International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 15 (2023) pp. 250-258

Received: 19 Feb.2023; Accepted:2 June, 2023; Published: 6 June, 2023

Ensemble Transfer Learning for Robust Human Activity Recognition from Images

Aayush Dhattarwal[1], Saroj Ratnoo[1], Anu Bajaj[2, *], Ajith Abraham[3]

[1] Department of Computer Science and Engineering,
Guru Jambheshwar University of Science and Technology, Hisar-125001, India

[2] Computer Science and Engineering Department,
Thapar Institute of Engineering and Technology, Patiala, India
[*] er.anubajaj@gmail.com

[3] Faculty of Computing and Data Science,
Flame Univeristy, Pune, India

***Abstract*: In recent years, the field of Human Activity Recognition (HAR) has witnessed a significant growth owing to the abundance of data and its practical applications in various real-world scenarios. The recognition of human activities from still images remains a challenging task due to the presence of class imbalance and limited intra-class variability. To address these issues, this work proposes an Ensemble Transfer Learning approach for image-based HAR. The proposed model employs an ensemble stacked averaging model consisting of well-known transfer learning architectures such as ResNet50V2, DenseNet169 and VGG19. The ensemble model can learn different features from different architectures, thus providing a robust recognition model. Additionally, data augmentation is employed to increase the diversity of the images in the datasets. The suggested model helps to mitigate the problems of class-imbalance and the lack of intra-class variability by generating new images with different variations of the original images. The model is evaluated on two benchmark datasets for image based HAR, namely, the PPMI action dataset and the Stanford 40 Actions dataset. The results demonstrate enhanced performance compared to a few of the related research works.**

***Keywords*:** Transfer Learning, Ensemble Learning, Data Augmentation, Human Activity Recognition (HAR), Computer Vision.

I. Introduction

Human Activity Recognition (HAR) is a domain of research that aims to classify human activities. It has broad spectrum of applications including healthcare (e.g., fall detection for older adults, gait analysis for patients with movement disorders), sports and fitness (e.g., tracking and analysis of athletic performance), and smart environments (e.g., ambient assisted living, human-computer interaction). HAR relies on a wide range of data sources, including sensor data, image data, audio data and video data. Due to the increasing availability of such datasets, the research in HAR has gained significant momentum in the last decade.

Convolutional neural networks (CNNs) have emerged as one of the most prevalent techniques for vision-based human activity recognition (HAR). CNNs are a type of deep learning model that excel at image recognition tasks by automatically learning and extracting features from images. In recent years, there have been several studies that have demonstrated the effectiveness of CNNs for image based HAR (e.g., [3]-[5]). One of the major limitations of CNNs is that they require a large amount of labeled data for training, which can be difficult and time-consuming to acquire. Furthermore, CNNs are prone to overfitting if the training data is not diverse enough. They also have limited generalizability, i.e. CNNs are typically trained on a specific dataset and task and may not generalize well to other datasets or related tasks.

It is important to have diversity in the image data to build CNN models with high generalization power. Therefore, data augmentation has been used to enhance the predictive performance of CNN models. Data augmentation includes techniques such as cropping, flipping, and rotating images to improve diversity in the training data. However, the process of applying various transformations costs additional computational resources, especially when dealing with large datasets. This can be a major limitation for real-time human activity recognition systems, where speed and efficiency are critical [1].

Also, Transfer learning has become increasingly popular as an alternative approach. It entails using a pre-trained model for one task as the foundation for training a model on a similar task. This can be particularly useful for HAR, as it allows researchers to leverage the large amounts of labeled data and computational resources needed to train deep learning models, without having to start from scratch. Several studies have shown that transfer learning can improve the performance of CNNs for still-image based HAR (e.g., [11]-[15]).Some of the limitations of transfer learning include overfitting, particularly when fine-tuning on small datasets [2]. This can lead to poor generalization and low performance on unseen data. Also, transfer learning can be sensitive to the choice of pre-trained

model and the amount of fine-tuning used [2]. This makes it difficult to achieve consistent performance across different datasets and applications.

The use of stacked ensemble averaging (SEA) can further improve the performance of classifiers. One approach is to train several models using the same dataset and then combine their predictions by averaging them to obtain a final prediction. This approach has been shown to be effective for vision based HAR also, particularly when the models are trained on different types of data or using different training algorithms (e.g., [16]-[17]).

In this paper, we propose a method based on transfer learning with ensemble stacked averaging for training the CNN architectures for image-based HAR. In this research, we have used pre-trained weights from ImageNet dataset and updated them on Stanford40 and PPMI datasets to overcome resource constraints. Additionally, data augmentation is used to enhance performance by increasing the diversity in the training data. By combining Data Augmentation techniques and Stacked Ensemble Averaging method, we got promising results on two benchmark datasets in image based HAR. The proposed method achieves competitive results than the other methods quoted in the literature.

