Neural Computing and Applications (2024) 36:5499–5513
https://doi.org/10.1007/s00521-023-09186-5

ORIGINAL ARTICLE

Multimodal fusion for audio-image and video action recognition

Muhammad Bilal Shaikh[1,4] · **Douglas Chai**[1] · **Syed Mohammed Shamsul Islam**[2,4] · **Naveed Akhtar**[3]

Received: 13 April 2023 / Accepted: 20 October 2023 / Published online: 9 January 2024

Abstract

Multimodal Human Action Recognition (MHAR) is an important research topic in computer vision and event recognition fields. In this work, we address the problem of MHAR by developing a novel audio-image and video fusion-based deep learning framework that we call Multimodal Audio-Image and Video Action Recognizer (MAiVAR). We extract temporal information using image representations of audio signals and spatial information from video modality with the help of Convolutional Neutral Networks (CNN)-based feature extractors and fuse these features to recognize respective action classes. We apply a high-level weights assignment algorithm for improving audio-visual interaction and convergence. This proposed fusion-based framework utilizes the influence of audio and video feature maps and uses them to classify an action. Compared with state-of-the-art audio-visual MHAR techniques, the proposed approach features a simpler yet more accurate and more generalizable architecture, one that performs better with different audio-image representations. The system achieves an accuracy 87.9% and 79.0% on UCF51 and Kinetics Sounds datasets, respectively. All code and models for this paper will be available at https://tinyurl.com/4ps2ux6n.

Keywords Deep learning · Human action recognition · Multimodal · Information fusion · CNN

1 Introduction

Recent advances in the deep learning architectures, improvements in communication mechanisms of Graphics Processing Unit (GPU) hardware, as well as software stacks have provided a boost in supporting computationally complex tasks such as Multimodal Human Action Recognition (MHAR). Understanding activities in a multimodal information setting are a complex and resource hungry task, one which has become an important research problem in computer vision. Audio and video-based action recognition has many potential applications in: visual acoustics-based platforms; as described by Gao and Grauman [9]; sports analytics as noted by Vinyes Mora and Knottenbelt [64]; smart social surveillance for COVID-19 as detailed by citehuu2022action; self-driving vehicles as identified by Kala [23]; and content-based search systems as examined by Gibbon and Liu [13].

An action could be defined as: “Action is the most elementary human -surrounding interaction with a meaning” [19, p.2]. Action recognition typically aims to discover a class of short, segmented, atomic actions. Due to their multifaceted nature, some of these approaches refer to action recognition as plan recognition, goal recognition, intent recognition, behavior recognition, location estimation, event recognition and interaction recognition, as evident in Shaikh and Chai [51], Vandersmissen et al [63], Slade et al [54] and Jing et al [21], respectively. Human Action Recognition (HAR) is the process of labeling

✉ Muhammad Bilal Shaikh
mbshaikh@our.ecu.edu.au

Douglas Chai
d.chai@ecu.edu.au

Syed Mohammed Shamsul Islam
syed.islam@ecu.edu.au

Naveed Akhtar
naveed.akhtar1@unimelb.edu.au

[1] School of Engineering, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Australia

[2] School of Science, Edith Cowan University, 270 Joondalup Drive, Joondalup, WA 6027, Australia

[3] School of Computing and Information Systems, The University of Melbourne, Melbourne Connect, 700 Swanston Street, Carlton, WA 3053, Australia

[4] Center for Artificial Intelligence and Machine Learning (CAIML), School of Science, Edith Cowan University, Perth, Australia

actions performed by humans within a given sequence of images, where it becomes the classification of goals of human agents in a series of image frames.

It is observable that many objects in our daily life operate concurrently to give meaning to an activity. This interaction relies on quantifiable, yet multimodal signals. Accordingly, joint learning of these multimodal quantities may result in better feature representation for any downstream analysis. Multimodal understanding is a natural human ability. In the context of MHAR, the goal is to improve recognition performance by collecting critical features from distinct modalities. These candidate features are then combined using an optimal aggregation strategy, which contributes to the overall learning of an action category. From an early age, humans become used to applying different senses to acquire visual as well as acoustics information from their surroundings in order to understand an event. However, recent approaches have generally ignored the contribution of audio data in enhancing action recognition. Further, multimodal data fusion has benefits that can be applied across different fields Boehm et al [3].

Audio signals have significant properties that help with efficient recognition in videos, where audio contains dynamics and rich contextual temporal information Gaver [12]. Most importantly, audio has a much lighter computational complexity as compared to video frames. Across an entire video, audio can also be beneficial for selecting critical features that are useful for recognition. For example, the sound of a person talking at the start of a video can suggest that the actual action has not yet started, while the sound of an electric saw may indicate that the action is taking place. Figure 1 shows a scenario where audio features combined with video features yield a significant improvement in recognizing classes where audio features are the discriminating factor. In Fig. 1a, the sample instance was incorrectly classified as a "Boxing Speed Bag" with video-only features. However, with the fusion of audio-image and video features, the same instance was correctly recognized as "Boxing Punching Bag" due to distinct audio features (shown in Fig. 1b).

Video modality contains spatial information, which is inherently helpful for CNN-based classification architectures. To better capture multimodality aspects of action data, a recent trend has been to combine information from different modalities, such as optical flow, RGB-difference and warped-optical flow Wang et al [66]. For example, as shown in Fig. 1, within a short clip of action of 'Boxing a Punching Bag', a single audio-image frame contains most of the dynamic contextual information contained in the audio, (i.e., the sound of a boxing glove hitting the bag), while the accompanying video clip contains useful cues of spatial dynamics.

Advanced computer vision technologies and artificial intelligence algorithms have been proposed to recognize human actions. Accordingly, Convolutional Neural Networks (CNNs) have been successful in image classification Paoletti et al [45], Seo and Shin [50], Wan et al [65], Sharma et al [53], digital recognition Baldominos et al [2], Tao et al [60], Kulkarni and Rajendran [27], object detection Jung et al [22], Deng et al [6] and many information retrieval domains. In action recognition research, CNN studies have predominantly used visual data, which is rich in spatial information. However, most of these works have not exploited the temporal and contextual information that lies in accompanying audio.

In this paper, we pose and seek to answer the following questions: (Q1) How well can CNN encode audio signals using image-based representations? (Q2) How does knowing information from one modality influences the model in learning the action efficiently? (Q3) How do we

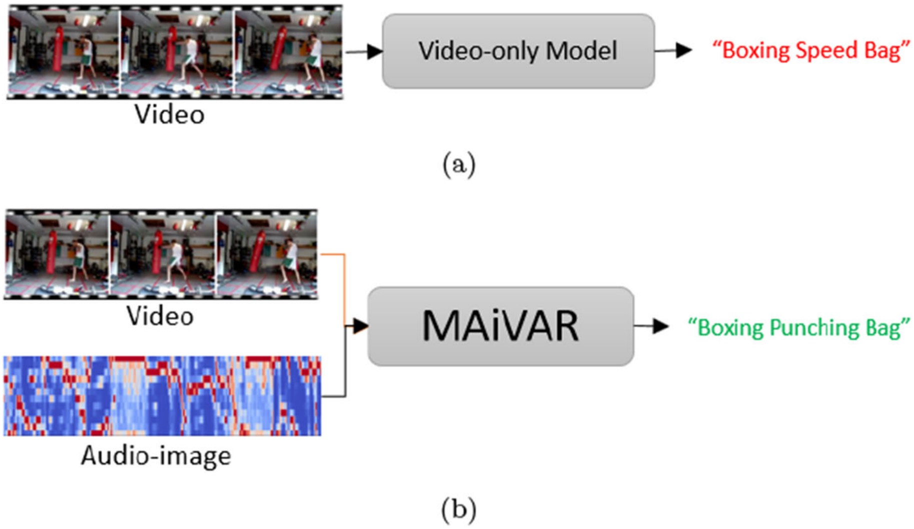


Fig. 1 Illustration of an example **a** miss-classified unimodal (Video-only) compared to **b** correctly classified example with multimodal data (Audio-image+Video) using proposed MAiVAR framework

fuse information found in both modalities so that it improves model performance? While related questions may been studied in the literature, to the best of our knowledge, no study has been conducted to answer these collective questions as a whole. Hence, we propose the Multimodal Audio-Image and Video Action Recognizer (MAiVAR) framework, a CNN-based audio-image to video fusion-based technique that can be applied to video and audio features for classifying actions. Additionally, we propose an approach to extract meaningful image representations of audio, which can significantly improve classification scores when used with CNN-based models.

