IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 13, 2022, accepted April 25, 2022, date of publication April 29, 2022, date of current version May 6, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3171263

Ensembled Transfer Learning Based Multichannel Attention Networks for Human Activity Recognition in Still Images

KOKI HIROOKA[1], MD. AL MEHEDI HASAN[1,2], JUNGPIL SHIN[1], (Senior Member, IEEE), AND AZMAIN YAKIN SRIZON[2]

[1]School of Computer Science and Engineering, The University of Aizu, Aizuwakamatsu 965-8580, Japan
[2]Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology (RUET), Rajshahi 6204, Bangladesh

Corresponding author: Jungpil Shin (jpshin@u-aizu.ac.jp)

This work was supported by the Competitive Research Fund of The University of Aizu, Japan.

ABSTRACT Human activity recognition is one of the most difficult tasks in computer vision. Due to the lack of time information, detecting human activities from still photos is more difficult than sensor-based or video-based techniques. Recently, various deep learning based solutions are being proposed one after another, and their performance is constantly improving. In this paper, we proposed a convolutional neural architecture by ensembling transfer learning based multi-channel attention networks. Here, four CNN branches were used to make feature fusion based ensembling and in each branch, an attention module was used to extract the contextual information from the feature map produced by existing pre-trained models. Finally, the extracted feature maps from four branches were concatenated and fed to fully connected network to produce the final recognition output. We considered 3 different datasets, Stanford 40 actions, BU-101 and Willow human actions datasets to evaluate our system. Experimental analysis showed that the proposed ensembled convolutional architecture outperformed previous works by a noteworthy margin.

INDEX TERMS Human activity recognition, multi-channel attention module, ensembling, InceptionV3, Xception, InceptionResnetV2, EfficientNetB7.

I. INTRODUCTION

In Human Activity Recognition (HAR), an action refers to an entity which can be observed by utilizing either human eye or a sensing device [1]. For example, walking is an action and it necessitates continual observation of a person in the field of view. Human activities can be classified into four broad groups based on the body components that are being used for action i.e., gesture, action, interaction and group activity [2]. Gestures involves face, hands or other body part movement for attaining nonverbal communication. Action refers to movements of human i.e., running, walking, jumping, crawling etc. Interaction denotes the actions between a human and an object or another human. Group activity refers to the cases where multiple persons conduct gestures, actions and interactions with diverse objects.

HAR has been an active area of research in computer vision and pattern recognition in recent years and has become a hot scientific topic in the computer vision community. It is involved in the development of many important applications such as human computer interaction (HCI) [3], [4], education [5], medical [6], entertainment [7], virtual reality [7], video surveillance and home monitoring [8], [9], security [10], video retrieval [11], abnormality identification [12] and so on. Therefore, the wide range of the activity recognition methods is directly linked to the application domain to which they are implemented [9].

HAR can be roughly categorized into two types based on the type of input data: sensor-based HAR and vision-based HAR [13], [14]. Furthermore, video-based HAR and image-based HAR are two different forms of vision-based HAR. Sensor-based HAR looks at raw data from biosensors and remote monitoring [15], [16], whereas vision-based HAR looks at pictures or videos captured by optical components [17], [18]. Since they are worn by people to automatically detect and track several actions such as sitting, walking, jumping, and relaxing, wearable devices are exemplary instances of sensor-based HAR [19], [20].

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

A sensor, however, will not operate if a person is either outside of its range [21] or doing unexplained activities [22]. CCTV systems, on the other hand, have long been used in vision-based HAR systems [17]. The identification of gestures and actions based on video analysis has been extensively researched [23], [24]. Furthermore, this issue is particularly beneficial to video surveillance [25], [26] and interactive media [27], [28]. Because vision-based data is less expensive and easier to acquire than sensor-based data, the great amount of research has concentrated on vision-based HAR in recent years. That's why, for this work, human activity recognition from still images has been considered which can be used for surveillance, robotic applications, human-computer interaction applications, annotating images using verbs, searching an image database using verbs, searching images online based on action queries, frame tagging, and so on in the domain of education, learning, and industry.

Video-based HARs, in particular, have been widely examined over the past decade, with excellent outcomes in every case [29]. This is because each video in the video-based technique is made up of numerous frames, each of which contains information on the subject's movement while also keeping track of time. Due to the absence of sufficient spatiotemporal information, image-based HARs are particularly complicated and difficult tasks compared to video-based HARs and sensor-based HARs. To put it another way, it is important to comprehend and distinguish human activities from a single visual in image-based HARs. Because the datasets and the number of pictures per label in each dataset are not large enough, many researchers haven't looked into this specific behavioral perception domain.

Machine learning methods have been used to tackle the HAR issue for decades, including random forest [30], Bayesian networks [31], Markov models [32], [33], and support vector machine [34], [35]. Traditional machine learning techniques have performed well in heavily controlled conditions with minimal input data. They do, however, need many pre-processing stages and rigorous hand-engineering, which is incredibly difficult and time-consuming [36]. Furthermore, the usage of shallow features results in poor performance utilizing unsupervised learning [36], [37]. In the realm of image recognition, however, emphasis has been concentrated on the construction of convolutional neural networks (CNN) and the creation of recipes [38]–[40] since AlexNet's debut in 2012. Convolutional neural networks, such as VGG [41] and Inception [42], have made it feasible to extract strong features from raw photos, resulting in great results and still improving recognition accuracy in large datasets. Despite CNN's significant performance, the absence of large volumes of labeled data in action recognition creates overfitting issues in deep CNN training [43].

After rigorous study of previous works, we discovered that the performance of human activity recognition can be boosted significantly by designing a new deep learning architecture. In this article, we used feature fusion-based ensembling technique to concatenate features that were produced from four CNN branches. Firstly, transfer learning [44] has been adopted using four pre-trained deep convolutional neural networks with ImageNet weights to overcome the problem of limited samples per class [45]. Therefore, the initial ImageNet weights were retrained for the considered datasets for each of the four pre-trained models. Secondly, by incorporating an attention mechanism, the extracted feature map was transformed into a more discriminative feature map. Thirdly, for each of the 4 models, after using a channel attention module, we added fully connected layers having dense layers with the ''SoftMax'' activation function. Here, SoftMax activation is producing a probabilistic confidence feature map that can be thought of as the feature map. Finally, we have combined the feature maps extracted from multiple channels (four channels) to perform feature fusion-based ensembled learning. Experimental analysis revealed that ensembling the final extracted feature map of four paths can boost the performance significantly. Details about the proposed methodology can be discovered in the ''Proposed Architecture'' subsection.

II. RELATED WORKS

As previously noted, a great number of experiments for both sensor-based and video-based HAR have been presented in the past. Because of the complexity of image-based HAR, little study has been done in this area. The bag-of-words (BoW) framework [46] is the standard technique to action recognition in still photos, and it is the most popular framework. A standard bag-of-words architecture collects characteristics from the whole image and encodes them as histogram representations. However, if the image comprises backgrounds or objects that are unrelated to the action categorization, noise-prone features will always be present, lowering classification accuracy. Previously, [47] used a bag-of-words and a support vector machine classifier to classify human activities. [48] suggested a single picture action recognition system based on semantic component actions. They proposed that, unlike previous part-based approaches, a mid-level semantic part action exists, and that human action is a mix of semantic part actions and context clues. They separated the body into seven sections (head, torso, two arms, two hands, and lower body) and utilized partial actions to predict the full body's action.

