Human-centric Computing and Information Sciences

January 2024 | Volume 14

www.hcisjournal.com

KIPS Korea Information Processing Society

KIPS CSWRG Korea Information Processing Society Computer Software Research Group

HCIS | Human-centric Computing and Information Sciences

RESEARCH (Research Manuscript) Open Access

Human-centric Computing and Information Sciences (2024) 14:04
DOI: https://doi.org/10.22967/HCIS.2024.14.004
Received: January 3, 2023; Accepted: May 4, 2023; Published: January 30, 2024

Attention-Based Deep Learning Framework for Action Recognition in a Dark Environment

Muhammad Munsif[1,2], Samee Ullah Khan[1,2], Noman Khan[1,2], and Sung Wook Baik[1,2,*]

Abstract

Human interactions and action recognition (HIAR) are critical tasks in video analytics and understanding. While HIAR has garnered significant attention in the visible spectrum, other modalities, such as infrared (IR), have posed significant challenges in achieving accurate action recognition (AR) due to hazy surface textures and inappropriate features. Current mainstream approaches typically consider either action or contextual information to recognize human action and interactions in the IR spectrum. We propose a convolutional block attention-based action and interaction recognition network (AIR-Net) that reflects both action, contextual features and can be utilized in the dark, as well as visually light environments in a broad range of applications such as surveillance, security, and healthcare. The proposed method is mainly categorized into three steps. Firstly, the video stream is passed through pre-processing, which transforms the IR input data into a sequence of frames. Next, the action features are extracted, aided by a backbone model followed by a convolutional block attention module, while contextual features are extracted using a fine-tuned InceptionV3 network. Finally, a two-stream bidirectional long short-term memory (BiLSTM) network is applied to acquire temporal patterns individually. These sequence patterns are intelligently fused and forwarded to the associated uni-stream BiLSTM, which learns expedient information for action and interaction recognition. A comprehensive ablation study over two benchmark datasets (InfAR and NTU-RGB+D) demonstrates that our proposed method outperforms state-of-the-art methods. Specifically, the recognition performance is marginally improved on InfAR by 2.5% and achieved 80.94% on NTU-RGB+D.

Keywords

Human Action Recognition, Infrared View Action Recognition, Action Recognition in Low Light, Activity Recognition, Cognitive Computing in Our Daily Lives, Video Processing, Action Recognition at Night, Human-Machine Interaction

1. Introduction

In recent years, among the research community, there has been a growing interest to develop automated recognition of human activities with a particular focus on HAIR. This interest has been driven by the increasing availability of smart devices, which enable the creation of cognitive and communicative

***Corresponding Author:** Sung Wook Baik (sbaik@sejong.ac.kr)
[1]Department of Software Convergence, Sejong University, Seoul, South Korea
[2]Department of Convergence Engineering for Intelligent Drone, Sejong University, Seoul, South Korea

environments. Cognitive computing, which combines artificial intelligence (AI), machine learning (ML), and neural networks, have been used to effectively address everyday issues. [1]. HAIR is a fundamental problem in computer vision (CV) research, which aims to recognize human-to-human interactions and actions. It has attracted significant attention from researchers due to its wide range of applications in fields such as urban security [2, 3], autonomous vehicles [4], human-computer interactions [5–8], and urban planning [9]. HAIR is a very challenging and complex task and has various complications including intra-class disparity, class viewpoint variations, action motion distinctions, contextual clutter, and human occlusions [10, 11]. Recent advances in ML and deep learning (DL), have led to the development of numerous approaches for visible spectrum-based HAIR, including classical ML and convolutional neural networks (CNNs) [12, 13]. Two streams of DL-based methods have improved the performance of HAIR in the visible spectrum by fusing the output feature vectors of separate networks. Therefore, one network learns spatial appearance and another extracts motion patterns. Furthermore, [14] it presented a multiple streams framework, which split each video sequence into several sequences and processes it by a two-stream architecture. The combination of the feature vectors at the end improved recognition performance. To enhance HAIR performance, many other studies such as [15, 16] have explored the pose sequence's skeletal information and aimed to extract robust representations of human actions. However, visible spectrum-based methods are unable to provide useful information in uncertain environments such as low-light conditions including smog, fog, and snow.

The existing state-of-the-art methods have achieved great success when using visible videos, however, there is still a gap causing an unsatisfactory performance in complex scenarios such as low-light conditions including dark, fog, or even smoky situations [17, 18]. In such conditions, IR or night vision technology streams are more appropriate due to their resilience to low lighting and background clutter conditions. Limited studies have addressed HAIR in IR videos [19–21]. For instance, [16] proposed a single-stream CNN-based framework for AR in the IR spectrum. Next, Jiang et al. [19] developed a two-stream architecture consisting of two-stream 3D CNN networks to learn to distinguish representations for each class.

The above-mentioned techniques have focused on integrating motion such as optical flow or contextual representation images [22] using single or multiple streams. However, only using one stream or growing the number of input streams and considering the same type of features is not enough to deliver robust and discriminative HAIR patterns. This is due to three main reasons. Firstly, most of the existing techniques do not discriminate and locate the action of the main participant in multi-person videos and only extract scene-level patterns, without giving attention to key actors participating in the ultimate action [23]. Most of the time, camera streams in real scenarios tend to contain multiple people interacting or cooperating and performing an ultimate action. The demonstration is given in InfAR [24] and NTU-RGB+D [25], for instance, in “Handshake,” and “Hug,” etc., where actors act and interact with each other in an indoor and outdoor environment, and one or two humans act and others make context for it. Samples of this demonstration are depicted in the results section. Thus, considering that, action features of the trained model on a single participant stream can create confusion sometimes between intra-classes when recognizing the action in the real scenario. It can mislead and avoid the model's attention from extracting the overall specific clues, along with keeping attention on the main actor to understand an action accurately. Secondly, these methods learn only redundant action features instead of focusing both on contextual and other relevant action patterns. Thirdly, most of the architectures do not allow information sharing among multiple sequence learning modules, encode only the actions or contextual features without ensuring the preservation of the distinguishing features and clues using attention or pooling technique.

To address mentioned limitations and gaps in accurately recognizing actions and interactions in the IR spectrum, we propose a context and action-aware HAIR network called AIR-Net. Our approach involves three stages: firstly, extracting key person actions from video streams using the EfficientNetB7 model extended by convolutional block attention module (CBAM) attention. This step allows for the exclusion of irrelevant information and focuses solely on action-related features. Secondly, overall scene information is extracted by considering the surroundings of the main actors, which can compensate for missing texture

and color features in IR-based sequences. To account for the failure of contextual feature extraction, InceptionV3 is utilized to leverage complementary distinguishing capabilities for IR learned from the visible spectrum. This allows the extraction of contextual features which can be analyzed alongside temporal action information. Lastly, the extracted features are processed sequentially to encode both contextual and action information, thereby obtaining more related patterns of the performed action, and passing them to the ultimate action and interaction predictor end. Furthermore, the summarized contributions of this study are as follows:

- A unified framework is proposed for HIAR for IR spectrums that simultaneously consider dual discriminator features; action features extraction draw attention to specific action portions throughout a sequence of frames, while context features extractions provide an understanding of the overall scene by keeping texture and background information in view for ultimate AR.
- Current motion features extraction techniques are computationally expensive, and it is difficult to extract meaningful information and manage them. To overcome this, action features are explored as they are acquired from CBAM attention, along with a backbone model that specifically focused on actions-related features in IR video sequences.
- To the best of our knowledge, context-aware features are not considered for IR streams. The proposed AIR-Net reflects context-aware features that are supportive of understanding the scene for accurate HAIR in different contexts.
- An extensive ablation study is conducted, where the influence of context and action-aware modules are deeply analyzed along with sequential modules (long short-term memory [LSTM] and bidirectional long short-term memory [BiLSTM]), revealing a convincing performance with a marginal improvement of 2.5% compared to existing state-of-the-art methods on InfAR, as well as an 80.94% benchmark performance achieved for the NTU-RGB+D IR.

The rest of the article is organized into Sections 2 and 3, containing the related work and the proposed method respectively, including a detailed discussion of different components. Section 4 consists of the evaluation of the proposed technique on various data and its results, while conclusions along the future work are presented in Section 5.

2. Related Work

In this section, a thorough review of HAIR's most pertinent work has been conducted, with particular emphasis on various feature extraction mechanisms for both visible and IR spectra. To enhance understanding of the existing research, this section has been structured into three main categories, namely conventional, DL, and attention-based approaches. A detailed account of the reviewed content for each category is provided below.