The benefits of the proposed MAiVAR are threefold. Firstly, MAiVAR has shown superior performance to outperform state-of-the-art models on the same data configuration for video representation when evaluated on UCF51 Takahashi et al [58] and Kinetics Sounds Arandjelovic and Zisserman [1] datasets. Secondly, MAiVAR naturally supports both audio and video inputs and can be applied to different tasks without any change of architecture. In particular, the models we used have different architectures for video and audio feature extraction. In contrast, existing video-only models typically require variants of RGB modality to obtain optimal performance, such as optical flow, warped-optical flow and RGB difference. Thirdly, compared with state-of-the-art video-based CNN models, experimental results show that MAiVAR features converge faster during training. To the best of our knowledge, MAiVAR is the first audio-image to video fusion-based action classification framework that uses image-based representations of audio to leverage CNN architecture for better action recognition. The key contributions of our work can be summarized as follows:

- We introduce MAiVAR, and a new multi-staged multimodal framework that supports novel dominant head audio-visual fusion. Our fusion approach eliminates expensive fusion operations, significantly reducing the computational complexity of the model. Our framework can be applied to different tasks with minimal changes to the overall architecture.
- We build a new feature representation strategy to select the most informative candidate representations for audio-visual fusion.
- Unlike previous methods, we propose a high-level weights assignment algorithm for better audio-visual interaction and convergence.
- We achieve state-of-the-art or competitive results on standard public benchmarks, validating the generalizability of our proposed approach through an extensive evaluation.

This paper is a significant extension of our previous work Shaikh et al [52] in a number of aspects. We extend the original MAiVAR framework to validate scalability and generalization by testing on a larger dataset (i.e., Kinetics-Sounds). We also provide more detailed discussion on technical aspects and include more comprehensive review of the related literature to better contextualize our contribution.

The remainder of the paper is structured as follows. Section 2 describes how our work differs from other closely related works on MHAR using audio and visual modalities. Section 3 describes our proposed methodology. Section 4 discusses the dataset and the training environment of the system. Section 5 provides an analysis and comparison of the results obtained using the proposed approach, and Sect. 6 presents the conclusion of this paper.

2 Related work

Much research has been conducted in the field of MHAR Takahashi et al. [58], Long et al. [40], Tian et al. [61], Brousmiche et al. [5], Li et al. [30–32, 35], Li and Tang [34], where methods have been designed to combine information from distinct modalities. This section uncovers some existing works related to our approach, examining them in terms of feature extraction and multimodal action recognition approaches.

2.1 Feature extraction

Feature extraction is a process for 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 has been rigorously used as an efficient video feature extractor for different problems. 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 Sudhakaran et al. [56], an accuracy improvement of 32% is achieved. Furthermore, Yang et al. [68] have used TSN with soft attention mechanism to capture important frames from each segment. Moreover, Zhang et al. [69] 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 Lei et al. [29], Girdhar et al. [14], Li et al. [33], Zhou et al. [70], Kwon et al. [28] and is typically used in conjunction with a succeeding module. In Kwon et al. [28], TSN was employed as a 2D CNN backbone to learn motion dynamics in videos. However, IRV2 has been used for feature extraction from images Mei et al. [43], helping with different image restoration and enhancement tasks Gu et al. [16], Yan et al. [67].

MAFnet Brousmiche et al. [5] are a multi-level attention-based fusion network that uses a lateral connection in the form of a Feature-wise Linear Modulation (FiLM) layer to incorporate modality conditioning among audio and video clip streams. MAFnet uses both intermediate feature fusion and late fusion of features to project the effect of the combined feature, which adds an overhead to the overall workflow. Spatial-Temporal Network (StNet) He et al. [17] adapt IRV2 to model local and global spatio-temporal features. The closest works to ours are Takahashi et al. [58] and Wang et al. [66]. Takahashi et al. [58] have classified audio events using 3D CNN with some representations of audio action data, while we have proposed a different audio representation strategy. Similarly, the temporal segment-based approach used by Wang et al. [66] applies optical flow modality variants with RGB, while our work uses RGB-based features using TSN as the feature extractor for the visual stream.

2.2 Multimodal action recognition

Multimodal action recognition employs multi-stream approaches to incorporate different modalities. Motivated by successes in image classification, CNNs have also been applied in different visual action recognition works, where several approaches have been designed to gain the benefits of appearance information. TSN Wang et al. [66], TRN Zhou et al. [70] and TSM Lin et al. [37] all are based on 2D CNNs. All three models employ a two-stream approach that uses both RGB and optical flow. Besides RGB and optical flow streams, Temporal Binding Networks (TBN) Kazakos et al. [25] adds audio as an additional modality as well. SlowFast Feichtenhofer et al. [8] uses two RGB streams with different resolutions and frame rates.

IMGAUD2VID Gao et al. [10] introduces a video skimming mechanism for untrimmed videos, aided by audio, to eliminate both short-term and long-term redundancies. Accordingly, it uses audio to extract dynamic scene information along with a single frame that captures most of the appearance information, in order to form an image-audio pair. These pairs are then used to select key moments from the video for action recognition. Unlike IMGAUD2VID, our idea captures more spatial information along with the holistic dynamic information of the scene from the image representations of audio.

An optimal strategy to fuse features from different modalities is critical to take maximum advantage of each modality. Existing multimodal action recognition approaches have used fusion at early Feichtenhofer et al. [7], middle Roitberg et al. [48] and late Patel et al. [47] stages of neural networks. Different levels of fusion in the network have been tested by Long et al. [39]. However, most of these approaches have used only visual modalities. In Long et al. [38], concatenation was used to fuse the visual and audio modalities. Complex fusion techniques, such as multimodal compact bilinear pooling (MCB) Gao et al. [11] or dual multimodal residual fusion (DMR) Tian et al [61] have also been studied.

In recent years, the surge in mobile device usage has underscored the need for robust security measures, particularly given the integral role smartphones play in our lives and the looming threats to user privacy. To address this, Li et al. [32] has introduced ADFFDA, an innovative mobile continuous authentication system that integrates an Adaptive Deep Feature Fusion scheme and a transformer-based GAN for Data Augmentation, employing common smartphone sensors like the accelerometer, gyroscope, and magnetometer, achieving a remarkably low mean equal error rate of 0.01%. Similarly, DeFFusion, another system by Li et al. [31], harnessed the same sensors but focused on CNN-based continuous authentication by transforming time domain data into the frequency domain, with an error rate of 1.00 % in a 5-second window. Furthermore, FusionAuth exploited these sensors to capture user behavior, employing two novel feature fusion strategies and achieving error rates as low as 1.47 % Li et al. [30]. Beside authentication, the massive influx of community-contributed images has led to advancements in image understanding. For instance, the Deep Collaborative Embedding (DCE) model, as proposed by citeli2018deep, seeks to understand these images by unifying the latent space for images and tags. Concurrently, the weakly supervised deep metric learning (WDML) algorithm leverages both visual content and user tags for more efficient image retrieval, addressing challenges like noisy or subjective tags Li and Tang [34]. These diverse yet interrelated studies collectively highlight the evolving landscape of device security and image understanding, stressing the importance of leveraging deep learning techniques and sensor data for better results.

