and rationale behind feature extraction in the context of multimodal human action recognition. Section 5 provides an analysis and comparison of a downstream task to establish a benchmark for our proposed dataset, and Section 6 presents the conclusion of this paper.

2. Related Works

2.1. Multimodal Recognition Methods

Feature extraction is a process of yielding critical information from raw instances, which in turn contributes to the learning process. Temporal Segment Network (TSN) is used as a feature extractor based on its temporal pooling of frame-level features, where it is rigorously used as an efficient video feature extractor for different problems. The Gate-Shift Module (GSM) can turn a 2D CNN into a highly efficient spatio-temporal feature extractor. For example, when TSN is plugged into GSM [3], an accuracy improvement of 32% is achieved. Furthermore, Yang et al. [4] used TSN with a soft attention mechanism to capture important frames from each segment. Moreover, Zhang et al. [5] have used the TSN model as a feature extractor with ResNet101 for efficient behavior recognition of pigs.

Recently, TSN has been adapted as a backbone in video understanding scenarios [6–10], and it is typically used in conjunction with a succeeding module. In [10], TSN was employed as a 2D CNN backbone to learn motion dynamics in videos. However, IRV2 has been used for feature extraction from images [11], helping with different image restoration and enhancement tasks [12,13]. In another work, Liu et al. [14] addressed a limitation in existing skeleton-based gesture recognition methods by introducing temporal-dependent adjacency matrices. This innovative approach enhanced the ability of GCN to model temporal information.

2.2. Audio-Image Representations

This subsection describes the six different image representations of audio signals.

2.2.1. Waveplot

A waveplot is a specialized graphical representation predominantly utilized in the field of signal processing and music technology for the analysis of audio data. This plot renders the temporal progression of an audio signal's amplitude, offering a vivid depiction of the audio properties and their fluctuations over time. In the construction of a waveplot, the horizontal axis, or the x-axis, symbolizes the dimension of time, while the vertical axis, or the y-axis, stands for amplitude. The fluctuations in the wave's amplitude, captured over time, generate an illustrative portrayal of the auditory characteristics of the sound, including its loudness and periods of silence. However, it is crucial to acknowledge that a waveplot, while informative, lacks the specificity to offer insights into an audio file's frequency content or pitch. For acquiring a more nuanced understanding of an audio file, analysts often resort to the usage of other types of plots such as spectrograms or mel spectrograms. These advanced graphical representations are capable of illuminating frequency-related information. The waveform provides a visual representation of the audio signal's temporal structure. This can be especially useful for recognizing actions that have distinct audio patterns or start and end abruptly. For example, the visual representation of waveform for a clapping action shows sharp spikes corresponding to claps.

2.2.2. Spectral Centroid

The spectral centroid is a measure of the center of "gravity" of the power spectrum of an audio signal [15]. Mathematically, the value of the spectral centroid (SC), for the kth frame is defined as

$$\mathrm{SC}_t = \frac{\sum_{k=1}^{N} m_t(k) \cdot k}{\sum_{k=1}^{N} m_t(k)}, \quad (1)$$

where SC_t is spectral centroid frequency at time t, k is the kth frequency bin, $m_t(k)$ is the power spectral density value at frequency k, and the summation is taken over all frequency bins. Essentially, this equation calculates the average frequency of a signal weighted by the power at each frequency.

In practice, the spectral centroid is usually computed using the Discrete Fourier Transform (DFT) of a short-time windowed segment of the audio signal. This results in a sequence of spectral centroids over time, which can be further processed and analyzed to extract useful features for various audio signal processing applications.

Overall, spectral centroid-based images provide an efficient, robust, and informative representation of the audio signal that can be used for human action recognition [16]. For example, a higher spectral centroid value often corresponds to a "brighter" or "sharper" sound, while a lower spectral centroid value usually indicates a "duller" or "muddier" sound [17]. By converting the spectral centroid over time into an image, we can capture spatial and temporal information that can be effectively processed by deep learning models [2]. Spectral centroid can also help in distinguishing actions based on their tonal or harmonic characteristics. For example, it can be valuable in recognizing actions involving musical instruments or vocalizations, where the timbre or brightness of the sound varies.