2.1 Conventional Approaches

Early works of the HAIR employed manually crafted features derived from a series of mathematical and statistical procedures, such as translations, rotations, and trajectory calculations of human body movements. In [26], a view-invariant motion descriptors-based technique was proposed for human AR (HAR), which accounted for changes in camera viewpoints relative to the central axis of the human body. Subsequently, in [27], the authors developed a technique to predict camera motion and accurately learn the action trajectory for HAR. They utilized a covariance matrix descriptor in conjunction with a classification technique for the estimation of skeletal joint movements, ultimately leading to the categorization of human actions. Despite their success, conventional methods relied on manually engineered features that only capture local patterns, thus limiting their effectiveness in accurately predicting complex human activities.

2.2 Deep Learning Approaches

DL-based techniques have emerged as the dominant approach in the field of CV, owing to their superior performance compared to handcrafted methods in various fields such as human-robot interaction [28–30], image representations [31–34], health care [35], including HAR [36, 37]. Several researchers have proposed CNN-based models for this task [38]. For instance, a study [39] incorporated CNNs for HAR by proposing a two-stream technique that leveraged RGB frames and optical flow features to extract spatial and temporal information for overall scene and specific AR, respectively. Similarly, Tran et al. [40] developed a network called 3D convolutional networks (ConvNets) that could simultaneously extract spatial and motion patterns through 3D ConvNets, demonstrating that their model was more appropriate for HAR than 2D CNNs. Another 3D ConvNet, called I3D-based two-stream networks, was proposed in [41]. The authors utilized special types of filters, such as inflate pooling filters primarily used for DL-based image classification, to extract robust spatial and temporal pattern information from video streams. However, limited DL-based research has been conducted on the IR spectrum. For instance, Gao et al. [24] proposed a CNN-based framework consisting of two streams for activity recognition in IR. Similarly, Jiang et al. [19] developed a model for IR data that incorporated optical flow features along with raw data, utilizing a two-stream 3D CNN with an extension of integration of a code layer to learn discriminative category-based patterns. Subsequently, various other multi-stream methods were proposed for accurate HAIR. For instance, Liu et al. [42] developed a three-stream architecture that could learn local, global, and spatial-temporal features for better human action representation. However, most of these methods have focused on the visual spectrum and only encode action or context features for AR, making them less robust in extracting meaningful action representations. Moreover, a type of network spatially designed for sequential data modeling, known as recurrent neural networks (RNNs), has also been considered suitable for AR, with several RNN-based techniques proposed for HAIR. For example, Gammulle et al. [43] proposed LSTM along with self-prediction to extract salient frames features in a video. Meanwhile, authors of [43] developed a deep fusion architecture that could extract spatial patterns from different levels of CNN layers and map them with temporal features using LSTM layers. However, RNN and LSTM base techniques often suffer from vanishing gradients with longtime stamps and are difficult to optimize [27]. Furthermore, these methods process data sequentially, which limits their ability to process it in parallel and focus on interest points in long sequences to extract appropriate contextual and action patterns from the entire scene.

2.3 Attention-Based Approaches

Recent advances in attention-based mechanisms have shown promise in achieving satisfactory performance in various spatial and temporal-based tasks, including image classification, audio data analysis and numerical data understanding [30], and AR. For instance, in AR, attention mechanisms can extract continuous sequences of streams without the need for processing data in a sequenced manner, which can be useful in identifying activities such as walking, jogging, or hugging [41]. However, limited research has been conducted on attention-based AR in both the visible and IR domains, particularly in the IR domain. In the visible spectrum, state-of-the-art performance has been achieved using attention mechanisms. For example, in [43], the authors proposed an attention mechanism-based technique for extracting informative video frames in the input sequence of visible video frames. Similarly, in [44], an attention layer was incorporated into a sequential network, the recurrent spatial-temporal network, for the extraction of contextual features for each timestep. Other attention-based methods, such as the cross-layer attention layer and the center-guided attention mechanism, have been proposed to aggregate important local features from each feature map and generate an ultimate representation of the input sequence [45]. Next, the transformer attention-based technique developed by Zhang et al. [31] for skeleton-based AR used a graph convolutional transformer for the extraction of spatial (position of the joints) as well as temporal (velocity of the motion) features extraction along sequences of skeleton information. Attention-based methods have also been explored for AR in the IR domain [21]. There, special features were extracted via a convolution layer, while triple layers convolutions were used for

temporal features extraction across the frames, and a ConvLSTM was used for the final prediction. This approach achieved better recognition scores, but it was limited to single-person actions with an airplane background and could not handle complex or interactive activities with high performance.

While all of these methods have been developed for HAR in the visible spectrum, few studies exist in the IR spectrum, despite its high potential for handling HAR in low-light conditions such as at night, in the fog, or in smoggy situations [24, 46]. Additionally, previous approaches have focused on integrating additional information streams, such as raw, optical flow, and motion visual streams of optical flow [47] but have not been able to extract more discriminative features that can accurately categorize an action. Hence, it is highly desirable to extract a more robust representation for ultimate accurate HAIR specifically in the IR spectrum, to utilize its full potential in a vast variety of applications. In this study, we investigate an effective way of action and contextual feature representations, as well as obtaining discriminative patterns for eventual action and interaction recognition in easy, as well as complex conditions.

3. Proposed Method

In this section, the proposed method for HAIR is introduced, which comprises of preprocessing, action, context feature extraction, and sequential learning modules. The factorial representation of the approach is presented in Fig. 1, the procedure is given in Algorithm 1 and the detailed explanation is provided for each subsection as follows.

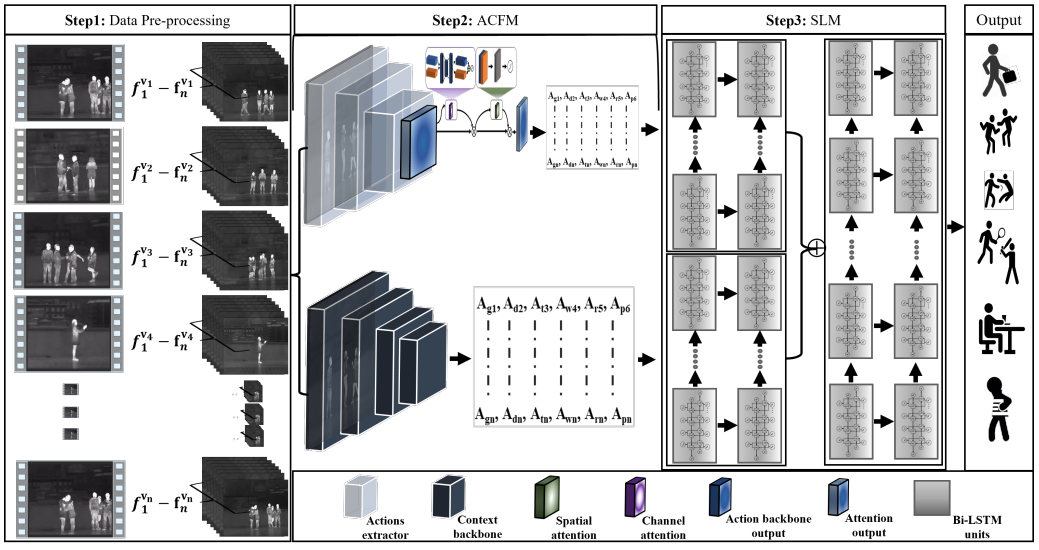

Fig. 1. Overview of the proposed framework, which mainly contains three steps. Step 1 pre-processes the acquired data by splitting it into a range of sequences. Next, the action and contextual features module (ACFM) is in step 2, where action features are extracted through a backbone convolution-based model extended by CBAM attention that extracts specific action features. It is followed by another state-of-the-art convolution-based module to compensate for the scene understanding information. Finally, the sequential learning module (SLM) decides a final prediction.

3.1 Pre-processing

Data pre-processing is one of the essential steps of the proposed method. As shown in Fig. 2, we first convert videos into frames to perform low-level pre-processing on each frame, including resizing the frame to 224 by 224 instead of processing the original resolution of frames, and ultimately reducing the

computational complexity of the proposed framework. The resized frames of each video are arranged in a proper sequence by considering Equations (1) and (2).

$$S = \left[\left[s_1^{v_1}, s_2^{v_1}, \ldots . . s_n^{v_1}\right], \left[s_1^{v_2}, s_2^{v_2} . . , s_n^{v_2}\right], \ldots . . \left[s_1^{v_n}, s_2^{v_n}, . . s_n^{v_n}\right]\right], \quad (1)$$

where S represents sequences of frames, and each sub-series indicates the number of sequences of frames that can be represented in Equation (2). Each sequence s in Equation (1) set of sequences of frames is denoted by f in Equation (2):

$$s_{s_n}^{v_n} = \left[f_1^{v_n}, f_2^{v_n}, \ldots \ldots f_n^{v_n}\right]. \quad (2)$$

We arranged the pre-processed frames in different lengths of sequences including 10, 20, 30, and 100, and selected 10 sequence lengths.