2.3 Chromagram representations

Chroma features have been widely used in action recognition because they are effective not only in capturing the musical and rhythmic structure of action, but also in the spectral information of an audio signal compactly and efficiently. The main justification for using chroma features for action recognition lies in their ability to represent the pitch content of an audio signal in a musically meaningful way. Chroma features are derived from the short-time Fourier transform (STFT) of an audio signal, where each frame of the STFT is mapped onto a 12-dimensional pitch class profile. This pitch class profile summarizes the energy of each pitch class in the frame, thereby providing a concise representation of the harmonic content of the signal.

For example, in a dance performance, the rhythm of the music and the dancer's movements is highly correlated. Chroma features can capture this correlation by highlighting the prominent pitch classes in the music and the corresponding rhythmic patterns in the dancer's movements. Similarly, in sports action recognition, the sound of the ball being hit or kicked can be captured by the chroma features, which can then be used to identify the type of sport being played. Another advantage of chroma features is their robustness to noise and variations in audio signal. Since chroma features only capture the pitch content of an audio signal, they are less sensitive to changes in the timbre or dynamics of a signal. This makes them particularly useful for recognizing actions in noisy or challenging environments where other features may be more prone to error.

Gouyon et al. [15] have shown that rhythmic descriptors based on chroma features are effective for musical genre classification. Moreover, chroma features are among the most effective audio features for human action recognition, particularly for recognizing actions that are accompanied by music. In addition, chroma features are robust to variations in the audio signal. Besides this, Lidy and Rauber [36] have evaluated feature sets and fusion strategies for genre classification of music, finding that chroma features are relatively robust to variations in the timbre and dynamics of the signal. Similarly, citeravanelli2018learning have used chroma features to learn speaker representations for speaker diarization and found that they are robust to noise and reverberation. Overall, the selection of chroma features for action recognition is justified by their ability to capture the musical and rhythmic structure of an action robustly and effectively. These characteristics make them a popular choice for many researchers and practitioners in the field.

3 Proposed methodology

In this section, we describe our overall proposed framework, which fuses multimodal (audio-image and video) data for action recognition. We start by providing a brief overview of our proposed approach. We then discuss individual components of the framework, which consist of video and audio streams. Furthermore, we discuss network architectures of visual and audio feature extractors. This is followed by a brief discussion about the multimodal fusion network, which fuses the features from individual streams.

3.1 Overview

Figure 2 presents the schematics of the MAiVAR framework. We split each video V into k number of segments s of equal duration so that an ith video sample could be defined as

$$V_i = \{s_1, s_2, \ldots, s_k\}. \tag{1}$$

A short snippet is then randomly selected from each segment. We then extract visual feature maps $(\{v_1, \ldots, v_T\}, v_t \in \mathbb{R}^{D_{video}})$ and audio feature maps $(\{a_1, \ldots, a_T\}, a_t \in \mathbb{R}^{D_{audio}})$ with pretrained modality specific CNNs. Features maps are then reduced using average pooling, which extracts $t = 1, \ldots, T$ temporal features across $k = 1, \ldots, K$ modalities. Accordingly, we obtain a richer multimodal representation of the entire video. These features are then passed through a Multimodal Fusion Network (MFN), which outputs the learned classifications. The output of the framework is $y \in \mathbb{R}^N$ with N being the number of classes. An example scheme of feature fusion is shown in Fig. 3. To go beyond simple fusion, we initialize the fusion network with weights from a trained Video Multi-Layer Perceptron (VMLP) model. The weights from VMLP are then directly injected into the fusion model, which helps the fusion model perform quicker convergence.

3.2 Visual stream

The duration of video input is t seconds, which is then converted into a frame-based representation that is compatible with TSN. This is then fed as an input to the TSN feature extractor. Each video clip thus represents an action example. We extract the feature from TSN using the *AvgPool2d* layer while pruning the original average consensus layer. TSN produces an embedding of size 1024×25 to allow the model to capture the spatio-temporal structure of the RGB video. The resulting sequence is then fed as input to the fusion model. The advantages of this setup are: 1) the standard TSN architecture is easy to implement and reproducible as it is off-the-shelf in TensorFlow and PyTorch, and 2) we intend to apply transfer learning for Fusion MLP, whereby a standard architecture makes transfer learning easier. We have adapted the MMAction2[1] interface for using TSN as a feature extractor.

3.3 Audio stream

For audio with a duration of t seconds, we convert it into an image-based representation. This results in a 299×299 image-based representation, referred to as an audio-image of the whole action data sample. This audio-image is then fed as an input to the IRV2 backbone. We then extract the audio features from IRV2 using the *AvgPool2D* layer while pruning the original classification layer. IRV2 produces an embedding of size 1536 to allow the model to capture the

[1] https://mmaction2.readthedocs.io/en/latest/.

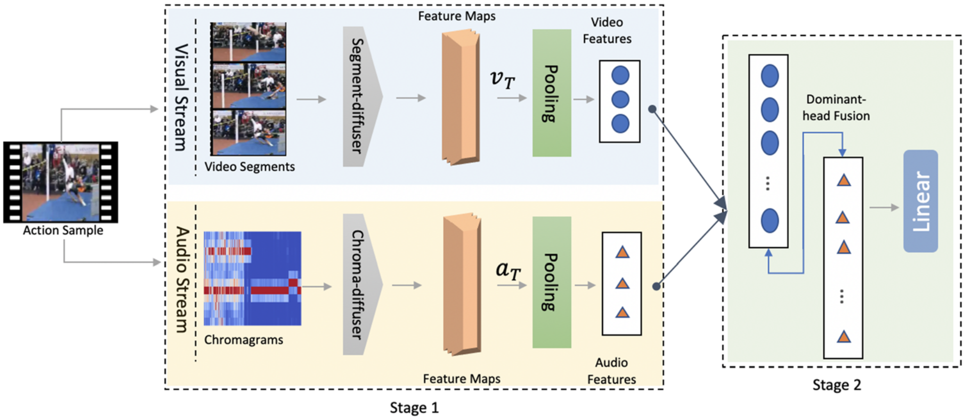


Fig. 2 Overview of our proposed multi-staged architecture. In stage 1, along the audio stream, the 2D audio is transformed into a grid-compatible representation, which is subsequently resized to 224 × 224 for feature extraction through chroma diffuser and then, linearly projected onto the stage 2 for Multimodal Fusion. Along the visual stream, the visual input is passed through a segment diffuser and then, linearly projected onto the stage 2 for fusion. The output of the stage 2 is used for classification with a linear layer. For the fusion network, weights from Video MLP were used to initialize the stage 2

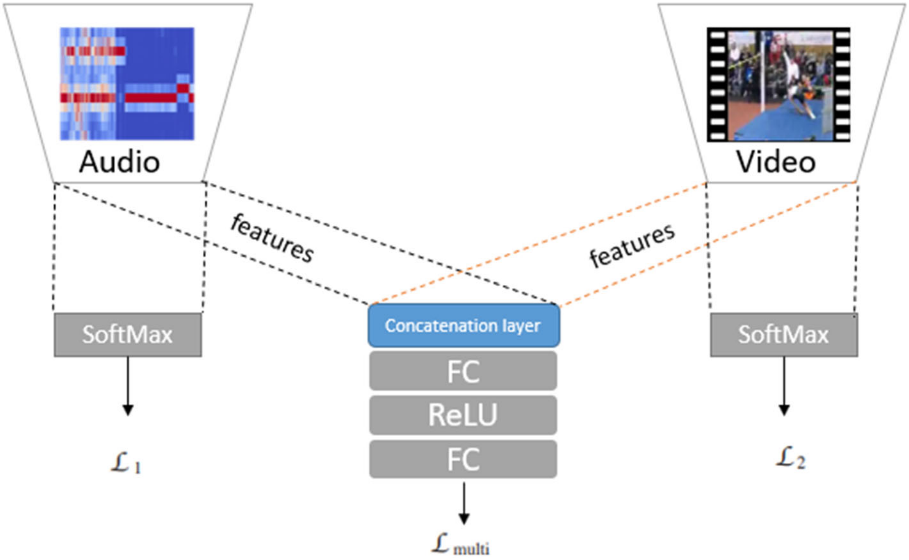


Fig. 3 Scheme showing the multimodal model architecture. The audio-DNN and video-DNN models output independent action prediction $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively, based on their respective input datasets. The features from the *average pooling* layers of the audio-DNN and video-DNN are combined to feed the fusion module, which outputs the multimodal action prediction $\mathcal{L}_{multi}$

spectral audio information through image-based representation. The resulting sequence is then fed as input to the fusion model. The advantage of this simple setup is that the standard IRV2 architecture performs better on audio-image representations as compared to several other CNN-based models (discussed in Sect. 5.2), whereby it is comparatively easy to reproduce, as it is off-the-shelf and available with PyTorch.[2]