2.2.3. Spectral Rolloff

Spectral rolloff is a measure in digital signal processing that provides an estimation of the frequency below which a specified percentage of the total spectral energy lies. In other words, it is the cutoff frequency where any additional increase in frequency contains less power or energy. Typically, spectral rolloff is expressed as a fraction of Nyquist frequency (half of the sampling rate), and it serves as an important feature in audio analysis for various tasks including music information retrieval, speech processing, and detection of musical onsets and offsets. Rolloff frequency can provide a sense of the bandwidth of the signal. A lower rolloff frequency often indicates a narrower bandwidth or a more tonal signal, while a higher rolloff frequency may suggest a broader bandwidth or a more noisy signal. Spectral rolloff can be relevant for recognizing actions based on the high-frequency content of the audio. For instance, actions that involve high-pitched sounds or actions that have significant energy in the higher frequency range can be distinguished using spectral rolloff.

2.2.4. Mel Frequency Cepstral Coefficients (MFCCs)

Mel Frequency Cepstral Coefficients (MFCCs) are a type of feature widely used in the field of digital signal processing and speech recognition. They provide a representation of the power spectrum of an audio signal that is more aligned with human auditory perception.

MFCCs are based on the known variation for the critical bandwidth of human ear. This variation is often expressed in terms of the Mel scale, which is a perceptual scale of pitches judged by listeners to be equal in distance from one another. Hence, the MFCCs take into account the non-linear human ear perception of frequencies, making them a robust feature for speech and music modeling.

The process to extract MFCCs involves several steps:

- Pre-emphasis: This step is performed to increase the signal's amplitude of the high-frequency part.
- Framing: The continuous signal is divided into frames of N samples, with adjacent frames being separated by M $(M < N)$.
- Windowing: Each frame is multiplied by a window function (Hamming window, for instance).

- Fast Fourier Transform (FFT): This step is taken to convert each frame from the time domain to the frequency domain.
- Mel Filter Bank Processing: The power spectrum is then multiplied with a set of Mel filters to obtain a set of Mel-scaled spectra.
- Discrete Cosine Transform (DCT): Finally, the log Mel spectrum is transformed to the time domain using the DCT. The result is called the Mel Frequency Cepstral Coefficients.

In the context of human action recognition, MFCCs can provide information about the rhythm, tempo, and acoustic cues related to actions.

2.2.5. MFCC Feature Scaling

MFCC Feature Scaling is a normalization process used when working with Mel Frequency Cepstral Coefficients (MFCCs) in machine learning applications, particularly in audio and speech processing.

The goal of feature scaling, also known as data normalization, is to normalize the range of feature values in order to promote computational efficiency and reduce the potential impact of the so-called "curse of dimensionality". This is especially critical in machine learning models, such as neural networks, where features with different scales can have a detrimental impact on the learning process.

When applied to MFCCs, feature scaling might take a couple of forms:

- Standardization: This technique scales the MFCC features so they have the properties of a standard normal distribution with a mean of zero and a standard deviation of one. This is achieved by subtracting the mean and then dividing by standard deviation.
- MinMax Scaling: Also known as normalization, this technique rescales the features to a fixed range, usually 0 to 1, or -1 to 1. The scaler subtracts the minimum value in the feature and then divides by range (max value—min value).

By applying MFCC Feature Scaling, it is possible to optimize the performance of machine learning models by ensuring that all MFCC features contribute equitably to the model's learning, preventing features with larger scales from dominating those with smaller scales.

2.2.6. Chromagram

A chromagram is a graphical representation of the chroma feature of an audio signal, utilized extensively in the field of music information retrieval. The term "chroma" pertains to the 12 different pitch classes in music, which correspond to the traditional Western music scale. In other words, it refers to the color of music that offers a sense of key and harmony.