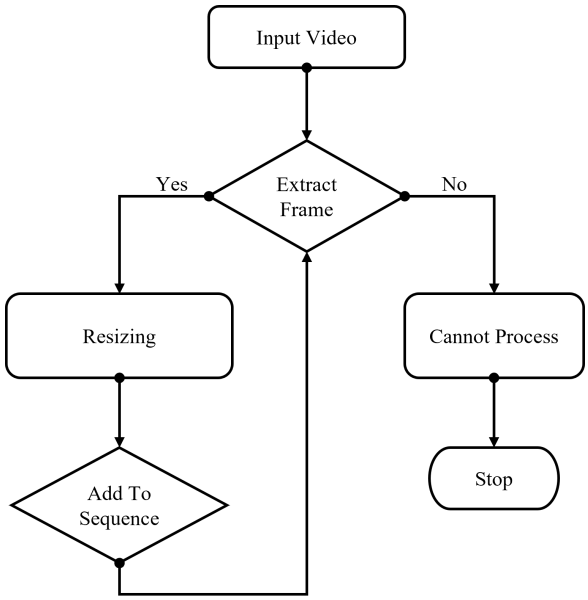

Fig. 2. The flow of the pre-processing procedure.

3.2 Action Aware Module

HAIR in complex and real-world scenarios mostly contains more than one person interacting with each other. In addition, it is also possible that not all of them will be involved in the ultimate action, making HAIR even more challenging. Therefore, it requires a deep analysis of the scenes. One solution to solve it can be to focus only on the actors presenting an action. However, it requires annotations in each frame of the action area that are expensive in terms of resources and time and are not easily available. Another technique is the automatic region of interest focusing, where models automatically draw attention to a certain area where the corresponding features are available. Inspired by the need for automatic key-actors' area discovery in scenes, we develop a new approach to focus only on the active region which locates the main actors involved in the action.

To identify the main action area that is considered the most informative and discriminative information, we train a DL model on preprocessed frames as shown in Fig. 1. The action submodule consists of the adopted SOTA backbone EfficientNetB7 [48] and CBAM introduced in [49], focusing only on the human instances performing actions. We extract features and these features are converted again into sequences, according to Equations (1) and (2). Next, the mathematical representation of the CBAM module is provided in Equations (3) and (4):

$$R_f = (S_A(F_{mc}) * F_{mc}) \tag{3}$$

$$F_{mc} = (C_A(F) * F) \tag{4}$$

where intermediate features are denoted by F, passed from the channel attention module (CAM) represented by C_A, followed by the spatial attention module (SAM) receiver's output of the CAM denoted by F_{mc}, as input and produce refined features. To further elaborate, the function of CAMs as shown in Fig. 3(a) calculates channel attention features by performing max and average pooling of the intermediate features generated by the backbone separately. Then, it feeds them to a sigmoid activation function after element-wise summation of the output of shared MLP layers. Finally, it produces a vector of features, representing channel-wise attention C_A.

Output maps are also called refined features represented by R_f, which are generated via the inter-spatial connection of features. It is different from the C_A which looks for "what" is to be obtained in the channel, while spatial attention (S_A) looks for "where" is an informative part of a channel. As shown in Fig. 3(a), S_A computes by applying average and max pooling on all the input channels which have proved as an effective approach towards the identification of the informative region in [33] and concatenate these to produce an efficient feature extraction descriptor. Furthermore, it creates a S_A at the end of the convolved concatenate and passes to a sigmoid activation, then generates an S_A 2D refined feature map. Finally, the output of the S_A concatenates with F_{mc} and produces CBAM attention, which is considered the final output of the action module.

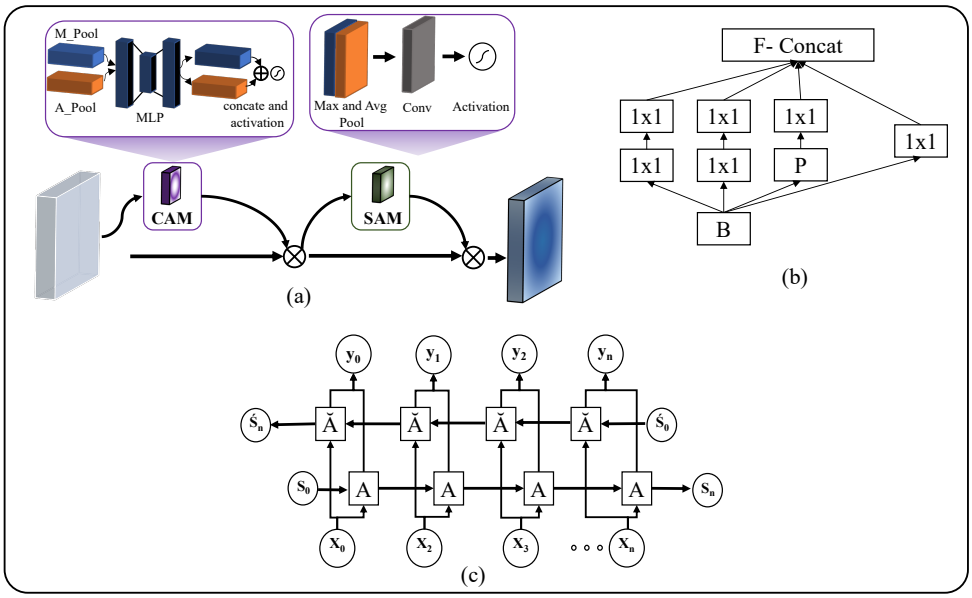

Fig. 3. The main components of the proposed features and SL modules include CBAM with channel and spatial sub-modules, InceptionV3 block for multiscale feature extraction, and a BiLSTM structure.

3.3 Context-Aware Features Extractions Module

The context in the action provides a clue for the ultimate HAIR model via contextual information encoding around the key actors. This information will enable the proposed model to deal with the effect of challenges surrounding the main action. We can call it the context of the action, and these representations are essential for the robust optimization of the ultimate analysis model. However, most researchers, especially those in the IR spectrum, did not consider the contextual information due to the unavailability of clear information in terms of texture and color of the surroundings. As provided in

Algorithm 1, we extract 2D features representing mostly the surroundings of actors in each frame via the state-of-the-art model InceptionV3 proposed in [50]. The basic block of the architecture of it is shown in Fig. 3(b) and contains different components connected with a skip connection. We extract features using the inception V3 model from the IR data by processing video sequence frames with width *W* and height *H* and acquiring desired features. Thus, across all frames, the temporal sequences locate the location of context features around the main actors at each frame, and these sequences of features are passed to SLM along with action features discussed in Section 3.2.

Algorithm 1. Pseudocode of the proposed AIR-Net for HAIR

Input: Video (VÏ) = {V1, V2,….Vn}
Output: Action label→ (AL) ={C_1, C_2, and C_n}
Pre-Initialization:
Load Training Dataset →(**(**TĎ)= {V_{T1}, V_{T2},… V_{Tn}}
(Validation)→(VĎ) = {V_{l1}, V_{l2},… V_{ln}}
Model Training Procedure:
Load contextual module → (CM)
Tune AND load action module (AM)
Sequential Model hyperparameter initialization →(ɵ)
Steps:
for VÏ →{V1, V2,… Vn}:
  while ***F*** ≠Fn do:
    extract context features → ĊM (F) → (ς)
    extract → action backbone features→ (Ɓ)
    → CBAM→ Ɓ→Č(Fm); Š(Fm)
    → early fusion→Ž→ (Č(Fm) Ө Š(Fm)) Ө (Frm))
    → Action Bi-LSTM → ABL (Ɓ) → Ars
    → Context Bi-LSTM → CBL → CBL(ς) → Crs
    → late fusion→L→ Ars Ө Crs → (FS)
    → FS → FC→FC→ Softmax
  end
end
Save sequential (model)→ **Ω**
Model Testing Procedure:
Load → ς
Load → Ɓ
Load →**Ω**
Load Testing Data →(**(**TestD)= {V_1, V_2,… V_n}
for (Vi→ Vn) → { Vt1, Vt2, Vt3 ... Vtn}
  → extract features → ς (Vi) & Ɓ (Vi)
  →Pass from **Ω** and predict AL
  →Store AL and Put text → Fn of Vn
End

3.4 Two-Stream Sequential Learning and Final Action Prediction Module

Since HAIR is formed from the set of actions that covers the consecutive frames of sequences, we process the extracted action and contextual features separately as shown in Algorithm 1, using two-stream BiLSTM to consider separate sequential patterns and then combining them in a single BiLSTM to learn the ultimate patterns for the final prediction. The prediction model consists mainly of the BiLSTM layers. BiLSTM is an advanced version of RNNs. RNNs models are also used for the sequence learning process by different researchers. However, RNN networks are unable to process long sequences due to the

vanishing and exploding gradient problem. LSTM is a special version of RNN that solves the issue to some extent, by introducing memory and forgetting cells. However, LSTM also has limited memory and fails when the sequence of input data increases due to vanishing and exploitation gradient problems [32].