The rest of the paper is organized as follows: Section II deals with the literature review. Section II describes the methodology and tools. Section IV presents the result. Section V concludes the research carried out in this paper.

II. Literature Review

A. Convolutional Neural Networks

Over the last decade, there has been a significant amount of research focused on still-image based human activity recognition (HAR). Convolutional neural networks (CNNs) have emerged as one of the powerful tools for this task due to their ability to directly learn complex patterns from images. Several studies have demonstrated the effectiveness of these approaches for image based HAR. For instance, Ji et al. (2012) published study in which CNNs were used to recognize human activities in surveillance videos of airports with very impressive accuracy [3].

Attique Khan et al. (2021) proposed a 26-layered CNN architecture for accurate complex action recognition, achieving 81.4%, 98.3%, 98.7%, and 99.2% accuracy respectively on HMDB51, KTH, Weizmann, and UCF Sports datasets, outperforming some of the contemporary works based on classical machine learning [4]. Ahmed et al. (2020) presented a motion classification method using a CNN to extract information through convolutional layers and a Softmax classifier in a fully connected layer to categorize human motion. The method achieved high success rates of 98.75% with KTH, 92.24% with Ixmas, and 100% with the Weizmann datasets [5]. M. Bilal et al. (2021) suggested a deep learning approach that combines CNN, LSTM, and RNN to minimize computational expenses and attain cutting-edge results in Human Action Recognition (HAR) for intersecting actions in temporal visual data streams that span long time periods [6].

Y. Lavinia et. al (2016) proposed a fusion-based method using multiple CNN models working on RGB and oRGB color spaces. Next, the features that were extracted are sent to SVM for recognition. This method has shown competitive performance on Stanford 40 and PPMI [7]. Moreover, Safaei et, al. (2019) proposed a novel CNN which uses spatial and temporal features of an image for single image activity recognition (STCNN). In the experiments, the STCNN demonstrated significant improvement over other state-of-the-art methods [8].

B. Data Augmentation

Several authors use data augmentation to avoid overfitting problem. Data augmentation converts single images into many images through certain transformation operations. Many research works show that data augmentation boosts the performance of HAR models. For Example, Alani et al. (2018) used Adapted Deep Convolutional Neural Network (ADCNN) in combination with data augmentation on 3750 hand gesture images to boost the performance of the model. The model outperformed other methods that do not employ data augmentation techniques [9]. Also Meng et al. (2019) proposed a novel data augmentation network called Sample fusion network, which uses LSTM Autoencoder to generate new samples. Experiments demonstrated that SFN greatly enhances the performance of the model, achieving an accuracy of 79.53% to 90.75% [10]. Also, Z. Islam (2019) presented a method for recognizing static hand gestures in still images using a CNN model and data augmentation techniques to increase the size of the training dataset resulting in the improved performance of the HAR model. They tested the method on a dataset of still images comprising of 10 classes and achieved an accuracy of 97.12%. The method outperforms CNN model without data augmentation [1]. Chakraborty et al. (2021) used transfer learning CNN models along with data augmentation. Their result showed improved accuracy on Stanford 40 and PPMI datasets [2]. Also, Sahoo et al. (2021) used sequential learning and depth-estimated history images with data augmentation to avoid overfitting, achieving the highest recognition rate of 97.67% on the KTH dataset [11].

C. Transfer Learning

Transfer learning, in which a model is pre-trained on a large dataset and fine-tuned on a smaller related dataset, has also been applied to still-image based HAR. Transfer learning approach has successfully reduced the amount of labeled data required for training and it has also improved the performance on the target task [2]. T. Ozcan et al. (2019) proposed a method which implements heuristic optimization algorithms for hyperparameter tuning of CNN structures based on the AlexNet model. The model that was suggested was evaluated on two datasets, namely the sign language digits dataset and Thomas Moeslund's gesture recognition dataset. Based on the experimental findings, the proposed artificial bee colony-based method produced an average accuracy of 98.40% for action classification, which surpasses the performance of prior studies on the sign language digits dataset. Additionally, for Thomas Moeslund's gesture recognition dataset, the advised approach attained an average accuracy of 98.09%, outdoing the top pre-existing method [12]. Also, Tan et al. (2018) found that deep transfer learning improves the classification accuracy of models compared to traditional machine learning methods [13]. The research done by Nawaratne et al. (2020) presented a transfer learning approach for human activity recognition and self-organizing maps (SOMs). The method involves first training a CNN on a source dataset, and then