A chromagram visually represents that the intensity of these pitch classes changes over time in a piece of music. Each row in a chromagram corresponds to one of the 12 pitch classes, and the columns correspond to points in time. The color or intensity at each point in the plot shows the degree to which that pitch class is present in the sound at that moment in time.

Generating a chromagram involves several steps:

- The audio signal is first converted into the frequency domain using Fourier Transform or a similar method.
- The resulting spectral information is then mapped onto the 12 pitch classes in an octave using a filter bank tuned to chroma frequencies.
- Over time, a 2D representation (time-pitch intensity) is obtained.

There are several benefits of using image-based representations for human action recognition, including:

- Efficient representation: Spectral centroid-based images provide efficient representation of the audio signal that can be easily processed by deep learning models. Unlike raw audio signals, which can be difficult to process due to their high dimensionality and variability, spectral centroid-based images provide a compact and informative representation that captures temporal dynamics of the audio signal.

- Robustness to noise: Spectral centroid-based images are less sensitive to noise and distortions than other audio features, such as the raw audio signal or Mel-frequency cepstral coefficients (MFCCs). This is because spectral centroids capture the "center of gravity" of the frequency content, which is less affected by noise and distortions than the fine-grained details of the audio signal. This makes them suitable for noisy environments where other audio features might be unreliable.
- Spatial information: Spectral centroid-based images provide spatial information that can be used by deep learning models to recognize human actions. By converting the spectral centroid over time into an image, we can capture the spatial and temporal information of the frequency distribution of the audio signal, which can be interpreted by deep learning models to recognize different human actions.
- Transfer learning: Spectral centroid-based images can be used for transfer learning, where pre-trained models can be fine-tuned on a specific task. This is because spectral centroid-based images provide a standardized and efficient representation that can be used to compare and combine data from dissimilar sources. This can be useful for tasks such as cross-dataset validation and transfer learning, where models trained on one dataset can be applied to another dataset.

3. Data Description

Data in this study were arranged in two directories: one for training and another for evaluating the model. Audio samples were extracted from videos lasting an average of 10 s (6837 samples overall with 4893 for training and 1944 for testing) [18].

Figure 1 shows a sample of action with different image representations. Images in both training and testing folders were organized in a format of {category}_{action group}_{sample number}.{file extension}, i.e., in ''ApplyEyeMakeup_g08_c01.png'', ApplyEyeMakeup is the class followed by ''g08", which is the supergroup of the sample, and then ''c01" is the sample number for this particular action class. The statistics describing all image representation samples employed in this experimental setting, including action class and a number of samples, are reported in Table 1.