To consider both forward and backward relations and avoid the vanishing gradient problem for input data, in this work we used BiLSTM as shown in Fig. 3(c), a special variant called BiLSTM using Equations (5) to (11). Where, U, ɵ are represents wights and C is the bias term:

$$I_t = ɘ(U_I * S_t + ɵ_I * h_{t-1} + C_F). \tag{5}$$

In Equation (5), it denotes the input gate at a time, ɘ is the sigmoid function, while S_t is the input feature at time t, and h_{t-1} represents the output last LSTM unit. This gate determines information about the previous unit that needed an update.

$$F_t = ɘ(U_F * S_t + ɵ_F * h_{t-1} + C_F). \tag{6}$$

At the same time, F_t shows the forget gate, computes the essential information, and forgets unnecessary old information.

$$è_t = \sigma(U_{ĉ} * S_t + ɵ_{ĉ} * h_{t-1} + C_{ĉ}), \tag{7}$$

$$ĉ_t = ɘ(F_t . ĉ_{t-1} + I_t . è_t). \tag{8}$$

Equation (7) calculates the candidate's state $ĉ_t$ using σ (tanh) function by matrix multiplication of weights, input features, and bias for each time stamp. While the present state $ĉ_t$ computes using Equation (8), where $ĉ_t$ denotes element-wise multiplication of forget, input gate.

$$O_t = ɘ(U_O * S_t + ɵ_I * h_{t-1} + C_t), \tag{9}$$

$$H_t = O_t \, . \sigma(ĉ_t). \tag{10}$$

Further Equations (9) and (10) calculate output gate O_t and the final output of the LSTM H_t, respectively, which considers normal LSTM called forward the sequence analysis LSTM. At the same time, it is clearly a big possibility to lose the information in a spectrum where the texture information is not available, being lost if features are considered only in the forward or back direction. We used BiLSTM as mentioned above, to make up back as well as forward directions' information. This enhances the generalization and robustness power of the proposed model, where $(\overleftarrow{H_t})$ represents back and $(\overrightarrow{H_t})$ forward variant of LSTM, and it is known as BiLSTM [51].

Finally, the output of the BiLSTM model flattens using the flatten function and is passed to a fully connected network to learn a complex curve of produced representation of sequence features. Additionally, it will create a final output in terms of a probability for each class of the data and optimize the whole action learning and prediction module, according to learning parameters discussed in the next section.

4. Results

This section is dedicated to the performance evaluation of the proposed framework. First, the experimental settings used throughout experiments are discussed and the data used for evaluation is discussed. At the end, a discussion is held on the ablation study and comparative analysis with the state-of-the-art methods.

4.1 Experimental Setting

The experiments were conducted on a PC with a Windows operating system, equipped with an NVIDIA GeForce-RTX 3090 GPU, 48 GB RAM, and an AMD Ryzen 93900X 12-Core Processor with

a clock speed of 3.79 GHz. The proposed framework was implemented and executed via Python 3.8, using the TensorFlow version 2.4.1 library [52]. Contextual features' extraction was performed using InceptionV3 [50], a residual CNN pre-trained model on ImageNet [53], while action aware feature extraction utilized the EfficientNetB7 [48] architecture with an extension of CBAM attention [49]. The ultimate action learning and prediction model on sequences of extracted features were trained with the Adam [54] optimizer and a set of hyperparameters, including β1 = 0.9, learning rate = 0.0001, were fixed after manual optimization. The prediction model was trained for 100 epochs with a batch size of 64, with each epoch requiring 1 minute for training. The summarized time complexity of the proposed model for each module used in the proposed model, including features extraction (action and contextual modules) and the final action prediction module, is tabulated in Table 1. It can be observed that the contextual module (CM) and Action module (AM) have more than 189 million parameters, while the sequential models have only 0.20 million (M) parameters. The sequential model's performance is dependent on the sequence of the input data, as longer sequences require more time for training and inferencing, and vice versa. In this case, our model took 1 minute (m) for training and 0.04 seconds (s) for inferencing on 20 sequences of frames. Further evaluation of the proposed technique is discussed in the upcoming subsections.

Table 1. Time complexity

Module	Number of parameters (M)	Training time per epoch (m)	Inference time per sequence (s)
Contextual module (CM)	189.53	-	2
Action module (AM)	190.18	-	2.5
Action prediction module (APM)	0.20	1	0.04

The various modules of the proposed AIR-Net consist of CM, AM, and APM.

4.2 Evaluation Metrics

The model evaluation was performed using standard metrics, including accuracy and precision. Average precision (AP) as [24, 55] is used for performance on InfAR. Besides this, the class-wise and average accuracy-based performance is also presented for both datasets. The mathematical representation of accuracy is presented in equation 11, where the total accurate prediction (CP) divides by the total number of predictions (AlP) and receive accuracy (Acc), while the AP formula is given in equation 12 where the true positive (TP) accumulation divide by the accumulated number of TP and false positive (FP).

$$Acc = \mathfrak{m}(\frac{CP}{AlP}), \tag{11}$$

$$AP = \mathfrak{m}\left(\frac{TP}{TP+FP}\right). \tag{12}$$

4.3 Datasets

Two different standard datasets, including InfAR [24] and NTU-RGB+D IR [25], are used throughout the evaluation of the proposed framework. InfAR has solo and interaction activities recorded in a complex outdoor background, while NTU-RGB+D IR has mutual actions recorded indoors and all the selected classes are from the interaction activity in which two actors are performing an activity. Samples frames of randomly selected activities of both datasets are shown in Fig. 4 and the distribution of data among classes of NTU-RGB+D dataset is shown in Fig. 5. Further details of each dataset are as follows.

4.3.1 InfAR

InfAR is one of the most challenging and popular IR activity datasets, consisting of 12 classes containing human interaction as well as solo activity data. The whole dataset contains 600 video clips, and each class

has 50 videos recorded with IR thermal imaging cameras. Each video is annotated with one activity from 12 specified human mutual or solo activity classes. The average recorded clip length is around 4 seconds with 25 frames per second speed, having a resolution of 293×256. Furthermore, as shown in Fig. 4, this dataset has some actors not involved directly in the ultimate activity but making the data challenging and providing a context for each class. Likewise, for better evaluation, each class video is divided into training (60%), validation (25%), and testing (15%), as recommended by most of the SOTA studies [24]. Training data is used for training the models and training time validation is performed via validation data, while test data is used for evaluating the model in real-time.

Fig. 4. Visual samples of different classes from both InfAR and NTU-RGB+D datasets. Both dataset samples are separated with a dotted line, above the line representing the InfAR and below the line showing samples of the second dataset.

4.3.2 NTU-RGB+D

RGB+D is one of the large scales publicly available datasets proposed especially for activities (daily, interaction, and medical activities) consisting of four modalities. One very essential modality is the IR modality, which we used in this study for evaluating the proposed framework. The data is recorded in video format, from 106 subjects of different genders (male and female), ages (19–57 years), and ethnic backgrounds (15 countries). The IR and others are recorded with Microsoft Kinect v2, and 155 distinct camera viewpoints considered 96 variations of illumination and backgrounds. These various subjects, viewpoints, backgrounds, and illuminations make it a very standard dataset to train and evaluate the

activities' prediction framework. The IR modality of this data consists of more than 0.1 million videos and 120 different classes discussed above, with an average range of 4 seconds recorded and 30 frames per second. Each class of data consists of an average of 882 videos and the detailed distribution of the data is given in Fig. 5.