[49] presented an unsupervised learning of a finite multivariate generalized Gaussian mixture model to recognize human actions. They focused on the estimate of the mixture model's parameters for a complete covariance matrix, which is a crucial cue in finite mixture models. They created a new learning technique that combines a fixed-point covariance matrix estimator with an expectation-maximization approach. [50] suggested that the appearance of inconsequential items and backdrops can readily be misinterpreted. They solved the problem by employing a bounding box for the target individual to extract only human features, although this method is inefficient since the

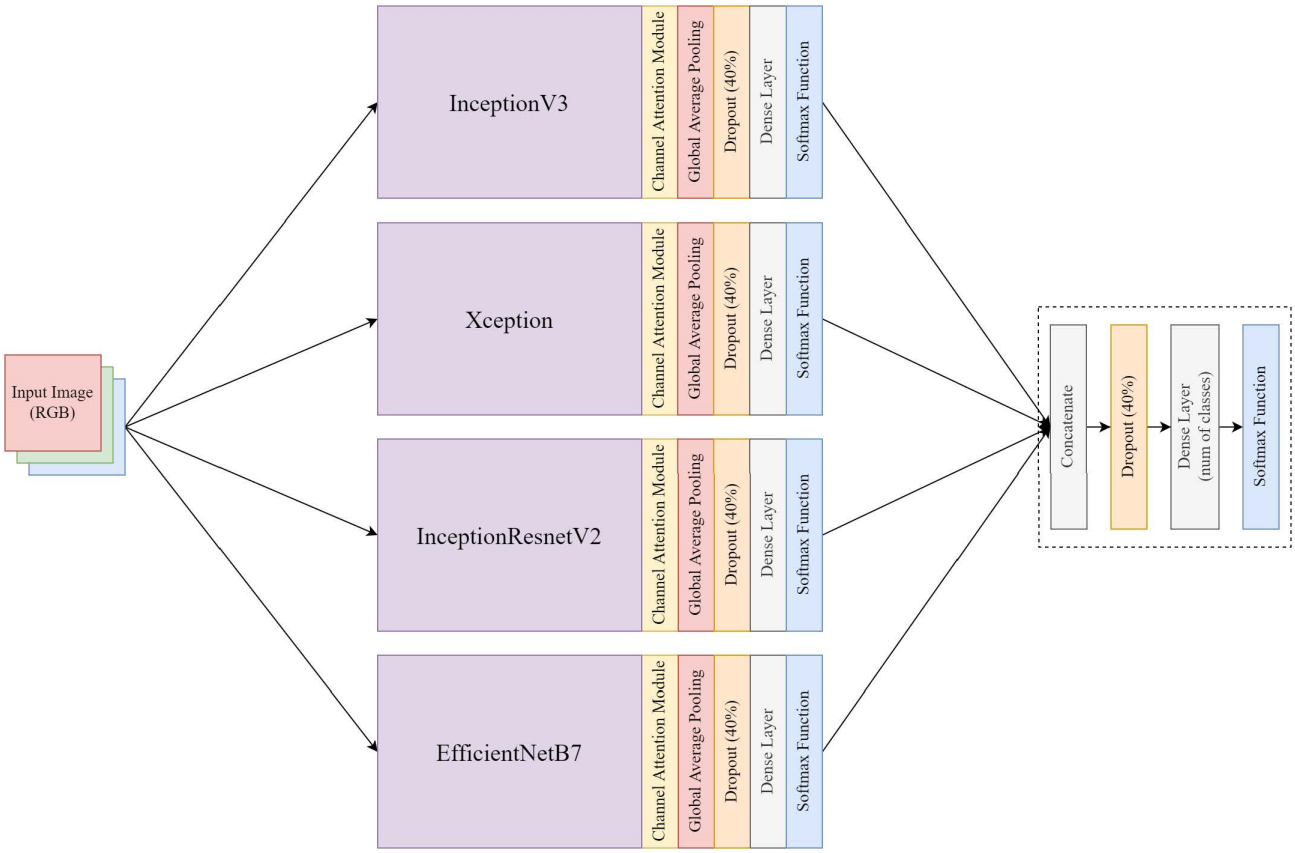


FIGURE 1. Proposed ensembled architecture consisting of pre-trained CNNs, attention module, fully connected layers and ensembling of feature maps.

bounding box may not be accessible. Individual mask loss was also implemented to automatically route feature map activation to the target human doing the action, eliminating misleading context activation. The fundamental challenge or difficulty with still image-based HAR, according to experiments done by [51], is the lack of temporal information. They presented a unique visual representation that captures the subject's look and predicts the actors' future movement patterns to solve this challenge. To add novel spatial-temporal CNNs, they adopted a transfer learning-based technique.

After the introduction of the multibranch attention networks, the performance of the human activity recognition was boosted by a big margin [52]. However, the channel attention module utilized there was not matured enough to beat the previously attained highest performance of [48]. A more recent work suggested some critical updates in the channel attention module that resulted in about 1.97% performance boost for the Stanford 40 actions dataset [53]. The critical changes were the usage of two consecutive dense layers with ReLU and sigmoid functions respectively. However, there were two major concerns in this attention module. Firstly, due to using two consecutive dense layers, the module may face overfitting problem as well as the vanishing gradient problem [54]. Secondly, sigmoid activations are easier to saturate. There is a comparatively narrow interval of inputs where the sigmoid's derivative is sufficiently nonzero [55]. In other words, once a sigmoid hits either the left or right plateau, a backward pass through it is practically pointless because the derivative is very close to 0.

In our work, we introduced a batch normalization layer between two dense layers of the attention module to prevent the overfitting and vanishing gradient problem [56]. Furthermore, instead of using sigmoid activation, we used the ReLU activation function as in practice, networks with ReLU perform better than sigmoid networks in terms of convergence [38]. Sparsity and a reduced likelihood of vanishing gradient are two other major advantages of ReLUs [55]. Moreover, we introduced four pre-trained models, fully connected layers with softmax activation and ensembling of the feature maps along with the channel attention module. All these elements of our proposed architecture has been described in details in the ''Proposed Architecture'' subsection.

III. MATERIALS AND METHODS

A. DATASET DESCRIPTION

In this research, three datasets have been considered: the Stanford 40 actions dataset, the BU-101 dataset and the Willow dataset. In this subsection, all of these datasets have been discussed in detail.

1) STANFORD 40 ACTIONS DATASETS

The Stanford 40 Action dataset [57] contains images of humans performing 40 actions. There are 9532 images in total, with 180-300 images per action class. Since the Stanford 40 Actions dataset is divided into the train set and the test set and provided by the author in this configuration, the experiment was conducted without changing the mentioned configuration.

2) BU-101 DATASET

The BU-101 action dataset [58] contains approximately 23.8K action images. BU-101 contains at least 100 images for each action. The actions in this dataset are divided into five categories: human-object interaction (HOI), body-motion only (BMO), playing musical instruments (PMI), human-human interaction (HHI), and sports (SPT). Since this dataset is not divided into train and test sets, it was randomly divided into train and test sets by keeping 80% data in the train set and the rest of the 20% data in the test set. In total, there are 20 human-object recognition, 16 body-motion only, 10 playing musical instruments, 5 human-human interaction, and 50 sports classes in the overall BU-101 dataset.