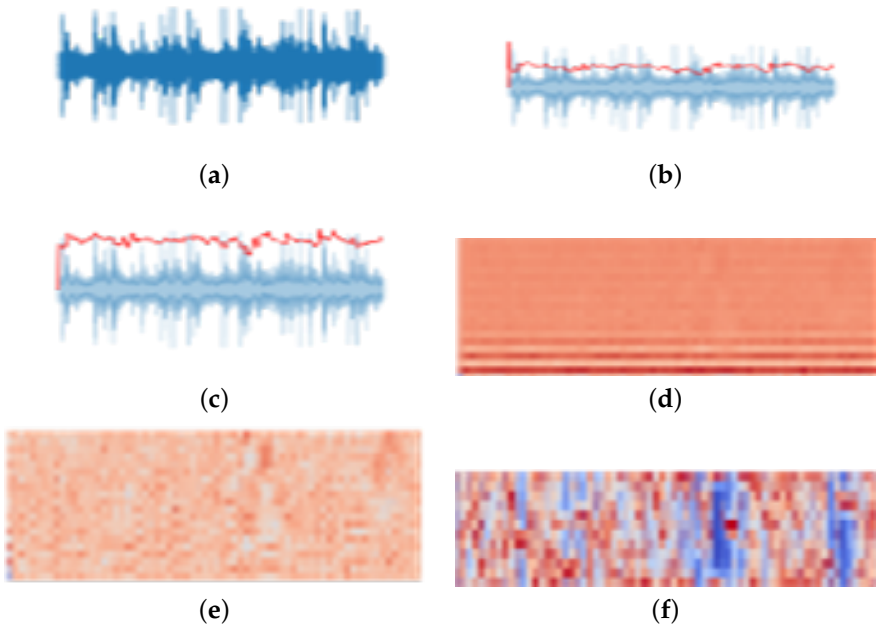

Figure 1. Six different audio-image representations of the same action. Each image represents different characteristics of the same audio signal (adopted from [19]). (**a**) Waveplot. (**b**) Spectral Centroids. (**c**) Spectral Rolloff. (**d**) MFCCs. (**e**) MFCCs Feature Scaling. (**f**) Chromagram.

Table 1. Statistics describing the image representations employed in the experimental setting: for all considered categories, we report the total number of training and testing samples.

#	Category	Train	Test	#	Category	Train	Test
1	HeadMassage	106	41	27	BandMarching	112	43
2	BoxingSpeedBag	97	37	28	CricketBowling	103	36
3	HandstandWalking	77	34	29	Basketball Dunk	90	37
4	Rafting	83	28	30	PlayingDaf	110	41
5	ApplyLipstick	82	32	31	FrisbeCatch	89	37
6	ParrallelBars	77	37	32	BodyWeightSquats	82	30
7	Haircut	97	33	33	Hammering	107	33
8	Typing	93	43	34	SumoWrestling	82	34
9	BoxingPunchingBag	114	49	35	CuttingInKitchen	77	33
10	StillRings	80	32	36	Archery	104	41
11	CricketShot	118	49	37	Mopping Floor	76	34
12	SkyDiving	79	31	38	Shotput	98	46
13	WritingOnBoard	107	45	39	HammerThrow	105	45
14	BlowingCandles	76	33	40	CliffDiving	99	39
15	IceDancing	112	46	41	PlayingSitar	113	44
16	BalanceBeam	77	31	42	BrushingTeeth	95	36
17	AppyEyeMakeup	101	44	43	WallPushups	95	35
18	TableTennisShot	101	39	44	Surfing	93	33
19	PlayingDhol	115	49	45	BabyCrawling	97	35
20	HandStandPushups	96	28	46	Bowling	112	43
21	UnevenBars	76	28	47	FrontCrawl	100	37
22	Playingflute	107	48	48	ShavingBeard	118	43
23	Playing Cello	120	44	49	LongJump	92	39
24	Floor Gymnastics	89	36	50	FieldHockeyPenalty	86	40
25	BlowDryHair	93	38	51	Knitting	89	34
26	SoccerPenalty	96	41				

4. Methodology

A high-level schematic of a prospective downstream multimodal task is illustrated in Figure 2. Audio samples for this dataset of human actions were extracted from videos with a sampling rate of 22,050 Hz. The process of extracting audio from UCF101 video dataset used the "ffmpeg" tool. The resulting audio file was saved separately. For each image representation, post-processing and metadata handling were applied. Particularly following best practices, for chromagram-based representation, a hop length of 512 was used. The extracted audio files were organized and stored according to UCF101 splits, and a quality control check was performed to ensure the audio met the desired standards. This process allowed for the isolation of the audio component from video data, making it available for various applications, including multimodal action recognition and standalone audio analysis.

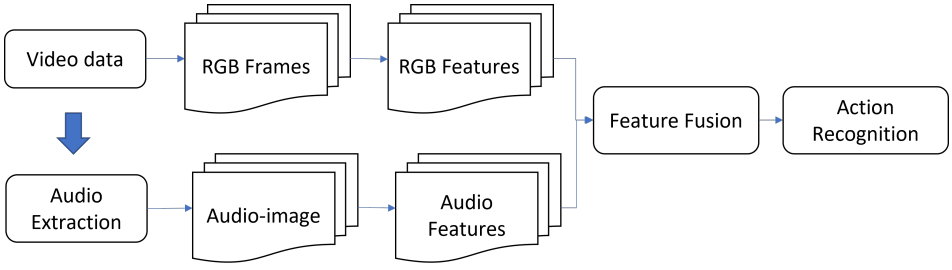

Figure 2. High-level schematic representation of our approach.