3.4 Segment diffuser

A segment diffuser uses a TSN-based Wang et al [66] feature extractor, pretrained with Kinetics-400 for projecting our visual embeddings. In TSN, the original BNInception backbone is used. The consensus layer is removed to expose features from the *AvgPool2D* layer of the TSN model. The features from TSN are then fed as input to the visual multilayer perceptron, which classifies visual features into class labels. The weights from the trained VMLP are then used to initialize the fusion network.

3.5 Chroma diffuser

A chroma diffuser uses an IRV2-based feature extractor initialized with weights from ImageNet for creating audio embeddings for the audio MLP. In IRV2, all gradients were kept frozen and the last layer was removed to expose features from the average pooling layers of the model. The

[2] https://pytorch.org/docs/stable/index.html.

features from IRV2 were then fed as input to the fusion module, which fuses it with visual features.

3.6 Dominant head fusion

Dominant Head Fusion (DHF) processes unimodal feature representations separately and then, learns a joint representation using a middle layer. DHF concatenates the visual and audio feature vectors and classifies the combined feature dimension into action labels. The VMLP model produces higher classification accuracy. Therefore, the weights from the trained visual network were used to initialize items and to fine-tune the DHF process.

All hidden layers, except the last fully connected layer, are equipped with Rectified Linear Unit (ReLU) nonlinearity. The fusion network was trained by minimizing the cross-entropy loss L with l_1 regularization using back-propagation:

$$\theta^* = \arg\min_{W} \sum_{i,j} L(x_j^i + z_j^i, y_j^i, W) \tag{2}$$

where x_j^i and z_j^i are the jth input vector from audio and video features, respectively, y_j is the corresponding class label, and W is the set of network parameters, respectively. For the audio-video fusion model, we used initialization weights from the video classification model (see Algorithm 1). Empirically, the weights from the video model boost the convergence speed of the fusion model.

4 Experimental setup

4.1 Datasets

We evaluated the MAiVAR on two popular action recognition datasets: UCF51 Takahashi et al [58] and Kinetics Sounds Arandjelovic and Zisserman [1].

UCF51 is a subset of standard benchmark dataset UCF101 Soomro et al [55] for action recognition. UCF101 consists of 13, 320 videos of 101 human action categories, such as Apply Eye Makeup, Blow Dry Hair and Table Tennis. However, 6,836 videos from 51 action classes have audio channels. The average video length is 7.0 s. This dataset was partitioned into three splits for training and testing. Number of samples per class is shown in Fig. 5.

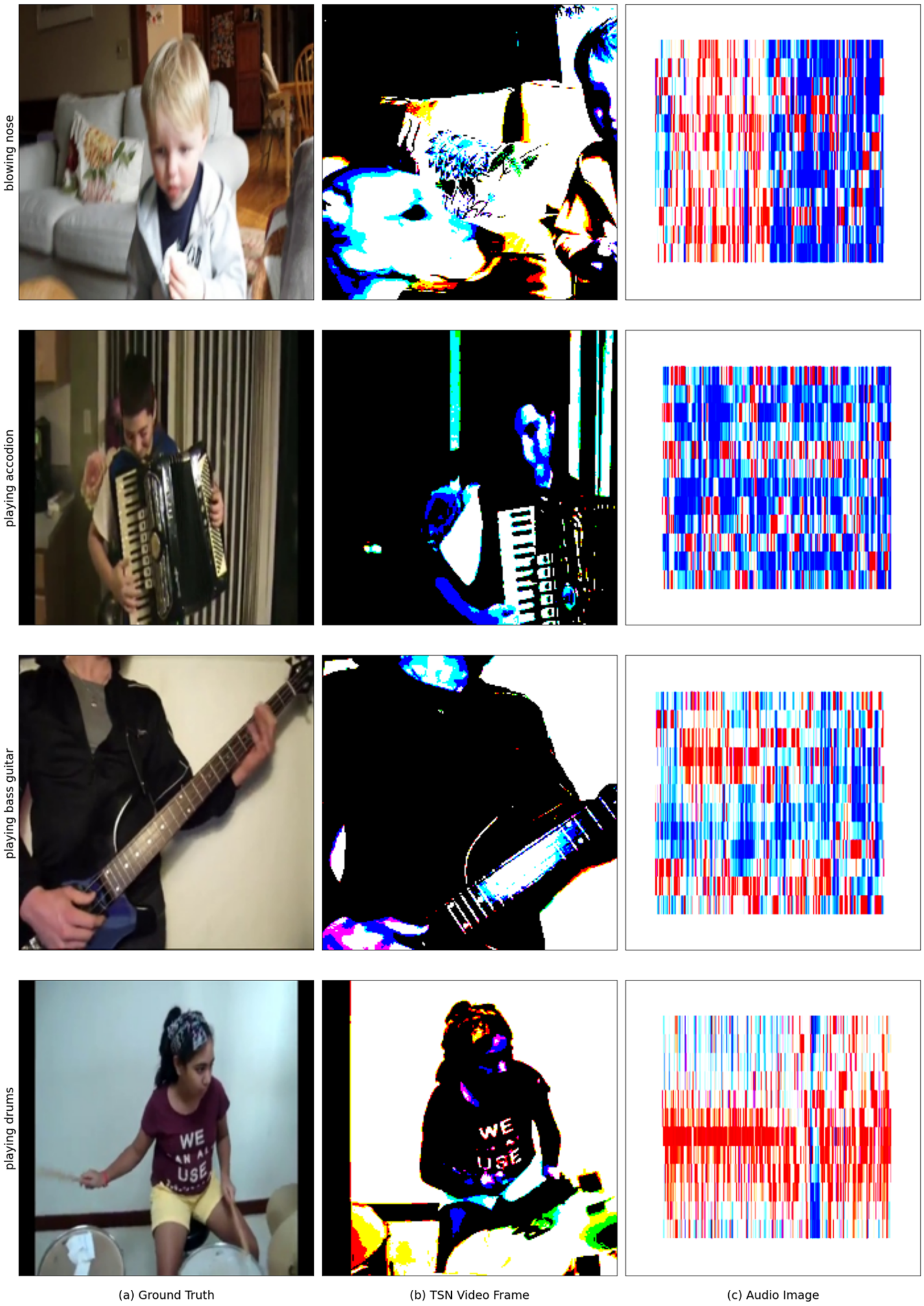
Kinetics Sounds is a subset of Kinetics Human Action Video dataset (referred as Kinetics-400) Kay et al [24]. The original version of Kinetics-400 dataset consists of 306, 245 videos divided into 400 action categories with $1 - 150$ clips for each action. However, Kinetics Sounds consists of only action classes that are both visually and aurally recognizable. The average video length is 9.7 seconds. The subset possesses more than 22, 000 videos from 27 different action classes.