3) WILLOW DATASET

The Willow action dataset [59] contains 911 images split into seven action categories: interacting with computer (Cat. 1), photographing (Cat. 2), playing music (Cat. 3), riding bike (Cat. 4), riding horse (Cat. 5), running (Cat. 6) and walking (Cat. 7). This dataset is labeled person by person, and that information has recorded in the annotations that accompany the dataset. Also, the division into the train set, validation set and test set is assigned for each cropped image. Therefore, we trimmed each image one by one based on the attached annotation, and then split it based on the split information provided by the author. The experiment was conducted without changing the configuration.

B. PROPOSED ARCHITECTURE

In this article, an ensembled multi-channel convolutional neural architecture has been proposed to recognize the human activities. While designing the architecture, first, four pre-trained convolutional neural networks were introduced. Input images were fed into each of these pre-trained convolutional neural architectures for discovering important features.

Secondly, after applying each of the pre-trained architectures, a channel attention module was utilized that has been previously proposed by squeeze and excitation networks [60]. The attention module can adaptively set weights to the channels of pre-trained feature maps to choose more strong features for using them in the classification section. The channel attention module catches the image's global context due to its vast effective field of view, and therefore, class-specific information can be discovered.

After applying the channel attention module to the output feature maps of the four pre-trained models, four new feature maps were generated. Thirdly, these four new feature maps were then passed to four fully connected layers to obtain the final feature maps. Finally, feature fusion was practiced by concatenating the final feature maps produced by the fully connected layers. After ensembling the features, another fully connected layer was added.

Experimental analysis showed that the proposed architecture had boosted the performance by a significant margin. Figure 1 illustrates the basic model structure. More details on how each module of the proposed convolutional neural architecture is designed have been discussed in later subsections.

1) PRE-TRAINED CNN

Transfer learning, often known as pre-training, is a machine learning approach for efficiently finding an effective hypothesis by transferring information learned in another activity. It is a technique where pre-trained models with previously trained weights i.e., ImageNet weights are utilized to retrain the models on another dataset [61]. By employing these strategies, it is feasible to shorten the time necessary for learning without using a huge quantity of data. Therefore, it is ideal for the applications of image-based human activity recognition with small datasets with a few images per class.

Four transfer-learned convolutional neural architectures were utilized in the pre-training section. These were InceptionV3 [62], Xception [63], InceptionResnetV2 [64] and EfficientNetB7 [65]. These four pre-trained models have their unique behavioral characteristic of extracting valuable discriminating features. The TensorFlow library has been utilized to implement these models while keeping the input shape at $512 \times 512 \times 3$. As mentioned earlier, the fully connected layers were intentionally removed by keeping the 'include top' feature as 'False'. These layers were swapped with the channel attention module and fully connected layers module that are described in the next subsections.

It should be noted that the pre-trained models were retrained on the considered datasets for this research separately and the learned weights were saved. These saved weights were then loaded while training the whole architecture while feature fusion-based ensembled learning and the weights were set to non-trainable (frozen). This process was adapted due to heavy memory requirements as training four pre-trained models at the same time needs tremendous weight update. Furthermore, by keeping the trained weights non-trainable, our architecture allowed the four pre-trained models to preserve their internal behavioral characteristics of extracting valuable features. Therefore, none of the pre-trained models were getting affected by one another while feature fusion based ensembling.

2) ATTENTION MODULE

As mentioned earlier, the pre-trained CNN models are capable of pulling strong properties that vary from one sample to another sample through several filters. However,

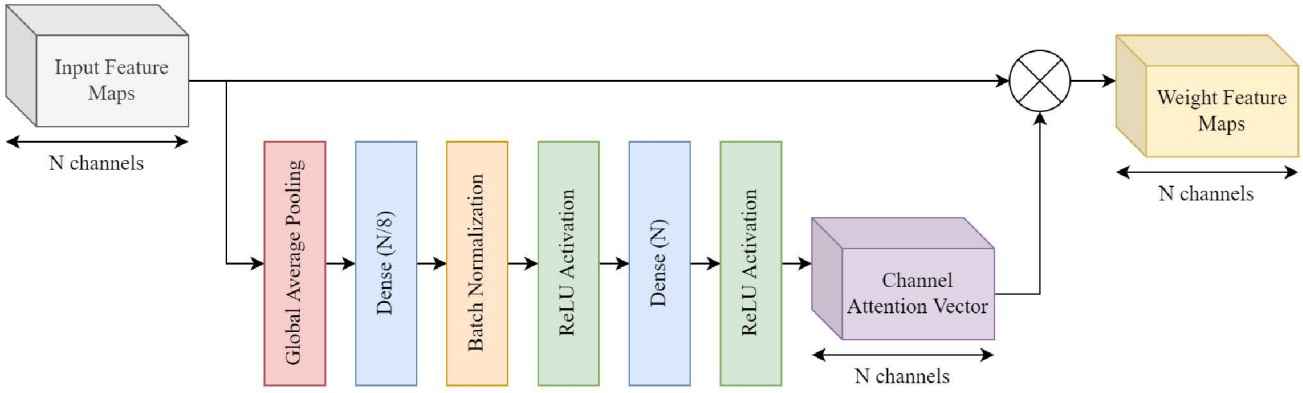


FIGURE 2. The channel attention module utilized in this research.

it has also been demonstrated that adding an attention module after pre-training results in the selection of more prominent features in a classification problem [52]. The channel attention module takes the extracted feature map, runs it through global average pooling, and outputs each channel's feature amount. Then, for each channel, two convolution layers are coupled, and ReLU activation function is utilized to produce a value of 0 or positive.

A more powerful feature map can be constructed by multiplying the result of the activation function by the pre-trained feature map. In other words, the channel attention module outputs positive value for significant features or channels and 0 for unnoticeable features or channels. To put it another way, when it is increased by stronger characteristics, remarkable features persist while unimportant features becomes zero. Two additional advantages of ReLU activation are sparsity and a reduced likelihood of vanishing gradient.

The attention module utilized in this research started with global average pooling of input feature maps from N channels. After that, a dense layer of size N/8 was added, followed by a batch normalization layer. The batch normalization layer was utilized here to solve the internal covariate shift problem. Batch normalization also prevented the gradient vanishing problem as it prevents the gradients from becoming too small. ReLU activation was utilized here. Next, another dense layer of size N was added with the ReLU activation again. The computational complexity of ReLUs is substantially lower than that of sigmoid. When dealing with large networks with many neurons, this benefit is enormous, and it can significantly shorten both training and evaluation times. Furthermore, the ReLU-trained model converges quickly. Figure 2 illustrates the proposed channel attention module that was employed in this experiment.

3) FULLY CONNECTED LAYERS

After applying the channel attention modules on the outputs of the pre-trained models, global average pooling was used and by doing so, the channel attention module's powerful feature maps were turned to feature vectors. Dropout layers were put before dense layers in the fully connected section to prevent overfitting. The dropout rate was maintained at 40%. The dense layers' activation function was set to softmax. Here, softmax activation is producing a probabilistic confidence map that can be thought as the feature map. The number of features generated by this probabilistic confidence map is equal to the number of classes of the corresponding dataset.