In this study, as mentioned above, our focus is on the complex interaction activities, therefore we only get the interaction activities classes which are 26 in total. Similar to the first dataset, we used 60% data for training, 25% for validation, and the remaining 15% is used for evaluating the model in real time.

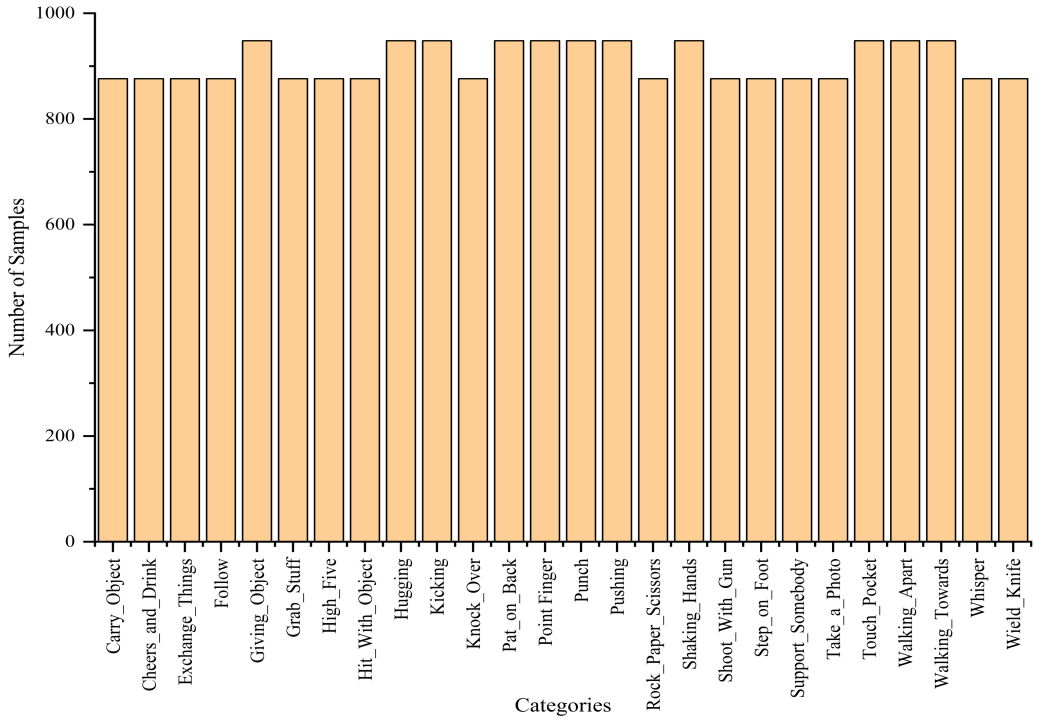

Fig. 5. Data distribution of NTU-RGB+D IR dataset.

4.4 Ablation Study

To achieve an optimized framework for action classification, we conducted extensive experiments including a comprehensive ablation study consisting of numerous experiments. We utilized various module combinations, such as features extraction, attentions, and sequential modules to show the significance of the consideration of action features extracted with EfficientNetB7, as well as an important type of attention module called CBAM, and a contextual module used to acquire context of the activity or interaction via a state-of-the-art InceptionV3 model. The results of the ablation study are tabulated in Table 2. Where both datasets (InfAR and NTU-RGB+D) are used throughout the experiments, it can be observed that the performance of the network is increased when we consider both action and contextual modules with the incorporation of SA, as well as CA. Furthermore, it is also found that the performance of the ultimate prediction of the proposed framework has a direct dependency on the nature and ability of the sequential models. We found an almost 8% increase in the performance of the HAIR framework by processing the extracted features with BiLSTM. The main reason for this is that the BiLSTM looks for both forward and backward sequences and most of the activities have very complex overlapping in terms of similarity, such as among push and punch classes, and are unable to distinguish only looking for forward sequences. The result, as shown in Table 2, achieved the highest accuracy of up to 86.75% and 80.94% on InfAR and NTU-RGB+D datasets, respectively, when using action modules with CBAM,

contextual modules, and BiLSTM, and put a benchmark for the IR spectrum HAIR. Next, class-wise and comparative evaluations with existing techniques are discussed below.

Table 2. Detailed ablation study using the different combinations of FEM including AM and CM, attention module, and sequential module on both infAR and NTU-RGB+D datasets

FEM	Attention module	Sequential module	InfAR	NTU-RGB+D
AM	-	LSTM	50.4	45.0
	SA	-	55.0	52.0
	CA	-	56.0	51.8
	CBAM	-	60.0	54.0
CM	-	-	47.5	43.2
	SA	-	51.2	47.3
	CA	-	45.8	39.7
	CBAM	-	49.0	45.6
AM+CM	SA	-	72.7	56.3
	CA	-	69.0	63.2
	CBAM	-	78.2	72.0
	CBAM	BiLSTM	86.7	80.9

FEM=feature extraction module.

To investigate the performance deeply in terms of accuracy and robustness of the proposed framework, calculate the class-wise accuracy as illustrated in Fig. 6 for InfAR and in Fig. 7, where the class-wise accuracy for NTU is depicted. It can identify that the proposed model has up to or more than 86% accuracy for most of the classes, for both NTU-RGB+D and InfAR dataset.

However, very few classes such as grabbing stuff and hitting with objects in NTU and Wave1 in InfAR datasets have the lowest accuracies due to high complexity, overlapping, and similarity with other classes. In NTU datasets, the highest accuracy has been achieved as up to 100% by patting on the back, touching pocket, and walking toward. However, the lowest performance has been found for grabbing stuff. On the other hand, for the InfAR, the framework has above 90% accuracy for handclapping, fighting, handshakes, and Wave2 classes, while the lowest performance has been found only for the wave1 class.

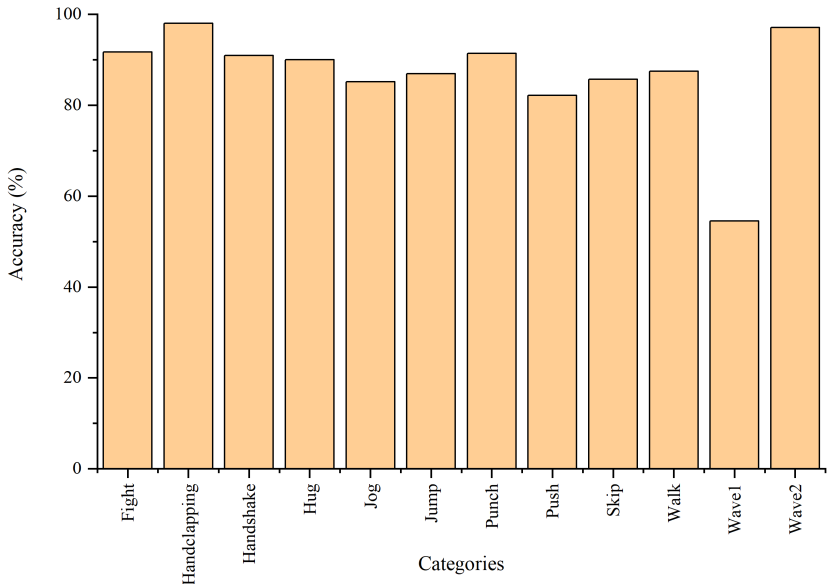

Fig. 6. The class-wise accuracy of the proposed framework on InfAR.

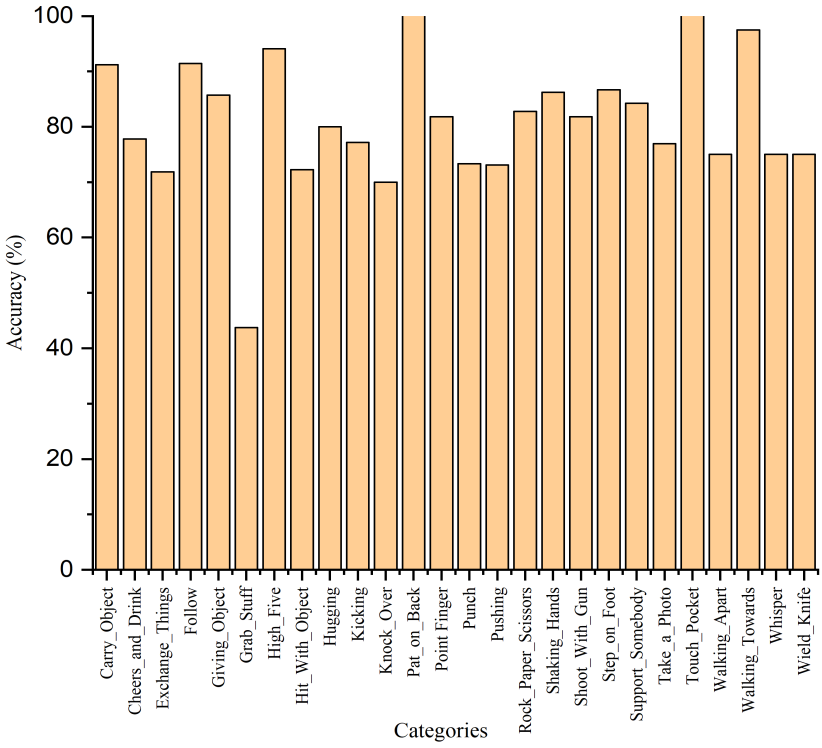

Fig. 7. The class-wise accuracy of the proposed framework on NTU-RGB+D.