These features were then projected onto images that could be processed by Convolutional Neural Networks (CNN) such as (IRV4) [20] or Transformers such as (AST) [21]. Samples that did not have any audio channels were removed from consideration. In to-

tal, 51 categories were analyzed to represent the audio-image features extracted from the audio signals.

Since the dataset delineates experimentations on human action recognition in daily life scenarios, all daily life actions occurring in action recognition were retained in order to inform the models (e.g., through fine-tuning) on specificities characterizing the audio at hand. Data were thus preserved in raw format, whereby no form of image normalization was undertaken, and no forms of pre-processing were applied to the collected data. No Data Augmentation (DA) approaches were adopted (such as horizontal flipping) in order to prevent injecting any kind of noise into the sample and to ensure the inclusion of extensively trimmed action sequences. DA was customarily performed through Rotation [22], Flipping [23], Cropping [24], Scaling [25], Translation [26], Noise Injection [27], Color Modification [28], and other modes. Carefully selecting appropriate data augmentation techniques ensures that modified images are still representative of the original dataset and do not introduce any unwanted biases. These types of processing can be easily completed with off-the-shelf software libraries, according to specific application needs by starting from our data.

5. Results

In the context of multimodal action recognition, as in Multimodal Audio-image and Video Action Recognition (MAiVAR) framework [2], these data are utilized, and they demonstrate superior performance compared to other audio representations. The study establishes a benchmark approach for using this dataset. According to Table 2, the data illustrate the performance of multimodal deep learning models using different audio representations, namely Waveplot, Spectral Centroids, Spectral Rolloff, and MFCCs. These representations are used in two scenarios: audio only and fusion of audio and video. Waveplot Representation shows mediocre performance in the audio-only scenario (12.08) but excels when combined with video, reaching a performance of 86.21 in the fusion scenario. However, Spectral Centroids Representation performs poorly in the audio-only scenario (13.22) but improves when combined with video, achieving a performance of 86.26 in the fusion scenario. In addition, Spectral Rolloff representation performs slightly better than the previous two in the audio-only scenario (16.46). Lastly, MFCC representation shows deficient performance in the audio-only scenario (12.96), and its performance in the fusion scenario (83.95) is also lower compared to that of other representations. In summary, all representations perform significantly better in the fusion scenario, indicating that the combined use of audio and video data enhances the effectiveness of these models. MFCCs representation, however, seems to be less effective when combined with video data compared to the others. This indicates that preprocessing steps for audio representations might play a crucial role in improving the model's performance.

Table 2. Comparing audio-image representations before and after fusion based on accuracy in percentage (adopted from [2]). Note: video-only accuracy is 75.67%.

Representation	Audio	Yield
Waveplot	12.08%	+10.54%
Spectral Centroids	13.22%	+10.59%
Spectral Rolloff	16.46%	+10.33%
MFCCs	12.96%	+8.28%

Finally, our previous work in [2] reports state-of-the-art results for action recognition on audio-visual datasets, highlighting the impact of this work in the research community. We use this dataset [29] to conduct an experiment for human action recognition. Extensive experiments are conducted in the following publications listed in Table 3 against several features.

Table 3. Prior publications produced using the proposed dataset.