4) FEATURE FUSION BASED ENSEMBLING

Instead of using output-based ensembling, feature fusion-based ensembling [66], [67] was utilized here. In output-based ensembling, the outputs of each layer were ensembled. Some popular ways of applying outcome-based ensembling are the voting method, softmax averaging, adding a dense layer with the outputs, etc. However, in feature fusion-based ensembling, feature maps are ensembled. In our case, the four final feature maps produced by the fully connected layers module were ensembled or concatenated, followed by a dropout of 40% and a dense layer with a softmax activation function.

It should be kept in mind that softmax values had been used as feature maps here, and therefore, these softmax values were getting concatenated. The reason behind using the softmax activation function rather than ReLU here is that softmax produces probability scores. Therefore, if one feature is prominent in a softmax output, other features are bound to be non-prominent, which allowed the proposed CNN model to converge with high confidence and boosted the performance.

As mentioned earlier, first, the pre-trained models were retrained separately and while training the whole architecture, the retrained weights were loaded and the parameters of the pre-trained models were set to non-trainable. That means, the training process had two stages. Firstly, the retraining of pre-trained models were utilized to extract valuable features using their unique architectural characteristics. Next, the weights of attention module and fully connected layers were

TABLE 1. Performance of different pre-trained CNNs and proposed ensembled model on stanford 40 actions dataset.

Methods	Accuracy(%)
InceptionV3	89.90
InceptionResNetV2	90.55
Xception	89.52
EfficientNetB7	92.44
Proposed ensembled model	93.76

TABLE 2. Performance comparison of Stanford 40 actions dataset with previous works.

Methods	Accuracy(%)
Yan et al. [52]	90.70
Zhao et al. [48]	91.20
Sina et al. [53]	93.17
Proposed ensembled model	**93.76**

trained by keeping the parameters of the pre-trained model frozen. This way of training not only reduces training time and memory consumption, but also it preserves the unique architectural characteristics of each pre-trained models to figure out the best set of features among all the extracted features of the pre-trained models.

C. PERFORMANCE METRICS

Every machine learning pipeline has performance metrics. Classification performance can be measured in a variety of ways. For performance measurements in this study, we employed accuracy, precision, mean average precision, and confusion matrix. The number of accurately anticipated data points out of all the data points is known as accuracy. The number of true positives (TP) and true negatives (TN) divided by the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) is how it's defined more formally. In simpler terms, accuracy can be expressed using formula (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision, on the other hand, is a metric that measures the number of correct positive predictions. The ratio of accurately predicted positive instances divided by the total number of positive examples predicted is used to compute it. Therefore, precision can be computed using formula (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

The mean average precision (mAP), often known as AP, is a widely used metric for assessing the performance of models performing document/information retrieval and object detection tasks. It can be calculated using formula (3).

$$mAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{3}$$

where Q is the total number of queries in the set and AveP(q) denotes the average precision (AP) for a single query, q.

Finally, a confusion matrix provides a summary of prediction outcomes of categorization problem. The number of correct and wrong predictions is calculated and broken down by class using count values in a confusion matrix.

IV. EXPERIMENTATION AND RESULT ANALYSIS

A. EXPERIMENTAL SETUP

The experimental environment was UbuntuOS, and the framework used was TensorFlow 2.5.0. Data augmentation was performed by using the ImageDataGenerator provided by the Tensorflow framework to make up for the lack of images in the dataset. Specifically, normalization, image rotation with a range of 0 to 30 degrees, random vertical shift with a range of 0% to 20%, random horizontal shift with a range of 0% to 20%, and horizontal inversion were utilized for augmentation. To train the proposed model, the learning rate was set to 0.0001 and SGD [77] was utilized with a momentum coefficient of 0.9. For the Stanford 40 actions dataset and BU-101 dataset, the batch size for the train and validation sets for InceptionV3, Xception, InceptionResNetV2 was kept at 8. On the other hand, the batch size for the train and validation sets for EfficientNetB7 was kept at 4 for the mentioned dataset due to lack of GPU memory. However, for the Willow dataset, batch size was kept at 4 for train and validation sets for all the considered CNN models. These values of batch size were selected using the grid search mechanism by keeping the limited GPU power in mind.

B. RESULTS FOR STANFORD 40 ACTIONS DATASET

Table 1 illustrates two sorts of results separated by double lines. First, the performance of individual pre-trained CNNs has been reported. It should be kept in mind that while applying the pre-trained CNN, no attention module, fully connected layers, or ensembling were involved. After that, the result of the proposed ensembling has been reported which involves attention module, fully connected layers, and previously trained frozen pre-trained CNNs. It can be noticed that EfficientNetB7 achieved the best accuracy of 92.44% among the four pre-trained CNN. However, the proposed ensembled model achieved 93.76% accuracy which is approximately 1.32% better than EfficientNetB7.

Table 2 illustrates a comparison of accuracy between the proposed ensembled model and previous methods. It can be noticed that pre-trained CNNs alone couldn't outperform the previous best result of 93.17%. However, the proposed ensembled model was able to outperform the best result by 0.59% due to the integration of the previously described attention module, fully connected layers, and ensembling. Figure 3 illustrates the confusion matrix for the Stanford 40 dataset. It can be observed that for each of the 40 actions, the proposed ensembled model is working equally fine.

C. RESULTS FOR BU-101 DATASET

Table 3 illustrates results for both the pre-trained CNNs and the proposed ensembled method. While applying the pre-trained CNN, no attention module, full connected layers,

TABLE 3. Performance of different pre-trained CNNs and proposed ensembled model on BU-101 dataset.

Methods	Precision for each category (%)				
	HOI (%)	BMO(%)	PMI(%)	HHI(%)	SPT(%)
InceptionV3	95.02	92.22	94.92	92.94	96.24
InceptionResNetV2	94.22	92.97	95.50	92.21	95.53
Xception	96.15	91.86	97.41	91.65	96.04
EfficientNetB7	96.33	92.71	96.84	94.59	95.53
Proposed Ensembled Model	97.98	94.02	97.43	93.74	97.79

*Here, HOI, BMO, PMI, HHI and SPT refer to human-object interaction, body-motion only, playing musical instruments, human-human interaction and sports respectively.

TABLE 4. Performance comparison of BU-101 dataset with previous works.

Methods	Precision for each category(%)				
	HOI (%)	BMO(%)	PMI(%)	HHI(%)	SPT(%)
Safaei, M et al. (2019) [51]	61.1	84.4	58.7	71.3	74.8
Safaei, M et al. (2020) [68]	59.6	93.8	68.9	67.0	74.7
Proposed Ensembled Model	**97.98**	**94.02**	**97.43**	**93.74**	**97.79**

*Here, HOI, BMO, PMI, HHI and SPT refer to human-object interaction, body-motion only, playing musical instruments, human-human interaction and sports respectively.

TABLE 5. Performance of different pre-trained CNNs and proposed ensembled model on willow dataset.

Methods	Precision for each label(%)							mAP(%)
	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	
InceptionV3	71.11	85.92	93.64	92.75	68.92	60.61	69.47	77.49
InceptionResNetV2	90.43	83.33	93.16	90.07	72.86	69.01	68.81	79.67
Xception	79.55	92.31	98.15	94.66	57.14	64.37	70.75	79.56
EfficientNetB7	71.79	88.89	97.92	50.94	82.61	88.89	57.89	76.99
Proposed Ensembled Model	87.18	89.04	92.44	90.85	76.47	73.08	73.45	83.21

*Here, mAP refers to mean average precision.