Algorithm 1 Weights assignment algorithm

```
Require: Initial weights for $\alpha^0$ audio, and $\nu^0$ video, $N$ Number of epochs
Ensure: $\omega^0$ (weights after training)

procedure WEIGHTS($\alpha^0, \nu^0, N$)
    Train $\alpha^0$ with $\{Wa_i\}_{i=1}^{k}$
    for $i = 1, 2, ..., N$ do
        process batch
        update weights $\alpha_i$
    end for
    Train $\nu^0$ with $\{Wv_i\}_{i=1}^{k}$
    for $i = 1, 2, ..., N$ do
        process batch
        update weights $\nu_i$
    end for
    $\omega^0 \leftarrow \nu^N$;
    Fusion model weights $\omega^0$ initialized with weights from trained video model $\nu^N$.
    return $\omega^0$ (weights for the fusion module)
end procedure
```

Fig. 4 Audio and video representations at different stages of the framework of data samples



4.2 Training

Input Pre-processing: We consider two modalities: RGB and audio. For RGB, we used video frames as input that are grouped into 25 frames for each segment. We followed Wang et al [66] for visual pre-processing and augmentation. For audio, we used Librosa library McFee et al [42] to generate six distinct image-based representations of audio samples.

Data Augmentation: Audio-image representations were normalized as per ImageNet configuration, with random horizontal and vertical flips. We have followed Wang et al [66] for visual data transformations.

Feature Extraction: The TSN-based Wang et al [66] feature extractor was used with BNInception Ioffe and Szegedy [20] as backbone for visual data and IRV2 Szegedy et al [57], which uses residual inception blocks for audio feature extraction.

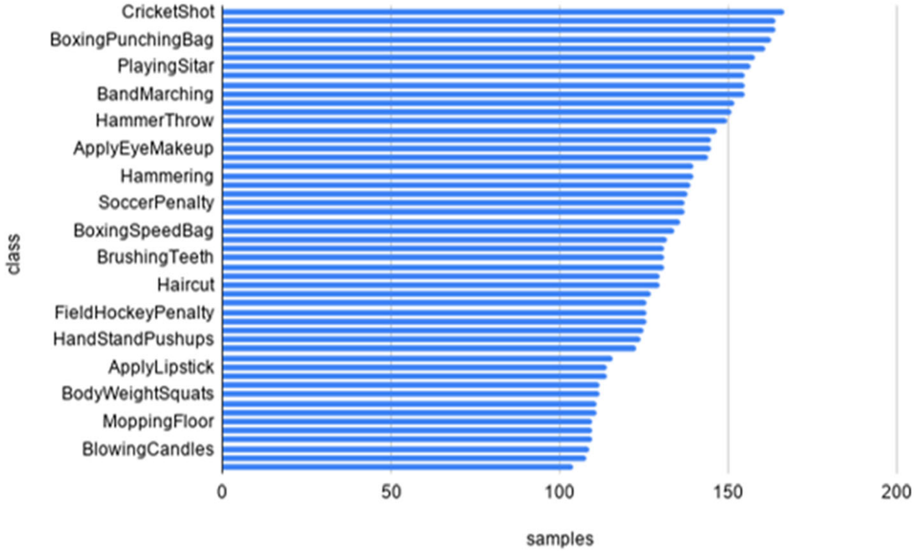


Fig. 5 A snapshot from UCF51 showing number of samples per action class with audio-channel

Table 1 Classification accuracy of MAiVAR compared to the state-of-the-art methods on UCF51 dataset

Year	Method	Accuracy [%]
2015	C3D Tran et al [62]	82.23
2016	TSN (RGB) Wang et al [66]	60.77
2017	C3D+AENet Takahashi et al [58]	85.33
2018	DMRNTian et al[61]	81.04
2018	DMRNTian et al[61] +Brousmiche et al [5] features	82.93
2020	Attention Cluster Long et al [40]	84.79
2020	IMGAUD2VID Gao et al [10]	81.10
2022	STA-TSN (RGB)Yang et al [68]	82.1
2022	MAFnet Brousmiche et al [5]	86.72
2023	MAiVAR-WP	86.21
2023	**MAiVAR-SC**	**86.26**
2023	MAiVAR-SR	86.00
2023	MAiVAR-MFCC	83.95
2023	MAiVAR-MFS	86.11
2023	**MAiVAR-CH**	**87.91**

Environment: The training environment consists of an NVIDIA GeForce GTX 1080 Ti GPU accelerator 12GB memory, Intel(R) Xeon(R) CPU E5-2650 v4 CPU.

Implementation: An IRV2 Szegedy et al [57] model pre-trained on the ImageNet Russakovsky et al [49] dataset was used to extract candidate feature representations from image representations of audio. The models were implemented using the PyTorch library Paszke et al [46]. Similar to Takahashi et al [58], we have used split 1 of standard train-test split provided by UCF101 Soomro et al [55].

Hyperparameters: The number of neurons in the hidden layer for MFN block was 1024. An optimal batch size of 768 for MFN and 512 for VMLP was used. The networks were trained using cross-entropy loss and Adam optimizer Kingma and Ba [26] with a learning rate of 3×10^{-5} for MFN and 10^{-4} for VMLP. We shifted learned weights from the video model to the fusion model for better weights initialization.

5 Results and discussion

Tables 1 and 2 compare recognition performance of the MAiVAR framework versus previous state-of-the-art methods on two datasets. Firstly, we compared the performance of the framework with baselines on both UCF51 and Kinetics Sounds datasets. Secondly, we demonstrate the considerations that evolved the design of our

Table 2 Classification accuracy of MAiVAR compared to the state-of-the-art methods on Kinetics Sounds dataset

Year	Method	Accuracy [%]
2017	Supervised direct Arandjelovic and Zisserman [1]	65
2017	Supervised pretraining Arandjelovic and Zisserman [1]	74
2017	L^3-Net Arandjelovic and Zisserman [1]	74
2020	Attention Cluster Long et al [40]	73.91
2018	DMRNTian et al [61]	77.5
2023	**MAiVAR-CH (Ours)**	**79.01**

framework. We present a comparison of different CNN-based feature extractors followed by study of the efficiency of different audio-image representations. Finally, we have analyzed the impact of multimodal fusion as compared to the unimodal alternatives. The benchmarks were reproduced using accuracy over the standard train and test split. Following the evaluation protocol of Takahashi et al [58], we used accuracy metrics to evaluate the performance of the models that could be calculated as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{I} TP_i}{\sum_{i=1}^{I} TP_i + \sum_{i=1}^{I} FP_i} \tag{3}$$

where respective TP_i and FP_i indicate the number of correct and wrong predictions in the ith class. Accordingly, I is the number of classes.

5.1 Comparisons with the state-of-the-art methods

The validation sample from the datasets action categories evaluate the system performance. Table 1 shows the MAiVAR performance compared to other methods that use audio-visual modalities including AENet Takahashi et al [58], C3D Tran et al [62], IMGAUD2VID Gao et al [10] and other TSN-based techniques Wang et al [66]. The UCF51 dataset is comprised of fewer classes with relevant audio information. Therefore, MAiVAR was able to exploit the significant features that lies in audio signals among all the architectures that use audio-visual features.