Ref.	Work	Venue
[19]	Multimodal Fusion for Audio-Image and Video Action Recognition	Neural Computing and Applications
[2]	MAiVAR: Multimodal Audio-Image and Video Action Recognizer	International Conference on Visual Communications and Image Processing (VCIP)
[30]	PyMAiVAR: An open-source Python suit for audio-image representation in human action recognition	Software Impacts
[29]	Spectral Centroid Images for Multi-class Human Action Analysis: A Benchmark Dataset.	Mendeley Data
[31]	Chroma-Actions Dataset—CAD.	Mendeley Data
[32]	Waveplot-based Dataset for Multi-class Human Action Analysis	Mendeley Data
[33]	Spectral Rolloff Images for Multi-class Human Action Analysis: A Benchmark Dataset	Mendeley Data
[34]	MFFCs for Multi-class Human Action Analysis: A Benchmark Dataset	Mendeley Data
[35]	MFCCs Feature Scaling Images for Multi-class Human Action Analysis: A Benchmark Dataset	Mendeley Data

We conducted comprehensive experiments on the proposed datasets and the results were derived against various features discussed in our prior publications listed in Table 4.

Table 4. Classification accuracy of MAiVAR using Chromagram representation and comparison to the state-of-the-art methods on the UCF51 dataset after fusion of audio and video features.

Year	Method	Accuracy (%)
2015	C3D [36]	82.23
2016	TSN (RGB) [37]	60.77
2017	C3D + AENet [38]	85.33
2018	DMRN [39]	81.04
2018	DMRN [39] + [40] features	82.93
2020	Attention Cluster [41]	84.79
2020	IMGAUD2VID [42]	81.10
2022	STA-TSN (RGB) [4]	82.1
2022	MAFnet [40]	86.72
2022	MAiVAR-WP [2]	86.21
2022	MAiVAR-SC [2]	86.26
2022	MAiVAR-SR [2]	86.00
2022	MAiVAR-MFCC [2]	83.95
2022	MAiVAR-MFS [2]	86.11
2022	MAiVAR-CH [2]	87.91
Ours	**MAiVAR-T [43]**	**91.2**

6. Conclusions

In conclusion, this paper presents an innovative dataset comprising spectral centroid images representing human actions, derived from audio signals of the UCF101 video dataset. These spectral centroid images provide a compact and information-rich representation of the temporal dynamics of human actions, making them robust to noise and distortion and highly suitable for diverse applications such as surveillance, healthcare monitoring, and robotics.

Moreover, the unique characteristics of the dataset allow for it to serve as a robust benchmark for assessing the efficacy of various machine learning models in human action recognition tasks. It also provides opportunities for cross-dataset validation and transfer

learning, opening avenues for fine-tuning pre-existing models on new tasks. Therefore, this dataset not only enhances the accuracy of human action-related tasks, but also provides a novel methodology that can contribute to the field of human action recognition.

In the future, subsequent investigations might center on the exploration of various large-scale multimodal datasets in conjunction with more efficient feature representations to extend and improve multimodal action recognition applications.

Author Contributions: Conceptualization, M.B.S. and D.C.; Methodology, M.B.S.; Software, M.B.S.; Validation, N.A.; Formal Analysis, M.B.S.; Investigation, M.B.S.; Data Curation, M.B.S.; Writing—Original Draft Preparation, M.B.S.; Writing—Review and Editing, S.M.S.I.; Visualization, M.B.S.; Supervision, D.C.; Project Administration, D.C.; Funding Acquisition, M.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Edith Cowan University (ECU), Australia and Higher Education Commission (HEC), Pakistan under GRANT No. PM/HRDI-UESTPs/UETs-I/Phase-1/Batch-VI/2018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: This study used data collected from a web platform that hosts publicly available videos of human actions. To ensure ethical data use, the study adhered to data redistribution policies established by the platform. The research team recognized the importance of respecting data privacy and intellectual property rights when working with publicly available data. Therefore, the study was conducted in compliance with ethical guidelines to protect the rights and interests of both data subjects and data owners.

Data Availability Statement: The data presented in this study are openly available in Mendeley Data at [29,31–35].