TABLE 6. Performance comparison of willow dataset with previous works.

Methods	Precision for each label (%)							mAP(%)
	Cat. 1	Cat. 2	Cat. 3	Cat. 4	Cat. 5	Cat. 6	Cat. 7	
Delaitre et al. [47]	58.2	35.4	73.2	82.8	69.6	44.5	54.2	59.6
Delaitre et al. [69]	56.6	37.5	72.0	90.4	75.0	59.7	57.6	64.1
Sharma et al. [70]	59.7	42.6	74.6	87.8	64.2	56.1	56.5	65.9
Sharma et al. [71]	64.5	40.9	75.0	91.0	87.6	55.0	59.2	67.6
Khan et al. [72]	61.9	48.2	76.5	90.3	84.3	64.7	64.6	70.1
Khan et al. [73]	66.8	48.0	77.5	**93.8**	**87.9**	67.2	63.3	72.1
Alexe et al. [74]	67.5	47.5	72.5	90.6	86.0	59.1	61.7	69.3
Uijlings et al. [75]	67.8	48.1	77.5	92.0	85.8	61.3	63.5	70.9
Zhao et al. [76]	67.9	49.1	86.5	93.0	86.2	65.7	72.6	74.4
Proposed Ensembled Model	**87.18**	**89.04**	**92.44**	90.85	76.47	**73.08**	**73.45**	**83.21**

*Here, mAP refers to mean average precision.

or ensembling were involved. On the other hand, while applying the proposed ensembled model the values of the trainable parameters of the pre-trained CNNs were kept frozen so that characteristics of each pre-trained CNN can be preserved. Unlike Stanford 40 actions dataset, precision has been calculated for the BU-101 dataset instead of accuracy for the proper comparison with the previous work. It can be observed that among the pre-trained models, EfficientNetB7 has outperformed others in the human-object interaction (HOI) and human-human interaction (HHI) category whereas InceptionResNetV2, Xception, and Inception V3 outperformed others in body-motion only (BMO), playing musical instruments (PMI) and sports (SPT) categories respectively. However, it can be noticed that the proposed ensembled model has outperformed all the pre-trained models in all categories except HHI with a very little difference margin.

Table 4 illustrates the performance comparison of the BU-101 dataset with the previous works in terms of precision. It can be seen that the proposed ensembled method has outperformed the previous works by a significant margin. To be precise, the proposed ensembled model achieved 36.88%, 0.22%, 28.53%, 22.44%, and 22.99% better precision than previous works in HOI, BMO, PMI, HHI, and SPT categories

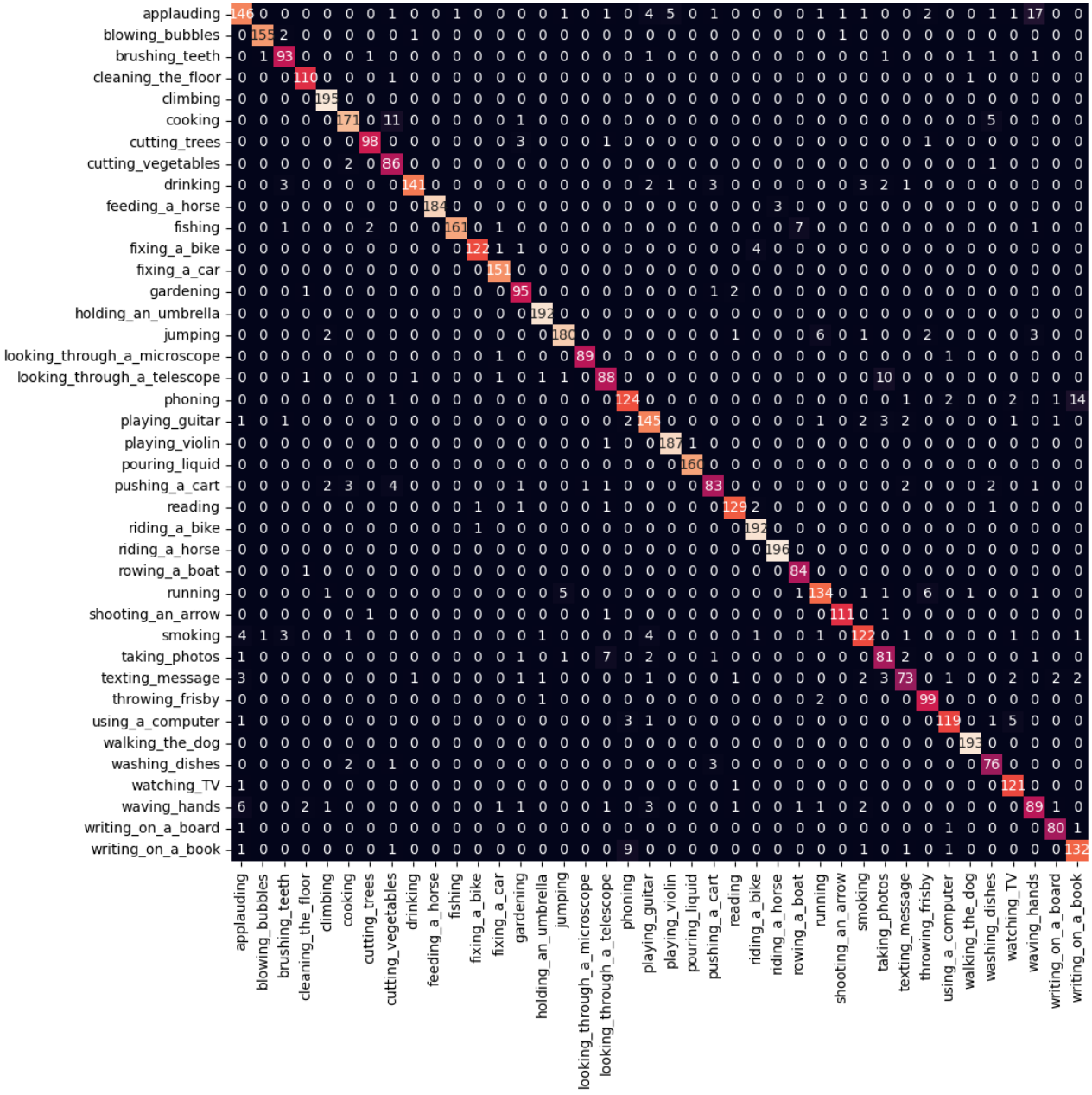


FIGURE 3. Confusion matrix for the Stanford 40 actions dataset.

respectively. There were a total of 101 classes in the BU-101 dataset, therefore, the size of the confusion matrix will be 101 × 101 which is quite large and difficult to understand. Hence, we haven't illustrated the confusion matrix here.