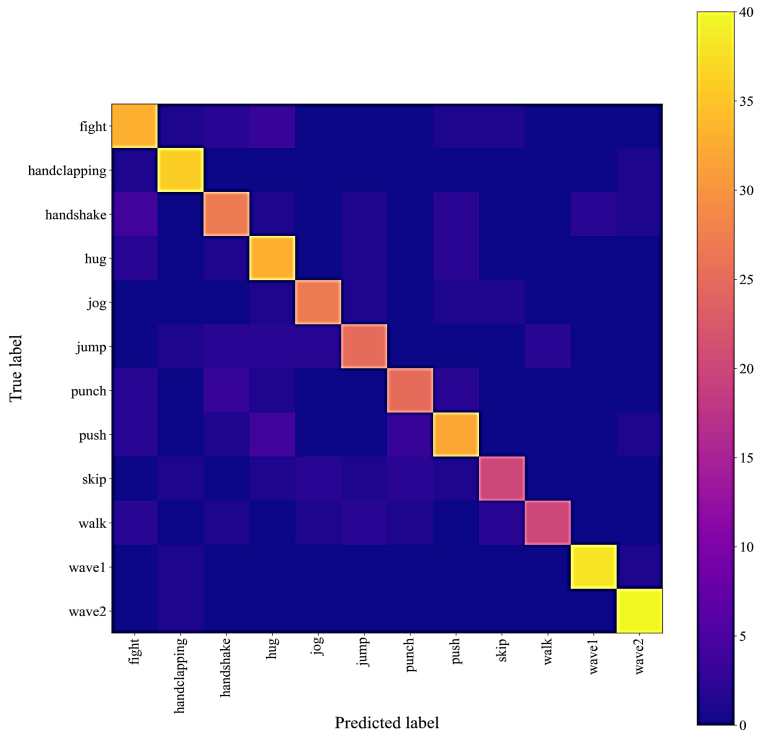

Fig. 8. Achieved confusion metrics using the proposed system on InfAR dataset.

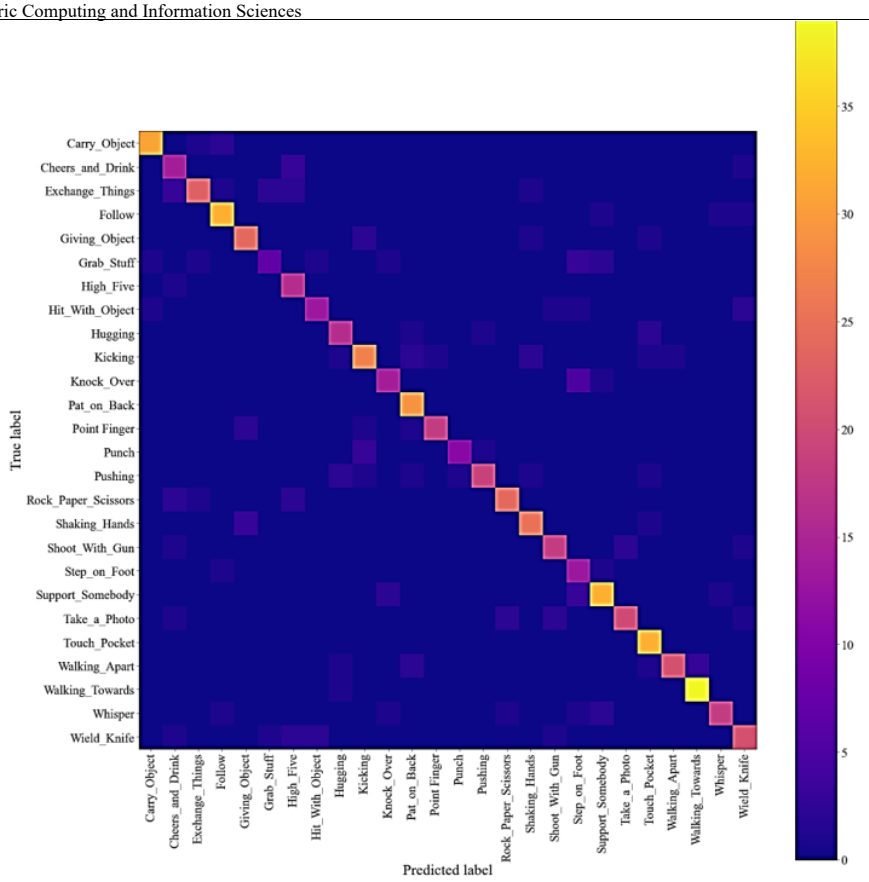

Fig. 9. Achieved confusion metrics using the proposed system on the NTU-RGB+D dataset.

Further investigation of the proposed framework has been obtained by prediction performance and for the actual and predicted we calculate confusion matrix for both datasets. As shown in Figs. 8 and 9, x-axes of the confusion matrix show the predicted labels, and the y-axis indicates the actual labels for both datasets, InfAR in Fig. 8 and NTU-RGB in Fig. 9. The diagonal elements of the confusion matrix correspond to the instances that were accurately classified for each class, while the confusion among classes is presented as rows of each class. In the InfAR, the fight is mostly confused with a hug due to their similarity and for other classes, most of the time model prediction is correct with little confusion. Similarly, for NTU-RGB+D, the grabbing stuff is confused with many classes, mostly with supporting somebody and patting on the back, confusion among classes for other categories are presented in Fig. 9.

4.5 Comparison with SOTA

A comparison with state-of-the-art techniques has been performed in Table 3, where we have selected techniques using the same data from the last 7 years (2016–2021) and have used various feature extraction and prediction techniques from one-stream to four-stream models [24, 42, 55–57]. Most of the techniques in them are using on action features and optical flow or other action features that are not robust. Meanwhile, we have considered action features extracted with CBAM, as well as contextual features to create context for action, achieving the highest accuracy with marginal improvements of up to 2.5% from the state-of-the-art technique. Due to the high performance, we will recommend our proposed model to be used in a real environment in the IR spectrum.

Table 3. Comparison of the proposed framework with state-of-the-art techniques on the InfAR dataset

Method	Year	Dataset	AP (%)
HOF [24]	2016	InfAR	68.58
Two-stream [24]	2016	-	76.66
CDFAG [56]	2018	-	78.55
TSTDDs [42]	2018	-	79.25
Four-stream CNN [57]	2019	-	83.50
SCA [55]	2021	-	84.25
Ours	-	-	86.75

4.6 Qualitative Analysis

To conduct a comprehensive evaluation of the efficacy of the proposed model for AR on both datasets, the NTU-RGB+D and InfAR predicted the test data of various classes. For visualization purposes, in Fig. 10, Hug and Shaking hands are selected from the NTU dataset, while from the InfAR, Jump, Jog, and Wave2 are chosen due to their overlapping with other classes and showcasing diverse examples. The prediction results of the model are shown in Fig. 9, where we selected samples from each class. On the NTU dataset, the proposed model accurately predicted the Hug class. However, the model sometimes exhibited confusion between the Shaking hands and High-five classes. This confusion could be attributed to the similarities in the visual appearance and motion patterns between the two classes, which can pose a challenge to the model's ability to distinguish between them. On the InfAR dataset, the proposed model achieved high accuracy in predicting the Jump and Jog classes. However, it occasionally predicted the Wave1 instead of the Wave2 class, indicating some confusion between these two classes. This confusion could be due to the similarities in motion patterns between them, which can make it challenging for the model to differentiate accurately. Overall, the proposed model exhibited high accuracy in AR on both datasets, but it had some limitations in differentiating between classes with similar motion patterns or visual appearance. Further research and model fine-tuning could potentially overcome these limitations and improve the model's performance.