Acknowledgments: Naveed Akhtar is a recipient of the Office of National Intelligence National Intelligence Postdoctoral Grant NIPG-2021-001 funded by the Australian Government.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Shaikh, M.B.; Chai, D. RGB-D data-based action recognition: A review. *Sensors* **2021**, *21*, 4246. [CrossRef] [PubMed]
2. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. MAiVAR: Multimodal Audio-Image and Video Action Recognizer. In Proceedings of the International Conference on Visual Communications and Image Processing (VCIP), Suzhou, China, 13–16 December 2022; pp. 1–5. [CrossRef]
3. Sudhakaran, S.; Escalera, S.; Lanz, O. Gate-shift networks for video action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1102–1111. [CrossRef]
4. Yang, G.; Yang, Y.; Lu, Z.; Yang, J.; Liu, D.; Zhou, C.; Fan, Z. STA-TSN: Spatial-Temporal Attention Temporal Segment Network for action recognition in video. *PLoS ONE* **2022**, *17*, e0265115. [CrossRef] [PubMed]
5. Zhang, K.; Li, D.; Huang, J.; Chen, Y. Automated video behavior recognition of pigs using two-stream convolutional networks. *Sensors* **2020**, *20*, 1085. [CrossRef] [PubMed]
6. Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T.L.; Bansal, M.; Liu, J. Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7331–7341. [CrossRef]
7. Girdhar, R.; Ramanan, D.; Gupta, A.; Sivic, J.; Russell, B. ActionVLAD: Learning spatio-temporal aggregation for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 971–980. [CrossRef]
8. Li, Y.; Li, W.; Mahadevan, V.; Vasconcelos, N. VLAD3: Encoding dynamics of deep features for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1951–1960. [CrossRef]
9. Zhou, B.; Andonian, A.; Oliva, A.; Torralba, A. Temporal relational reasoning in videos. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 803–818. [CrossRef]
10. Kwon, H.; Kim, M.; Kwak, S.; Cho, M. Learning Self-Similarity in Space and Time As Generalized Motion for Video Action Recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), Montreal, QC, Canada, 10–17 October 2021; pp. 13065–13075. [CrossRef]
11. Mei, X.; Lee, H.C.; Diao, K.Y.; Huang, M.; Lin, B.; Liu, C.; Xie, Z.; Ma, Y.; Robson, P.M.; Chung, M. Artificial intelligence–enabled rapid diagnosis of patients with COVID-19. *Nat. Med.* **2020**, *26*, 1224–1228. [CrossRef] [PubMed]