D. RESULTS FOR WILLOW DATASET

Table 5 illustrates the performance of different pre-trained CNNs along with the performance of the proposed ensembled model in terms of precision. As the previous works calculated the mean average precision, we also calculated the mean average precision for proper comparison. Willow dataset is a very complex dataset as there are 7 categories with a total of only 911 images. Among these 7 categories, some categories are hard to discriminate due to lack of dissimilarity among extracted features. In Table 5, it can be noticed that each of the pre-trained CNNs is struggling in recognizing some categories. InceptionV3 is struggling for category 5, 6 and 7, InceptionResNetV2 in category 6 and 7, Xception in category 5 and 6, and EfficientNetB7 in category 4 and 7. However, while applying the proposed ensembled model, it was discovered that for each of the categories, the proposed architecture was achieving decent

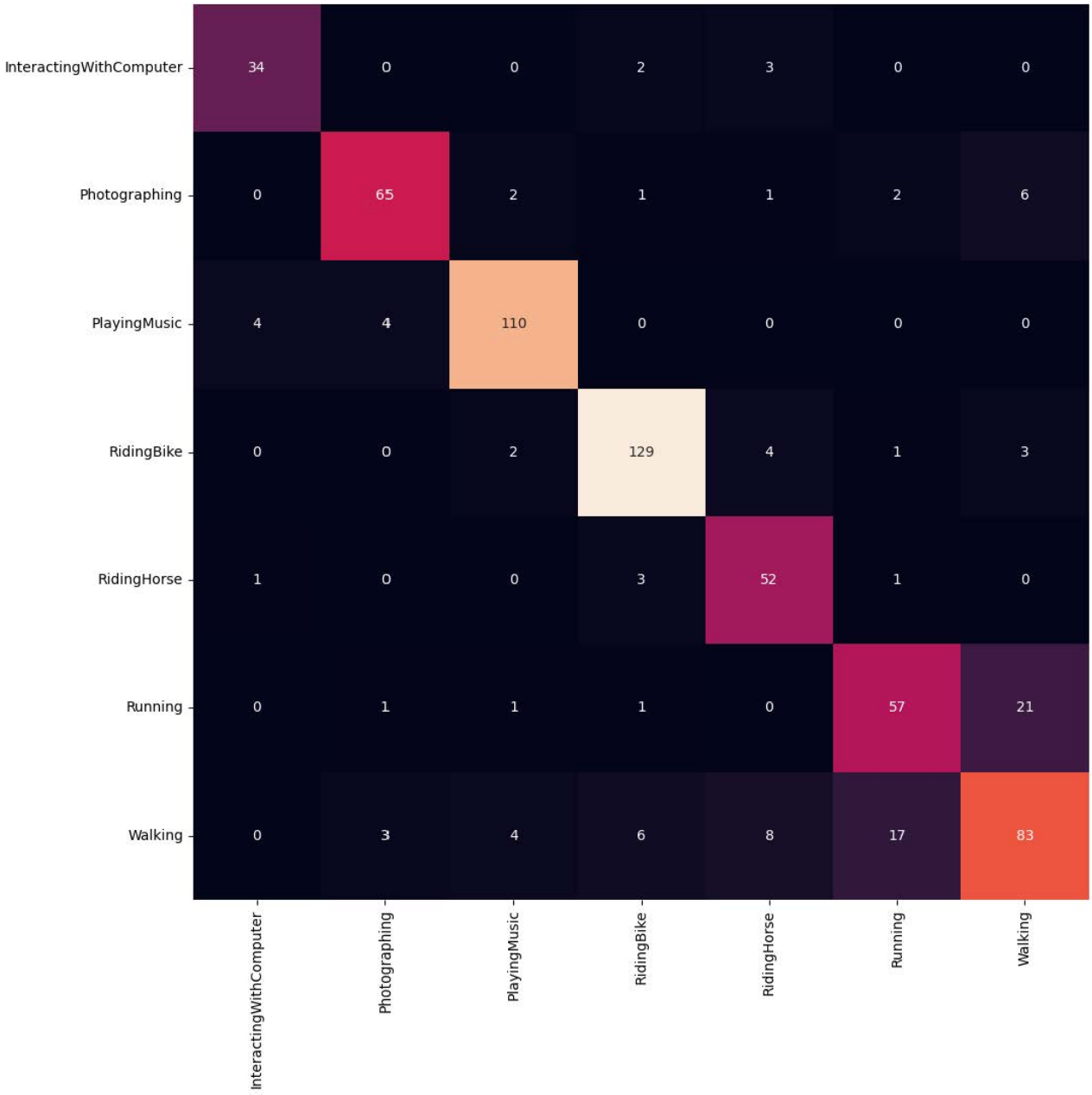


FIGURE 4. Confusion matrix for the Willow dataset.

performance. After calculating the mean average precision, it was observed that the proposed ensembled model achieved the highest mean average precision (mAP). Therefore, the proposed ensembled model is more capable of discriminating between extracted features than the single pre-trained models.

Table 6 illustrates the performance comparison of the Willow dataset with previous works in terms of mean average precision. It can be noticed that the proposed ensembled model has outperformed all the previous works in terms of mean average precision (mAP). If class-wise precision is considered, the proposed ensembled model has outperformed previous works in categories 1, 2, 3, 6, and 7. For the other 2 classes, the proposed approach achieved decent precision. However, it should be noted that for categories 1, 2, and 6 the previous highest precision values were very low i.e., 67.9, 49.1, and 67.2 respectively. In Table 6, it can be noticed that the performance has improved by a huge margin for these categories when the proposed approach was practiced. In simpler words, the proposed ensemble CNN is not affected by overfitting like the previous works and therefore, is capable of finding better discriminating features. For a more clear understanding, we have provided the confusion matrix in Figure 4.

V. DISCUSSION

After comparing the performance of our approach with the performances of previous works, as illustrated in Table 2, Table 4 and Table 6, it can be noticed that the proposed method performed favorably against all previous works. The boost in performance was obtained for four major reasons. Firstly, due to using the pre-training technique, the models needed little training time and examples to figure out the significant patterns. Secondly, using an effective attention mechanism ensures that the prominent features will survive and the less important features will be discarded. Thirdly, the fully connected layers involving softmax activation ensures that the most prominent features get highest values than others. Finally, because of the proposed ensembled feature fusion technique, the performance was boosted even further.

VI. CONCLUSION

Previously, many researchers have conducted experiments on sensor-based HARs and video-based HARs. However, the domain of experimenting on HARs involving still images has not been explored enough due to the lack of time-series data for still images. Another major problem of human activity recognition from still images is the lack of images per class and the complexities of human actions. In this article, we addressed this problem domain and proposed ensembled transfer learning-based multi-channel attention networks for human activity recognition in still images. Firstly, four pre-trained models were utilized to address the 'lack of images per class' problem. Next, an attention module was added followed by fully connected layers for each of the four pre-trained models. Finally, we applied feature fusion by ensembling or concatenating the final feature maps. While designing the attention module, we reduced the chance of occurring vanishing gradient problem, the covariate problem and sparsity which were previously undiscovered. Moreover, the well-thought integration of fully connected layers and ensembling outperformed all previous works by a noteworthy margin. In the future, there is scope for figuring out a way to obtain the skeletal joint points of human actions from still images to achieve a higher recognition rate.

In the domains of education, learning, and industry, human activity recognition from still images can be utilized for image annotation, action behavior-based image retrieval, human-computer interaction, and frame reduction in videos. Moreover, human activity recognition can be used in active and assisted living (AAL) systems for smart homes, surveillance and monitoring systems, and healthcare monitoring applications. We introduced a methodology in this paper that is capable of recognizing human activities from still images more accurately than previous significant works. We hope that with the increasing usage of this technology, our work will have a significant impact on education, learning, and industry.