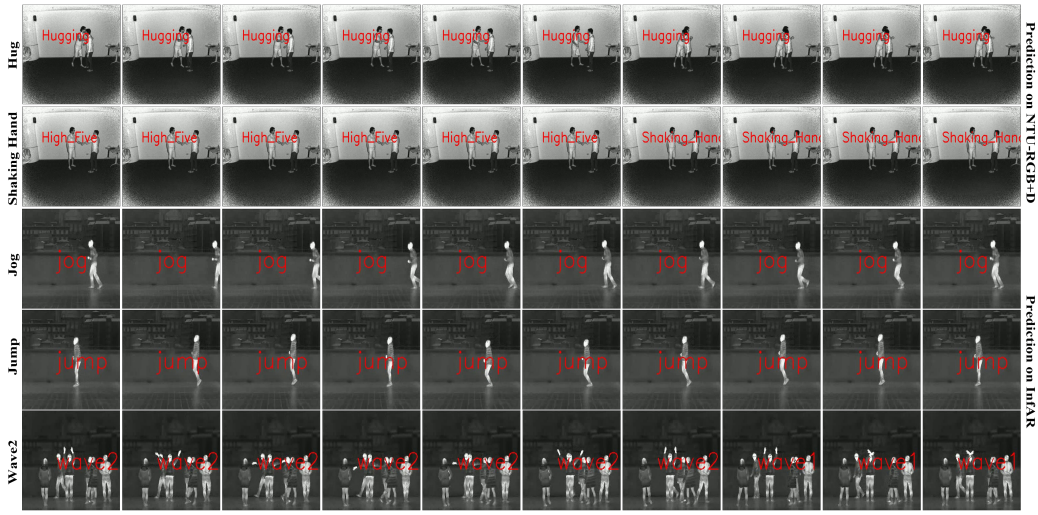

Fig. 10. Prediction performance of the proposed model on various classes.

5. Conclusion

This article proposed a new framework for complex HAIR in the IR spectrum. Action and contextual features are extracted using two different backbones. The action information extraction module, enriched with CBAM attention, specifically focused on the action portion of the input frames. The actions and contextual modules effectively extract discriminative IR features by considering action and contextual information and combined them using two-stream BiLSTM in a compact representation, which shows considerable improvements. Furthermore, the proposed approach evaluated from various aspects, including an extensive ablation study and comparison with existing state-of-the-art techniques, highlights the satisfactory performance of the proposed framework. Experiments utilized two standard IR datasets including the InfAR and NTU-RGB+D IR portion. The proposed method achieved the highest performance in terms of accuracy, up to 86.75% on InfAR and 80.94% on the NTU-RGB+D dataset, with nearly 3% improvements in accuracy achieved compared to existing SOTA approaches.

The proposed framework leverages advanced high-weight feature extraction backbones in conjunction with a sequential BiLSTM model. However, due to the sequential nature of the model, parallelization during training, as well as inferencing are not feasible, which hinders the possibility of achieving a high real-time response. We intend to expand this work in the future by incorporating lightweight and more robust backbone networks and adopting transformer structures in place of the current sequential models, which will enable us to obtain discriminative features for both high accuracy and real-time response. Moreover, the present framework exhibits exceptional flexibility, and its generalization potential extends to the IR spectrum multi-person sequences analysis. As a result, it is poised to be an invaluable tool in various domains, including AR in various locations and big data analytics.

Author's Contributions

Conceptualization, MM, NK; Data curation, MM; Formal analysis, MM, SUK; Funding acquisition, SWB; Methodology, MM, SUK; Project administration, SWB; Software, MM, SUK; Supervision, SWB; Validation, MM, SUK; Writing_original draft, MM, NK; Writing_review and editing, MM, SUK, NK.

Funding

This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (Grant No. 2023R1A2C1005788).

Competing Interests

The authors declare that they have no competing interests.

References

[1] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5281-5292, 2022. https://doi.org/10.1109/TCSVT.2022.3142771

[2] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer LSTM," *IEEE Transactions on Industrial Electronics,* vol. 66, no. 12, pp. 9692-9702, 2019. https://doi.org/10.1109/TIE.2018.2881943

[3] S. Ul Amin, M. Ullah, M. Sajjad, F. A. Cheikh, M. Hijji, A. Hijji, and K. Muhammad, "EADN: an efficient deep learning model for anomaly detection in videos," *Mathematics*, vol. 10, no. 9, article no. 1555, 2022. https://doi.org/10.3390/math10091555

[4] R. Quintero Minguez, I. Parra Alonso, D. Fernandez-Llorca, and M. A. Sotelo, "Pedestrian path, pose, and intention prediction through gaussian process dynamical models and pedestrian activity recognition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1803-1814, 2019. https://doi.org/10.1109/TITS.2018.2836305

[5] S. Sudhakaran, S. Escalera, and O. Lanz, "LSTA: long short-term attention for egocentric action recognition," in *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 9954-9963. https://doi.org/10.1109/CVPR.2019.01019

[6] N. Hussain, M. A. Khan, S. Kadry, U. Tariq, R. R. Mostaf, J. I. Choi, and Y. Nam, "Intelligent deep learning and improved whale optimization algorithm based framework for object recognition," *Human-centric Computing and Information Sciences*, vol. 11, article no. 34, 2021. https://doi.org/10.22967/HCIS.2021.11.034

[7] H. Khan, M. Ullah, F. Al-Machot, F. A. Cheikh, and M. Sajjad, "Deep learning based speech emotion recognition for Parkinson patient," *Electronic Imaging,* vol. 35, article no. IPAS-298, 2023. https://doi.org/10.2352/EI.2023.35.9.IPAS-298

[8] S. U. Amin, A. Hussain, B. Kim, and S. Seo, "Deep learning based active learning technique for data annotation and improve the overall performance of classification models," *Expert Systems with Applications*, vol. 228, article no. 120391, 2023. https://doi.org/10.1016/j.eswa.2023.120391

[9] K. Corona, K. Osterdahl, R. Collins, and A. Hoogs, "MEVA: a large-scale multiview, multimodal video dataset for activity detection," in *Proceedings of 2021 IEEE Winter Conference on Applications of Computer Vision (WACV),* Waikoloa, HI, USA, 2021, pp. 1059-1067. https://doi.org/10.1109/WACV48630.2021.00110

[10] J. Yang, J. Fan, Y. Wang, Y. Wang, W. Gan, L. Liu, and W. Wu, "Hierarchical feature embedding for attribute recognition," in *Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 13052-13061. https://doi.org/10.1109/CVPR42600.2020.01307

[11] J. Weng, C. Weng, J. Yuan, and Z. Liu, "Discriminative spatio-temporal pattern discovery for 3D action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 4, pp. 1077-1089, 2019. https://doi.org/10.1109/TCSVT.2018.2818151

[12] A. Hussain, S. U. Khan, N. Khan, I. Rida, M. Alharbi, and S. W. Baik, "Low-light aware framework for human activity recognition via optimized dual stream parallel network," *Alexandria Engineering Journal*, vol. 74, pp. 569-583, 2023. https://doi.org/10.1016/j.aej.2023.05.050

[13] W. Ahmad, M. Munsif, H. Ullah, M. Ullah, A. A. Alsuwailem, A. K. J. Saudagar, K. Muhammad, and M. Sajjad, "Optimized deep learning-based cricket activity focused network and medium scale benchmark," *Alexandria Engineering Journal*, vol. 73, pp. 771-779, 2023. https://doi.org/10.1016/j.aej.2023.04.062

[14] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: towards good practices for deep action recognition," in *Computer Vision – ECCV 2016*. Cham: Springer, 2016. pp. 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[15] S. Asghari-Esfeden, M. Sznaier, and O. Camps, "Dynamic motion representation for human action recognition," in *Proceedings of 2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Snowmass, CO, USA, 2020, pp. 546-555. https://doi.org/10.1109/WACV45572.2020.9093500

[16] S. Lamghari, G. A. Bilodeau, and N. Saunier, "ActAR: actor-driven pose embeddings for video action recognition," in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* New Orleans, LA, USA, 2022, pp. 398-407. https://doi.org/10.1109/CVPRW56347.2022.00055

[17] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, "Current advances and future perspectives of image fusion: a comprehensive review," *Information Fusion*, vol. 90, pp. 185-217, 2023. https://doi.org/10.1016/j.inffus.2022.09.019

[18] Y. Cao, Q. Tang, X. Wu, and X. Lu, "EFFNet: enhanced feature foreground network for video smoke source prediction and detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1820-1833, 2022. https://doi.org/10.1109/TCSVT.2021.3083112