To demonstrate the generalization of MAiVAR, we conducted experiments on a larger dataset, Kinetics-Sounds, which is also focused toward human actions in daily life and has potentially more recognizable samples with both acoustic and visual information. Based in Table 2, our proposed approach yielded competitive results compared to other methods. MAiVAR is better capable of

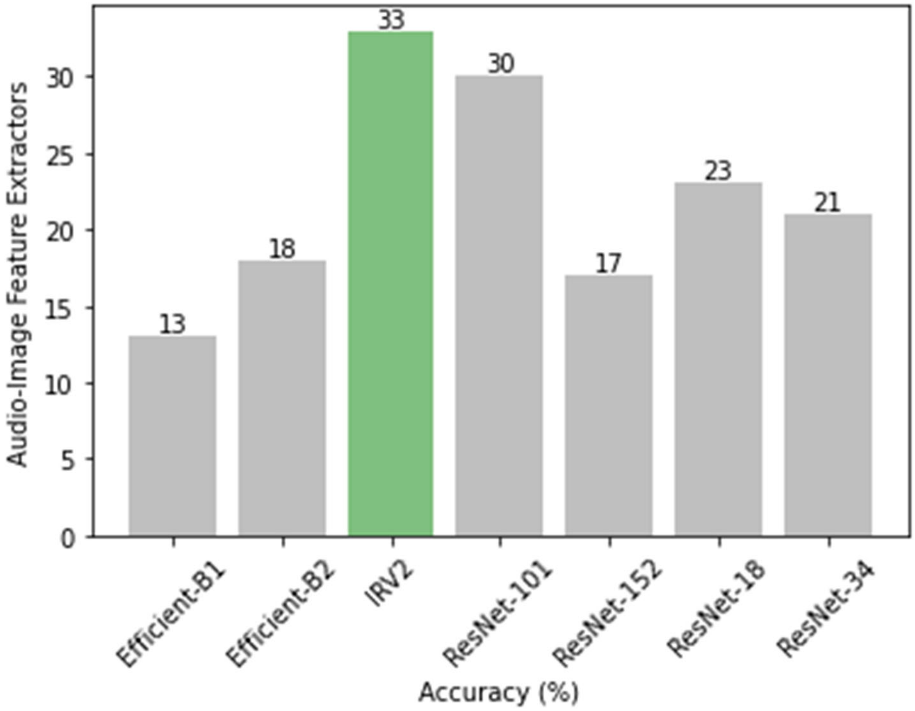


Fig. 6 Performance analysis of audio-image feature extractors on the UCF51 dataset

getting benefits from both modalities. Moreover, the proposed framework demonstrated better mixing of audio and visual information 3.6 than other baseline models.

5.2 Ablation study

5.2.1 Audio-image feature extraction

Several experiments were conducted to determine the optimal design configuration for the proposed framework, which can be broken down into the following points: (1) different feature extractors for audio-image representations, (2) optimal audio-image representation, and (3) analyzing the influence of hyper-parameter settings. We tried several CNN-based feature extractors for audio-image representations including: the smaller as well as wider ResNet He et al [18], EfficientNet Tan and Le [59] and Inception ResNet Szegedy et al [57]. Compared to IRV2, ResNet101 and ResNet18 showed relatively better performance (see Fig. 6).

5.2.2 Convergence of the audio representations

We also evaluated the performance of six different audio-image representations for fusion with video features including: (i) Wave plot, (ii) MFCC, (iii) MFCC feature scaling, (iv) Spectral Centroids, (v) Spectral Rolloff and (vi) Chromogram. A sample example of each audio-image representation is shown in Fig. 9, along with a visual representation of same data sample. As the structure of the chromagram is well-suited for CNN-based models, chromagram-based representations performed better than other competitor representations in the fusion process. Convergence of each audio-image representation after fusion with video modality is illustrated in Fig. 7. However, it can also be observed that a unimodal representation learning with higher accuracy is not the optimal candidate for achieving better performance after multimodal fusion. As presented in Table 3, it is evident that MFCCs Feature Scaling-based representation produced the best audio-level accuracy, after which fusion with chromagram-based representation produced optimal features for the fusion.

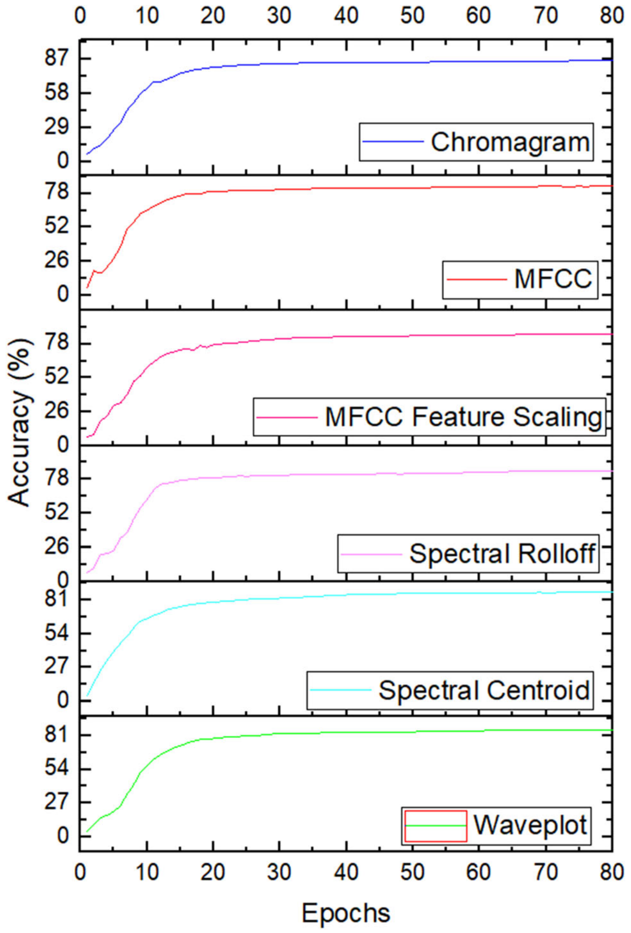


Fig. 7 Convergence of audio representations on the UCF51 dataset

Table 3 Validation accuracy of different audio representations before and after fusion on UCF51

Representation	Modalities	
	Audio	Fusion
Waveplot	12.08	86.21
Spectral Centroids	13.22	86.26
Spectral Rolloff	16.46	86.00
MFCCs	12.96	83.95
MFCCs Feature Scaling	17.43	86.11
Chromagram	15.48	**87.91**

5.2.3 Multimodal fusion analysis

In this section, we analyze the impact of feature fusion. It was observed that fusing extracted features from audio-image and video provides better results than without fusion. However, empirically it has been shown that beyond simple concatenation, weights from one modality can influence fusion with the other modality. The best results are obtained when visual features and weights from the VMLP model interact with audio-image features. We learned that vision features and weights contribute more than acoustic signals in this regard. This is aligned with previous observations in Brousmiche et al [4, 5].

Figure 8 compares the embedding of the different features extracted at different phases of the framework (shown