REFERENCES

[1] P. Pareek and A. Thakkar, "A survey on video-based human action recognition: Recent updates, datasets, challenges, and applications," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 2259–2322, Mar. 2021.
[2] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 1–43, 2011.
[3] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Process.*, vol. 93, no. 6, pp. 1445–1457, Jun. 2013.
[4] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 2, pp. 224–241, Feb. 2011.
[5] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Inf. Fusion*, vol. 46, pp. 147–170, Mar. 2019.
[6] M. M. Hassan, S. Ullah, M. S. Hossain, and A. Alelaiwi, "An end-to-end deep learning model for human activity recognition from highly sparse body sensor data in internet of medical things environment," *J. Supercomput.*, vol. 77, no. 3, pp. 2237–2250, Mar. 2021.
[7] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, pp. 88–131, Jun. 2013.
[8] A. Poulos, C. Brown, D. McCulloch, and J. Cole, "Context-aware augmented reality object commands," U.S. Patent 9 791 921, Oct. 17, 2017.
[9] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, no. 1, pp. 1–54, Jan. 2012.
[10] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4657–4666.
[11] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Understand.*, vol. 117, no. 6, pp. 633–659, Jun. 2013.
[12] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Eng. Appl. Artif. Intell.*, vol. 77, pp. 21–45, Jan. 2019.
[13] Z. S. Abdallah, M. M. Gaber, B. Srinivasan, and S. Krishnaswamy, "Activity recognition with evolving data streams: A review," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–36, 2018.
[14] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image Vis. Comput.*, vol. 60, pp. 4–21, Apr. 2017.
[15] C. Xu, L. N. Govindarajan, and L. Cheng, "Hand action detection from ego-centric depth sequences with error-correcting Hough transform," *Pattern Recognit.*, vol. 72, pp. 494–503, Dec. 2017.
[16] J. Qi, P. Yang, M. Hanneghan, S. Tang, and B. Zhou, "A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1384–1393, Apr. 2019.
[17] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognit.*, vol. 61, pp. 295–308, Jan. 2017.
[18] X. Yang and Y. Tian, "Super normal vector for human activity recognition with depth cameras," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 1028–1039, May 2017.
[19] B. Alsinglawi, Q. V. Nguyen, U. Gunawardana, A. Maeder, and A. J. Simoff, "RFID systems in healthcare settings and activity of daily living in smart homes: A review," *E-Health Telecommun. Syst. Netw.*, vol. 6, no. 1. pp. 1–17, 2017.
[20] M. Kirchhof, L. Schmid, C. Reining, M. T. Hompel, and A. Pauly, "Chances of interpretable transfer learning for human activity recognition in warehousing," in *Proc. Int. Conf. Comput. Logistics*. Cham, Switzerland: Springer, 2021, pp. 163–177.
[21] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 3rd Quart., 2013.
[22] M. Cornacchia, K. Ozcan, Y. Zheng, and S. Velipasalar, "A survey on activity detection and classification using wearable sensors," *IEEE Sensors J.*, vol. 17, no. 2, pp. 386–403, Jan. 2017.
[23] A. Prati, C. Shan, and K. I.-K. Wang, "Sensors, vision and networks: From video surveillance to activity recognition and health monitoring," *J. Ambient Intell. Smart Environ.*, vol. 11, no. 1, pp. 5–22, 2019.
[24] K. S. Kumar and R. Bhavani, "Human activity recognition in egocentric video using HOG, GiST and color features," *Multimedia Tools Appl.*, vol. 79, no. 5, pp. 3543–3559, 2020.

[25] P. K. Roy and H. Om, "Suspicious and violent activity detection of humans using hog features and SVM classifier in surveillance videos," in *Advances in Soft Computing and Machine Learning in Image Processing*. Cham, Switzerland. Springer, 2018, pp. 277–294.

[26] A. Thyagarajmurthy, M. Ninad, B. Rakesh, S. Niranjan, and B. Manvi, "Anomaly detection in surveillance video using pose estimation," in *Emerging Research in Electronics, Computer Science and Technology*. Cham, Switzerland: Springer, 2019, pp. 753–766.

[27] L. Martínez-Villaseñor and H. Ponce, "A concise review on sensor signal acquisition and transformation applied to human activity recognition and human–robot interaction," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 6, 2019, Art. no. 1550147719853987.

[28] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, and S. J. Maybank, "Asymmetric 3D convolutional neural networks for action recognition," *Pattern Recognit.*, vol. 85, pp. 1–12, Jan. 2019.

[29] Y. Yan, T. Liao, J. Zhao, J. Wang, L. Ma, W. Lv, J. Xiong, and L. Wang, "Deep transfer learning with graph neural network for sensor-based human activity recognition," 2022, *arXiv:2203.07910*.

[30] C. Hu, Y. Chen, L. Hu, and X. Peng, "A novel random forests based class incremental learning method for activity recognition," *Pattern Recognit.*, vol. 78, pp. 277–290, Jun. 2018.

[31] Q. K. Xiao and R. Song, "Action recognition based on hierarchical dynamic Bayesian network," *Multimedia Tools Appl.*, vol. 77, pp. 6955–6968, Sep. 2018.

[32] C. A. Ronao and S.-B. Cho, "Recognizing human activities from smartphone sensors using hierarchical continuous hidden Markov models," *Int. J. Distrib. Sensor Netw.*, vol. 13, no. 1, Jan. 2017, Art. no. 155014771668368.

[33] P. Sok, T. Xiao, Y. Azeze, A. Jayaraman, and M. V. Albert, "Activity recognition for incomplete spinal cord injury subjects using hidden Markov models," *IEEE Sensors J.*, vol. 18, no. 15, pp. 6369–6374, Aug. 2018.

[34] B. M. Abidine, L. Fergani, B. Fergani, and M. Oussalah, "The joint use of sequence features combination and modified weighted SVM for improving daily activity recognition," *Pattern Anal. Appl.*, vol. 21, no. 1, pp. 119–138, Feb. 2018.

[35] C. Zhangjie and Y. Wang, "Infrared–ultrasonic sensor fusion for support vector machine–based fall detection," *J. Intell. Mater. Syst. Struct.*, vol. 29, no. 9, pp. 2027–2039, 2018.

[36] I. Portugal, P. Alencar, and D. Cowan, "The use of machine learning algorithms in recommender systems: A systematic review," *Expert Syst. Appl.*, vol. 97, pp. 205–227, May 2015.

[37] T. N. Nguyen, H. Nguyen-Xuan, and J. Lee, "A novel data-driven nonlinear solver for solid mechanics using time series forecasting," *Finite Elements Anal. Des.*, vol. 171, Apr. 2020, Art. no. 103377.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[39] D. Bhatt, C. Patel, H. Talsania, J. Patel, R. Vaghela, S. Pandya, K. Modi, and H. Ghayvat, "CNN variants for computer vision: History, architecture, application, challenges and future scope," *Electronics*, vol. 10, no. 20, p. 2470, Oct. 2021.

[40] C. Patel, D. Bhatt, U. Sharma, R. Patel, S. Pandya, K. Modi, N. Cholli, A. Patel, U. Bhatt, M. A. Khan, S. Majumdar, M. Zuhair, K. Patel, S. A. Shah, and H. Ghayvat, "DBGC: Dimension-based generic convolution block for object recognition," *Sensors*, vol. 22, no. 5, p. 1780, Feb. 2022.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 56, pp. 1929–1958, 2014.