[19] Z. Jiang, V. Rozgic, and S. Adali, "Learning spatiotemporal features for infrared action recognition with 3D convolutional neural networks," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Honolulu, HI, USA, 2017, pp. 309-317. https://doi.org/10.1109/CVPRW.2017.44

[20] A. Rezaei, M. C. Stevens, A. Argha, A. Mascheroni, A. Puiatti, and N. H. Lovell, "An unobtrusive human activity recognition system using low resolution thermal sensors, machine and deep learning," *IEEE*

Transactions on Biomedical Engineering, vol. 70, no. 1, pp. 115-124, 2023. https://doi.org/10.1109/TBME.2022.3186313

[21] M. Ding, Y. Ding, L. Wei, Y. Xu, and Y. Cao, "Individual surveillance around parked aircraft at nighttime: thermal infrared vision-based human action recognition," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 2, pp. 1084-1094, 2023. https://doi.org/10.1109/TSMC.2022.3192017

[22] X. Sui, S. Li, X. Geng, Y. Wu, X. Xu, Y. Liu, R. Goh, and H. Zhu, "CRAFT: cross-attentional flow transformer for robust optical flow," in *Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* New Orleans, LA, USA, 2022, pp. 17581-17590. https://doi.org/10.1109/CVPR52688.2022.01708

[23] Y. Ji, Y. Yang, F. Shen, H. T. Shen, and W. S. Zheng, "Arbitrary-view human action recognition: a varying-view RGB-D action dataset," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 1, pp. 289-300, 2021. https://doi.org/10.1109/TCSVT.2020.2975845

[24] C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, and A. G. Hauptmann, "InfAR dataset: infrared action recognition at different times," *Neurocomputing*, vol. 212, pp. 36-47, 2016. https://doi.org/10.1016/j.neucom.2016.05.094

[25] J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan, and A. C. Kot, "NTU RGB+ D 120: a large-scale benchmark for 3D human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2684-2701, 2020. https://doi.org/10.1109/TPAMI.2019.2916873

[26] S. A. W. Talha, M. Hammouche, E. Ghorbel, A. Fleury, and S. Ambellouis, "Features and classification schemes for view-invariant and real-time human action recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 4, pp. 894-902, 2018. https://doi.org/10.1109/TCDS.2018.2844279

[27] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of 2013 IEEE International Conference on Computer Vision*, Sydney, NSW, Australia, 2013, pp. 3551-3558. https://doi.org/10.1109/ICCV.2013.441

[28] Q. Gao, J. Liu, Z. Ju, and X. Zhang, "Dual-hand detection for human–robot interaction by a parallel network based on hand detection and body pose estimation," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9663-9672, 2019. https://doi.org/10.1109/TIE.2019.2898624

[29] Q. Gao, J. Liu, and Z. Ju, "Robust real-time hand detection and localization for space human–robot interaction based on deep learning," *Neurocomputing,* vol. 390, pp. 198-206, 2020. https://doi.org/10.1016/j.neucom.2019.02.066

[30] S. Wang, L. Huang, D. Jiang, Y. Sun, G. Jiang, J. Li, et al., "Improved multi-stream convolutional block attention module for sEMG-based gesture recognition," *Frontiers in Bioengineering and Biotechnology*, vol. 10, article no. 909023, 2022. https://doi.org/10.3389/fbioe.2022.909023

[31] J. Zhang, W. Xie, C. Wang, R. Tu, and Z. Tu, "Graph-aware transformer for skeleton-based action recognition," *The Visual Computer*, vol. 39, no. 10, pp. 4501-4512, 2023. https://doi.org/10.1007/s00371-022-02603-1

[32] Y. Zhang, Z. Gao, X. Wang, and Q. Liu, "Image representations of numerical simulations for training neural networks," *Computer Modeling in Engineering & Sciences*, vol. 134, no. 2, pp. 821-833, 2023. https://doi.org/10.32604/cmes.2022.022088

[33] J. Yun, D. Jiang, Y. Liu, Y. Sun, B. Tao, J. Kong, et al., "Real-time target detection method based on lightweight convolutional neural network," *Frontiers in Bioengineering and Biotechnology,* vol. 10, article no. 861286, 2022. https://doi.org/10.3389/fbioe.2022.861286

[34] Y. Zhang, Z. Gao, X. Wang, and Q. Liu, "Predicting the pore-pressure and temperature of fire-loaded concrete by a hybrid neural network," *International Journal of Computational Methods*, vol. 19, no. 08, article no. 2142011, 2022. https://doi.org/10.1142/S0219876221420111

[35] M. Wozniak, M. Wieczorek, and J. Silka, "BiLSTM deep neural network model for imbalanced medical data of IoT systems," *Future Generation Computer Systems*, vol. 141, pp. 489-499, 2023. https://doi.org/10.1016/j.future.2022.12.004

[36] H. Basak, R. Kundu, P. K. Singh, M. F. Ijaz, M. Wozniak, and R. Sarkar, "A union of deep learning and swarm-based optimization for 3D human action recognition," *Scientific Reports*, vol. 12, no. 1, article no. 5494, 2022. https://doi.org/10.1038/s41598-022-09293-8

[37] G. Yan and M. Wozniak, "Accurate key frame extraction algorithm of video action for aerobics online teaching," *Mobile Networks and Applications*, vol. 27, no. 3, pp. 1252-1261, 2022. https://doi.org/10.1007/s11036-022-01939-1

[38] A. H. Pavan, P. Anvitha, A. P. Sai, I. Sunil, Y. Maruthi, and V. Radhesyam, "Human action recognition in videos using deep neural network," in *Evolution in Signal Processing and Telecommunication Networks*. Singapore: Springer, 2022. pp. 335-341. https://doi.org/10.1007/978-981-16-8554-5_31

[39] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, 2014. pp. 568-576.

[40] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV),* Santiago, Chile, 2015, pp. 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[41] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, USA, 2017, pp. 4724-4733. https://doi.org/10.1109/CVPR.2017.502

[42] Y. Liu, Z. Lu, J. Li, T. Yang, and C. Yao, "Global temporal representation based CNNs for infrared action recognition," *IEEE Signal Processing Letters*, vol. 25, no. 6, pp. 848-852, 2018. https://doi.org/10.1109/LSP.2018.2823910

[43] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream LSTM: a deep fusion framework for human action recognition," in *Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision (WACV),* Santa Rosa, CA, USA, 2017, pp. 177-186. https://doi.org/10.1109/WACV.2017.27

[44] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1347-1360, 2018. https://doi.org/10.1109/TIP.2017.2778563

[45] L. Zhu, H. Fan, Y. Luo, M. Xu, and Y. Yang, "Temporal cross-layer correlation mining for action recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 668-676, 2022. https://doi.org/10.1109/TMM.2021.3057503

[46] C. Zhao, F. Zhou, K. Lu, S. Yang, B. Tan, W. Sun, L. Shangguan, H. Y. Wang, and Y. Liu, "Near-infrared fluorescent probe for in vivo monitoring acetylcholinesterase activity," *Sensors and Actuators B: Chemical*, vol. 360, article no. 131647, 2022. https://doi.org/10.1016/j.snb.2022.131647

[47] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3200-3225, 2023. https://doi.org/10.1109/TPAMI.2022.3183112

[48] M. Tan and Q. V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning (ICMN)*, Long Beach, CA, USA, 2019, pp. 6105-6114.

[49] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018*. Cham, Switzerland: Springer, 2018. pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[51] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," 2015 [Online]. Available: https://arxiv.org/abs/1508.01991.

[52] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Deanet al., "TensorFlow: a system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI),* Savannah, GA, USA, 2016, pp. 265-283.

[53] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[54] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014 [Online]. Available: https://arxiv.org/abs/1412.6980.

[55] X. Chen, C. Gao, C. Li, Y. Yang, and D. Meng, "Infrared action detection in the dark via cross-stream attention mechanism," *IEEE Transactions on Multimedia*, vol. 24, pp. 288-300, 2022. https://doi.org/10.1109/TMM.2021.3050069

[56] Y. Liu, Z. Lu, J. Li, C. Yao, and Y. Deng, "Transferable feature representation for visible-to-infrared cross-dataset human action recognition," *Complexity*, vol. 2018, article no. 5345241, 2018. https://doi.org/10.1155/2018/5345241

[57] J. Imran and B. Raman, "Deep residual infrared action recognition by integrating local and global spatio-temporal cues," *Infrared Physics & Technology*, vol. 102, article no. 103014, 2019. https://doi.org/10.1016/j.infrared.2019.103014