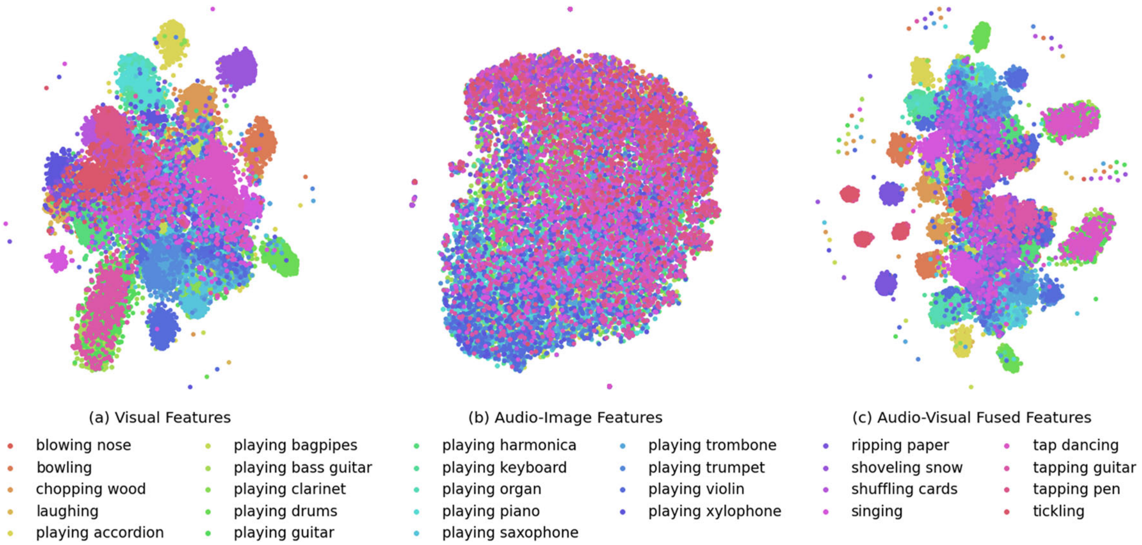
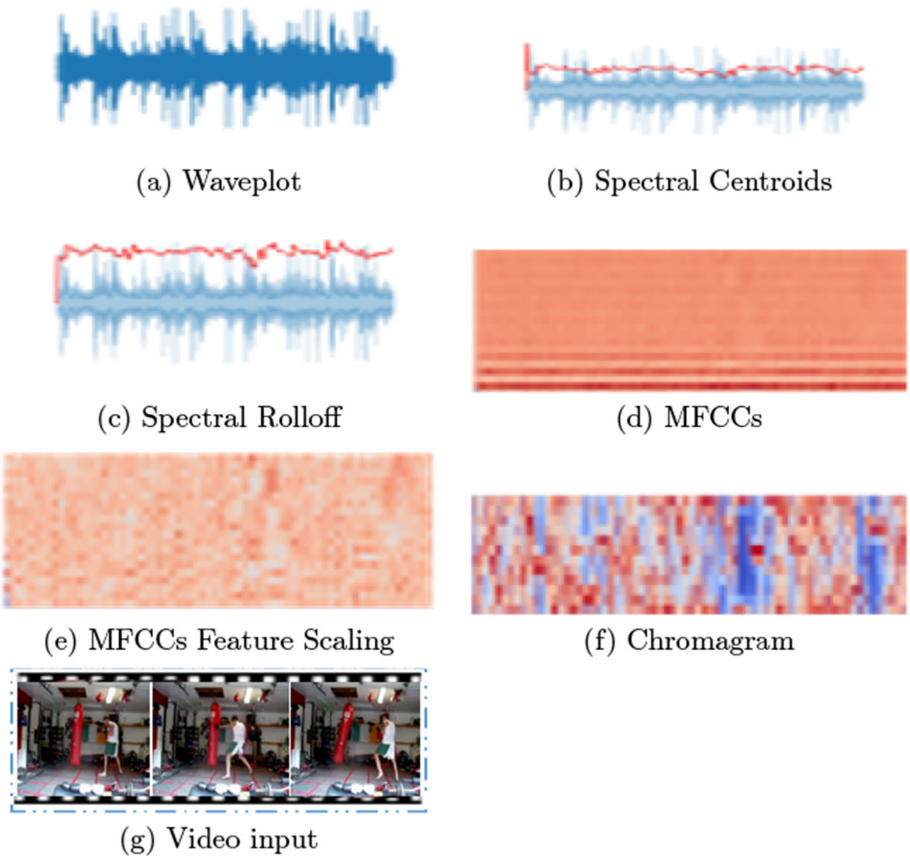


Fig. 8 t-SNE visualization of the Kinetics Sounds embedding of different features (video (left), audio (middle) and fused (right)) in the downstream path. Note that t-SNE embeddings do not use the class labels. Labels are only used during final visualization

Fig. 9 Examples of segmented video input and six different audio-image representations of the same action



in Fig. 2) on Kinetics-Sounds. For compatibility with t-SNE der Maaten and Hinton [41], we transformed feature maps from the average pooling layer to reduce the embedding dimensions to 2D. We observed a better clustering of the different action classes after the fusion of audio and video features in the framework.

6 Conclusion

In this paper, we proposed a multimodal audio-image to video action recognition framework called the Multimodal Audio-image and Video Action Recognizer (MAiVAR). We generated several audio-image representations and compared their efficiency. We then extracted audio and visual features using CNN-based models and fused them. The framework achieves 87.91% accuracy on the UCF51 dataset and 79.01% accuracy on the Kinetics Sounds dataset. Experiments demonstrated that visual modalities are not indispensable. MAiVAR performed better in comparison with other baseline methods.

Future research could focus on exploring different transformer-based architectures such as Video Transformer Network Architecture (VidTr) Neimark et al [44], with other optimal feature representations. The drawback of the fusion-based model is that using complex modality representations greatly influence computational complexity. To address this issue, data and model-parallelism could be used to simultaneously compute audio-visual modality streams instead of synchronous execution.

Acknowledgements This work is jointly supported 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. Dr. Akhtar is a recipient of Office of National Intelligence National Intelligence Postdoctoral Grant # NIPG-2021-001 funded by the Australian Government. The author would also like to thank School of Science, Edith Cowan University for providing GPUs facilities.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability Data will be available upon the request to the corresponding author.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Arandjelovic R, Zisserman A (2017) Look, listen and learn. In: IEEE, Proceedings of the ICCV, pp 609–617
2. Baldominos A, Saez Y, Isasi P (2018) Evolutionary convolutional neural networks: an application to handwriting recognition. Neurocomputing 283:38–52
3. Boehm KM, Aherne EA, Ellenson L et al (2022) Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nat Cancer 3(6):723–733
4. Brousmiche M, Rouat J, Dupont S (2019) Audio-visual fusion and conditioning with neural networks for event recognition. In: IEEE, Proceedings of the machine learning for signal processing (MLSP) Workshop, pp 1–6
5. Brousmiche M, Rouat J, Dupont S (2022) Multimodal attentive fusion network for audio-visual event recognition. Inf Fusion 85:52–59
6. Deng Z, Lei L, Sun H, et al (2017) An enhanced deep convolutional neural network for densely packed objects detection in remote sensing images. In: IEEE, proceedings of the remote sensing with intelligent processing (RSIP) workshops, pp 1–4
7. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: IEEE, Proceedings of The CVPR, pp 11933–11941
8. Feichtenhofer C, et al (2019) Slowfast networks for video recognition. In: Proceedings of the ICCV, pp 6202–6211
9. Gao R, Grauman K (2021) VisualVoice: Audio-visual speech separation with cross-modal consistency. IEEE, Proceedings of the CVPR, pp 15495–15505, https://doi.org/10.1109/CVPR46437.2021.01524
10. Gao R, et al (2020) Listen to look: action recognition by previewing audio. In: IEEE Proceedings of the CVPR, pp 10457–10467
11. Gao Y, Beijbom O, Zhang N, et al (2016) Compact bilinear pooling. In: IEEE, Proceedings of the CVPR, pp 317–326
12. Gaver WW (1993) What in the world do we hear?: an ecological approach to auditory event perception. Ecol. Psychol. 5(1):1–29
13. Gibbon DC, Liu Z (2008) Introduction to video search engines. Springer. https://doi.org/10.1007/978-3-540-79337-3
14. Girdhar R, et al (2017) ActionVLAD: Learning spatio-temporal aggregation for action classification. In: IEEE, Proceedings of the CVPR, pp 971–980
15. Gouyon F, Dixon S, Pampalk E, et al (2004) Evaluating rhythmic descriptors for musical genre classification. In: Proceedings of the AESIC, p 204
16. Gu J, et al (2021) NTIRE 2021 challenge on perceptual image quality assessment. In: IEEE, Proceedings of the CVPR, pp 677–690
17. He D, et al (2019) StNet: Local and global spatial-temporal modeling for action recognition. In: Proceedings of the AAAI conference on artificial intelligence, pp 8401–8408
18. He K, Zhang X, Ren S, et al (2016) Deep residual learning for image recognition. In: IEEE, Proceedings of the CVPR, pp 770–778, https://doi.org/10.1109/CVPR.2016.90
19. Herath S, Harandi M, Porikli F (2017) Going deeper into action recognition: a survey. Image Vis Comput 60:4–21
20. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: PMLR, Proceedings of the ICML, pp 448–456, https://doi.org/10.5555/3045118.3045167
21. Jing C, Wei P, Sun H et al (2020) Spatiotemporal neural networks for action recognition based on joint loss. Neural Comput Appl 32:4293–4302