[44] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey, PA, USA: IGI Global, 2010, pp. 242–264.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[46] G. Csurka, C. Dance, L. Fan, J. Willamowski, and A. Bray, "Visual categorization with bags of keypoints," in *Proc. Workshop Stat. Learn. Comput. Vis. (ECCV)*, vol. 1, Prague, Czech Republic, 2004, pp. 1–2.

[47] V. Delaitre, I. Laptev, and J. Sivic, "Recognizing human actions in still images: A study of bag-of-features and part-based representations," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2010, pp. 1–12.

[48] Z. Zhao, H. Ma, and S. You, "Single image action recognition using semantic body part actions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3391–3399.

[49] F. Najar, S. Bourouis, N. Bouguila, and S. Belghith, "Unsupervised learning of finite full covariance multivariate generalized Gaussian mixture models for human activity recognition," *Multimedia Tools Appl.*, vol. 78, no. 13, pp. 18669–18691, Jul. 2019.

[50] L. Liu, R. T. Tan, and S. You, "Loss guided activation for action recognition in still images," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer. 2018, pp. 152–167.

[51] M. Safaei and H. Foroosh, "Still image action recognition by predicting spatial-temporal pixel evolution," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 111–120.

[52] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Multibranch attention networks for action recognition in still images," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 4, pp. 1116–1125, Dec. 2018.

[53] S. Mohammadi, S. G. Majelan, and S. B. Shokouhi, "Ensembles of deep neural networks for action recognition in still images," in *Proc. 9th Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2019, pp. 315–318.

[54] H. H. Tan and K. H. Lim, "Vanishing gradient mitigation with deep learning neural network optimization," in *Proc. 7th Int. Conf. Smart Comput. Commun. (ICSCC)*, Jun. 2019, pp. 1–4.

[55] H. Ide and T. Kurita, "Improvement of learning for CNN with ReLU activation by sparse regularization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2684–2691.

[56] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[57] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei, "Human action recognition by learning bases of action attributes and parts," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1331–1338.

[58] S. Ma, S. A. Bargal, J. Zhang, L. Sigal, and S. Sclaroff, "Do less and achieve more: Training CNNs for action recognition utilizing action images from the web," *Pattern Recognit.*, vol. 68, pp. 334–345, Aug. 2017.

[59] B. Ham, M. Cho, C. Schmid, and J. Ponce, "Proposal flow: Semantic correspondences from object proposals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 7, pp. 1711–1725, Jul. 2018.

[60] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[62] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[63] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[64] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.

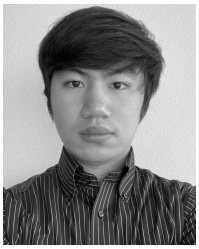
[65] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[66] C. I. Patel, D. Labana, S. Pandya, K. Modi, H. Ghayvat, and M. Awais, "Histogram of oriented gradient-based fusion of features for human action recognition in action video sequences," *Sensors*, vol. 20, no. 24, p. 7299, Dec. 2020.

[67] C. I. Patel, S. Garg, T. Zaveri, A. Banerjee, and R. Patel, "Human action recognition using fusion of features for unconstrained video sequences," *Comput. Electr. Eng.*, vol. 70, pp. 284–301, Aug. 2018.

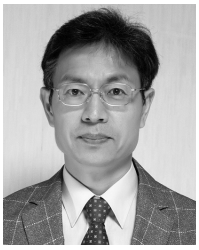
[68] M. Safaei, P. Balouchian, and H. Foroosh, "UCF-STAR: A large scale still image dataset for understanding human actions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 2677–2684.

[69] V. Delaitre, J. Sivic, and I. Laptev, "Learning person-object interactions for action recognition in still images," in *Proc. 25th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2011, pp. 1–9.
[70] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3506–3513.
[71] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 652–659.
[72] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg, "Coloring action recognition in still images," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 205–221, Dec. 2013.
[73] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, "Semantic pyramids for gender and action recognition," *IEEE Trans. Image Process.*, vol. 23, no. 8, pp. 3633–3645, Aug. 2014.
[74] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 73–80.
[75] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.
[76] Z. Zhao, H. Ma, and X. Chen, "Generalized symmetric pair model for action classification in still images," *Pattern Recognit.*, vol. 64, pp. 347–360, Apr. 2017.
[77] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Cham, Switzerland: Springer, 2010, pp. 177–186.



KOKI HIROOKA was born in Aizumisato-machi, Fukushima, Japan. He received the bachelor's degree in computer science and engineering from The University of Aizu (UoA), Japan, in March 2022. He is currently pursuing the master's degree. He joined the Pattern Processing Laboratoratory, UoA, in April 2021, under the supervision of Prof. Dr. Jungpil Shin and Prof. Dr. Md. Al Mehedi Hasan. His research interests include computer vision, pattern recognition, and deep learning. He is currently working on human activity recognition from still images.

MD. AL MEHEDI HASAN received the B.Sc., M.Sc., and Ph.D. degrees in computer science & engineering from the Department of Computer Science & Engineering, University of Rajshahi, Rajshahi, Bangladesh, in 2005, 2007, and 2017, respectively. He became a Lecturer, an Assistant Professor, an Associate Professor, and a Professor with the Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology (RUET), Rajshahi, in 2007, 2010, 2018, and 2019, respectively. His research interests include bioinformatics, artificial intelligence, pattern recognition, medical image, signal processing, machine learning, computer vision, data mining, big data analysis, probabilistic and statistical inference, operating systems, computer networks, and security. He has coauthored more than 100 publications published in widely cited journals and conferences.



JUNGPIL SHIN (Senior Member, IEEE) received the B.Sc. degree in computer science and statistics and the M.Sc. degree in computer science from Pusan National University, South Korea, in 1990 and 1994, respectively, and the Ph.D. degree in computer science and communication engineering from Kyushu University, Japan, in 1999, under the scholarship from the Japanese Government (MEXT). He was an Associate Professor, a Senior Associate Professor, and a Professor with the School of Computer Science and Engineering, The University of Aizu, Japan, in 1999, 2004, and 2019, respectively. He has coauthored more than 270 publications published in widely cited journals and conferences. His research interests include pattern recognition, image processing, computer vision, and machine learning, human–computer interaction, non-touch interface, human gesture recognition, automatic control, Parkinson's disease diagnosis, ADHD diagnosis, user authentication, machine intelligence, handwriting analysis, recognition, and synthesis. He is a member of ACM, IEICE, IPSJ, KISS, and KIPS. He served several conferences as a program chair and a program committee member for numerous international conferences. He serves as an Editor for journals of IEEE and *Sensors* (MDPI). He serves as a reviewer for several IEEE and SCI major journals.

AZMAIN YAKIN SRIZON received the B.Sc. degree in computer science & engineering. He is currently pursuing the M.Sc. degree in computer science & engineering with the Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Kazla, Rajshahi, Bangladesh. In 2021, he joined as a Lecturer with the Computer Science and Engineering Department, Rajshahi University of Engineering & Technology. He is the coauthor of 23 publications. His research interests include machine learning, deep learning, transfer learning, computer vision, pattern recognition, bioinformatics, biomedical, medical image processing, and human activity recognition & gesture recognition.