12. Gu, J.; Cai, H.; Dong, C.; Ren, J.S.; Timofte, R.; Gong, Y.; Lao, S.; Shi, S.; Wang, J.; Yang, S. NTIRE 2021 challenge on perceptual image quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 677–690. [CrossRef]
13. Yan, C.; Teng, T.; Liu, Y.; Zhang, Y.; Wang, H.; Ji, X. Precise no-reference image quality evaluation based on distortion identification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2021**, *17*, 1–21. [CrossRef]
14. Liu, J.; Wang, X.; Wang, C.; Gao, Y.; Liu, M. Temporal decoupling graph convolutional network for skeleton-based gesture recognition. *IEEE Trans. Multimed.* **2023**, *26*, 811–823. [CrossRef]
15. Giannakopoulos, T.; Pikrakis, A., Eds. Introduction. In *Introduction to Audio Analysis*; Academic Press: Oxford, UK, 2014. [CrossRef]
16. Imtiaz, H.; Mahbub, U.; Schaefer, G.; Zhu, S.Y.; Ahad, M.A.R. Human Action Recognition based on Spectral Domain Features. *Procedia Comput. Sci.* **2015**, *60*, 430–437. [CrossRef]
17. Peeters, G. A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO Ist Proj. Rep.* **2004**, *54*, 1–25.
18. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
19. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. Multimodal Fusion for Audio-Image and Video Action Recognition. *Neural Comput. Appl.* **2024**, 1–14. [CrossRef]
20. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, California USA, 4–9 February 2017; pp. 4278–4284.
21. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. In Proceedings of the Interspeech, Brno, Czech Republic, 30 August–3 September 2021; pp. 571–575. [CrossRef]
22. Chen, T.; Zhai, X.; Ritter, M.; Lucic, M.; Houlsby, N. Self-supervised GANs via auxiliary rotation loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12154–12163. [CrossRef]
23. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *Information* **2020**, *11*, 125. [CrossRef]
24. Takahashi, R.; Matsubara, T.; Uehara, K. Data augmentation using random image cropping and patching for deep CNNs. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 2917–2931. [CrossRef]
25. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 1–48. [CrossRef]
26. Anoosheh, A.; Sattler, T.; Timofte, R.; Pollefeys, M.; Van Gool, L. Night-to-day image translation for retrieval-based localization. In Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 5958–5964. [CrossRef]
27. Alharbi, Y.; Wonka, P. Disentangled image generation through structured noise injection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5134–5142. [CrossRef]
28. Liao, X.; Yu, Y.; Li, B.; Li, Z.; Qin, Z. A new payload partition strategy in color image steganography. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 685–696. [CrossRef]
29. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. Spectral Centroid Images for Multi-Class Human Action Analysis: A Benchmark Dataset. Mendeley Data. 2023. Available online: https://data.mendeley.com/datasets/yfvv3crnpy/1 (accessed on 29 October 2023).
30. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. PyMAiVAR: An open-source Python suite for audio-image representation in human action recognition. *Softw. Impacts* **2023**, *17*, 100544. [CrossRef]
31. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. Chroma-Actions Dataset: Acoustic Images. Mendeley Data. 2023. Available online: https://data.mendeley.com/datasets/r4r4m2vjvh/1 (accessed on 29 October 2023).
32. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. Waveplot-Based Dataset for Multi-Class Human Action Analysis. Mendeley Data. 2023. Available online: https://doi.org/10.17632/3VSZ7V53PN.1 (accessed on 29 October 2023).
33. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. Spectral Rolloff Images for Multi-class Human Action Analysis: A Benchmark Dataset. Mendeley Data. 2023. Available online: https://data.mendeley.com/datasets/3vsz7v53pn/1 (accessed on 29 October 2023).
34. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. MFFCs for Multi-Class Human Action Analysis: A Benchmark Dataset; Mendeley Data, 2023. Available online: https://data.mendeley.com/datasets/6ng2kgvnwk/1 (accessed on 29 October 2023).
35. Shaikh, M.B.; Chai, D.; Islam, S.M.S.; Akhtar, N. MFCCs Feature Scaling Images for Multi-Class Human Action Analysis: A Benchmark Dataset. Mendeley Data. 2023. Available online: https://data.mendeley.com/datasets/6d8v9jmvgm/1 (accessed on 29 October 2023).
36. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (CVPR), Santiago, Chile, 7–13 December 2015; pp. 4489–4497. [CrossRef]
37. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 20–36. [CrossRef]

38. Takahashi, N.; Gygli, M.; Van Gool, L. AENet: Learning deep audio features for video analysis. *IEEE Trans. Multimed.* **2017**, *20*, 513–524. [CrossRef]
39. Tian, Y.; Shi, J.; Li, B.; Duan, Z.; Xu, C. Audio-visual event localization in unconstrained videos. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 247–263. [CrossRef]
40. Brousmiche, M.; Rouat, J.; Dupont, S. Multimodal Attentive Fusion Network for audio-visual event recognition. *Inf. Fusion* **2022**, *85*, 52–59. [CrossRef]
41. Long, X.; De Melo, G.; He, D.; Li, F.; Chi, Z.; Wen, S.; Gan, C. Purely attention based local feature integration for video classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2140–2154. [CrossRef] [PubMed]
42. Gao, R.; Oh, T.H.; Grauman, K.; Torresani, L. Listen to Look: Action Recognition by Previewing Audio. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10457–10467. [CrossRef]
43. Shaikh, M.B.; Chai, D.; Shamsul Islam, S.M.; Akhtar, N. MAiVAR-T: Multimodal Audio-image and Video Action Recognizer using Transformers. In Proceedings of the 2023 11th European Workshop on Visual Information Processing (EUVIP), Gjovik, Norway, 11–14 September 2023; pp. 1–6. [CrossRef]