22. Jung D, Son JW, Kim SJ (2018) Shot category detection based on object detection using convolutional neural networks. In: IEEE, Proceedings of the ICACT, pp 36–39
23. Kala R (2016) On-road intelligent vehicles: motion planning for intelligent transportation systems. Butterworth-Heinemann, OXford
24. Kay W, Carreira J, Simonyan K, et al (2017) The kinetics human action video dataset. arXiv preprint arXiv:1705.06950
25. Kazakos E, et al (2019) EPIC-Fusion: audio-visual temporal binding for egocentric action recognition. In: Proceedings of the ICCV, pp 5492–5501
26. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980https://doi.org/10.48550/arXiv.1412.6980
27. Kulkarni SR, Rajendran B (2018) Spiking neural networks for handwritten digit recognition-supervised learning and network optimization. Neural Netw 103:118–127
28. Kwon H, Kim M, Kwak S, et al (2021) Learning self-similarity in space and time as generalized motion for video action recognition. In: Proceedings of the ICCV, pp 13065–13075
29. Lei J, Li L, Zhou L, et al (2021) Less is more: clipbert for video-and-language learning via sparse sampling. In: IEEE, Proceedings of the CVPR, pp 7331–7341
30. Li Y, Zou B, Deng S et al (2020) Using feature fusion strategies in continuous authentication on smartphones. IEEE Internet Comput 24(2):49–56
31. Li Y, Tao P, Deng S et al (2021) Deffusion: Cnn-based continuous authentication using deep feature fusion. ACM Trans Sens Netw (TOSN) 18(2):1–20
32. Li Y, Liu L, Qin H et al (2022) Adaptive deep feature fusion for continuous authentication with data augmentation. IEEE Trans Mobile Comput. https://doi.org/10.1109/TMC.2022.3186614
33. Li Y, et al (2016) VLAD3: encoding dynamics of deep features for action recognition. In: IEEE, Proceedings of the CVPR, pp 1951–1960
34. Li Z, Tang J (2015) Weakly supervised deep metric learning for community-contributed image retrieval. IEEE Trans Multimed 17(11):1989–1999
35. Li Z, Tang J, Mei T (2018) Deep collaborative embedding for social image understanding. IEEE Trans Pattern Anal Mach Intell 41(9):2070–2083
36. Lidy T, Rauber A (2005) Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: Proceedings of the ISMIR, pp 34–41
37. Lin J, Gan C, Han S (2019) TSM: Temporal shift module for efficient video understanding. In: Procedings of the ICCV, pp 7083–7093
38. Long X, Gan C, De Melo G, et al (2018a) Attention clusters: purely attention based local feature integration for video classification. In: IEEE, Proceedings of the CVPR, pp 7834–7843
39. Long X, Gan C, Melo G, et al (2018b) Multimodal keyless attention fusion for video classification. In: No. 1 in Proceedings of the AAAI
40. Long X, De Melo G, He D, et al (2020) Purely attention based local feature integration for video classification. IEEE TPAMI pp 2140 – 2154
41. der Maaten LV, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(86):2579–2605
42. McFee B, Raffel C, Liang D, et al (2015) Librosa: audio and music signal analysis in python. In: Proceedings of the python in science conference, pp 18–25
43. Mei X, Lee HC, Ky Diao et al (2020) Artificial intelligence-enabled rapid diagnosis of patients with covid-19. Nat Med 26(8):1224–1228
44. Neimark D, Bar O, Zohar M, et al (2021) Video transformer network. In: Proceedings of the ICCV, pp 3163–3172, https://doi.org/10.1109/ICCVW54120.2021.00355
45. Paoletti M, Haut J, Plaza J et al (2018) A new deep convolutional neural network for fast hyperspectral image classification. ISPRS J Photogramm Remote Sens 145:120–147. https://doi.org/10.1016/j.isprsjprs.2017.11.021
46. Paszke A et al (2019) PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8024–8035
47. Patel CI, Garg S, Zaveri T et al (2018) Human action recognition using fusion of features for unconstrained video sequences. Comput Electr Eng 70:284–301
48. Roitberg A, Pollert T, Haurilet M, et al (2019) Analysis of deep fusion strategies for multi-modal gesture recognition. In: IEEE, Proceedings of The CVPRW, pp 198–206
49. Russakovsky O, Deng J, Su H et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115(3):211–252
50. Seo Y, Ks Shin (2019) Hierarchical convolutional neural networks for fashion image classification. Expert Syst Appl 116:328–339. https://doi.org/10.1016/j.eswa.2018.09.022
51. Shaikh MB, Chai D (2021) RGB-D data-based action recognition: a review. Sensors 21(12):4246
52. Shaikh MB, Chai D, Islam SMS, et al (2022) Maivar: multimodal audio-image and video action recognizer. In: IEEE, Proceedings of the VCIP, pp 1–5
53. Sharma N, Jain V, Mishra A (2018) An analysis of convolutional neural networks for image classification. Procedia Comput Sci 132:377–384
54. Slade S, Zhang L, Yu Y et al (2022) An evolving ensemble model of multi-stream convolutional neural networks for human action recognition in still images. Neural Comput Appl 34(11):9205–9231
55. Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402
56. Sudhakaran S, Escalera S, Lanz O (2020) Gate-shift networks for video action recognition. In: IEEE, Proceedings of the CVPR, pp 1102–1111
57. Szegedy C, et al (2017) Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proceedings of the AAAI, pp 4278–4284, https://doi.org/10.5555/3298023.3298188
58. Takahashi N, Gygli M, Van Gool L (2017) AENet: learning deep audio features for video analysis. IEEE TMM 20(3):513–524
59. Tan M, Le Q (2019) EfficientNet: rethinking model scaling for convolutional neural networks. In: Proceedings of the ICML, pp 6105–6114, https://doi.org/10.48550/arXiv.1905.11946
60. Tao W, Leu MC, Yin Z (2018) American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. Eng Appl Artif Intell 76:202–213
61. Tian Y, Shi J, Li B, et al (2018) Audio-visual event localization in unconstrained videos. In: Proceedings of the ECCV, pp 247–263
62. Tran D, Bourdev L, Fergu R, et al (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the ICCV, pp 4489–4497, https://doi.org/10.1109/ICCV.2015.510

63. Vandersmissen B, Knudde N, Jalalvand A et al (2020) Indoor human activity recognition using high-dimensional sensors and deep neural networks. Neural Comput Appl 32:12295–12309
64. Vinyes Mora S, Knottenbelt WJ (2017) Deep learning for domain-specific action recognition in tennis. In: IEEE, Proceedings of the CVPR Workshops, pp 114–122, https://doi.org/10.1109/CVPRW.2017.27
65. Wan S, Liang Y, Zhang Y (2018) Deep convolutional neural networks for diabetic retinopathy detection by image classification. Comput Electr Eng 72:274–282
66. Wang L, et al (2016) Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of the ECCV, pp 20–36, https://doi.org/10.1007/978-3-319-46484-8_2
67. Yan C, Teng T, Liu Y et al (2021) Precise no-reference image quality evaluation based on distortion identification. ACM Trans Multimed Comput Commun Appl(TOMM) 17(3s):1–21
68. Yang G et al (2022) STA-TSN: spatial-temporal attention temporal segment network for action recognition in video. PloS one 17(3):1–19
69. Zhang K, Li D, Huang J et al (2020) Automated video behavior recognition of pigs using two-stream convolutional networks. Sensors 20(4):1085
70. Zhou B, et al (2018) Temporal relational reasoning in videos. In: Proceedings of the ECCV, pp 803–818

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.