fine-tune the network on a target dataset using SOMs to map the features from the CNN to the target dataset [14]. A novel meta-learning model was proposed by Jang et al. (2019) and it determines which layers and features of a teacher network to match and learn depending on the importance of the classification task's features. The model underwent comprehensive testing on several datasets, including STL-10, CIFAR-100, Stanford 40, Stanford Dogs, MIT 67 and CUB 200, and was proven effective [15]. Finally, K. K. Verma et. al (2021) proposed a deep transfer learning method, with a multiclass SVM for extracting features for activity recognition from videos. A fine-tuned VGG-19 architecture was applied for visual feature extraction, and a multi-class Support Vector Machine was used for classification. It achieved 97.13% accuracy on the UCF Sports Action dataset, outperforming handcrafted feature engineering methods [16].

D.Ensemble Learning

Further, there is a growing trend of using Ensemble learning in vision based HAR. Ensemble Learning is a well-known methodology to enhance the classification performance. The ensemble learning reduces the model variance as well. Following the stacked ensemble approach, in which multiple models are trained and their predictions are combined, has also been applied to image based HAR. This technique improves the robustness and accuracy of the final model by leveraging the strengths of multiple individual models. Gour et al. (2022) proposed a stacked ensemble method with CNNs for image classification. The method is evaluated on three publicly available datasets and outperforms existing methods [17]. Also, Dhakate et al. (2020) proposed a method for detection of distracted driver to reduce accidents using stacking ensemble technique using various CNNs. The method achieved an overall accuracy of 97% outperforming the standard CNN models [18].

E. Hybrid Methods

In addition to CNNs, some hybrid methods that use additional spatial & temporal information have been proposed for HAR. For example, Liu et al. (2018) propose a multi-task learning approach that combines CNNs with temporal modeling for HAR in videos [19]. Moreover, in a research by Sharma et al. (2012), a method based on discriminative features and intra-class variations in subject poses is presented which uses Spatial Saliency maps to weigh visual features [20]. Also, Zhao et al. (2017) introduced a Generalized Symmetric Pair model that integrates a max-margin classifier and exhibited better outcomes compared to several standard techniques on datasets for still image based HAR [21]. J. Zhang (2016) improved feature recognition using tensor descriptors and tucker decomposition, outperforming previous works on 3 publicly available datasets [22]. L. Zhang et al (2016) improved upon Vector of Locally Aggregated Descriptor (VLAD) by addressing empty cavity and assignment ambiguity using middle-level assignments. The proposed method was proven effective on Stanford40 and PPMI dataset outperforming related studies [23]. Moreover, in a study by F.S. Khan et al. (2014), a semantic pyramid method was introduced to normalize the subject's positions, utilizing pre-trained detectors to extract semantic information of full-body, upper-body, and face regions for classification [24]. Additionally, in [25], Safaei et al. (2019) proposed a novel CNN which uses spatial and temporal features of an image for single image activity recognition (STCNN). In the experiments, the STCNN demonstrated significant improvement over other existing methods.

We have learnt from the literature review that though many above mentioned methods in HAR have shown promising results, nonetheless these require large amounts of annotated data and are computationally intensive [28-34]. Also, in some cases the methods are prone to overfitting. Further, there is a need to develop flexible approaches to addresses these issues.

III. Methodology and Experimental Setup

This Section describes the proposed model used for Human Activity Recognition in this study. The overall research methodology is given in the Figure 1.

1) The Datasets

The following Datasets are used for evaluating the proposed stack ensemble model:

Stanford 40: The Stanford 40 Action dataset comprises 9532 images featuring 40 distinct human action classes, with varying sizes in Joint Photographic Experts Group (JPEG) format. This dataset is typically bifurcated into two subsets: the Body Motion dataset and the Non-body Motion dataset. The Body Motion dataset contains images of 11 activities, including Climbing, Cleaning floor, Waving Hands, Jumping, Riding Horse, Throwing Frisbee, and Riding Bike. The remaining 29 activities are classified under the Non-body Motion dataset. The Stanford 40 dataset is a widely recognized benchmark for Human Activity Recognition [26].

People Playing Musical Instruments (PPMI): This dataset is comprised of two distinct sets of data: Play Instrument and With Instrument. These datasets contain images that depict individuals interacting with twelve different musical instruments. Within the Play Instrument dataset, the images showcase individuals playing the musical instruments, while the With Instrument dataset includes images of individuals holding the musical instruments [27]. The sample images from Stanford 40 and PPMI are shown in Figure 1.

Stanford 40 Dataset sample

PPMI Dataset sample

Figure 1. Stanford 40 and PPMI Dataset

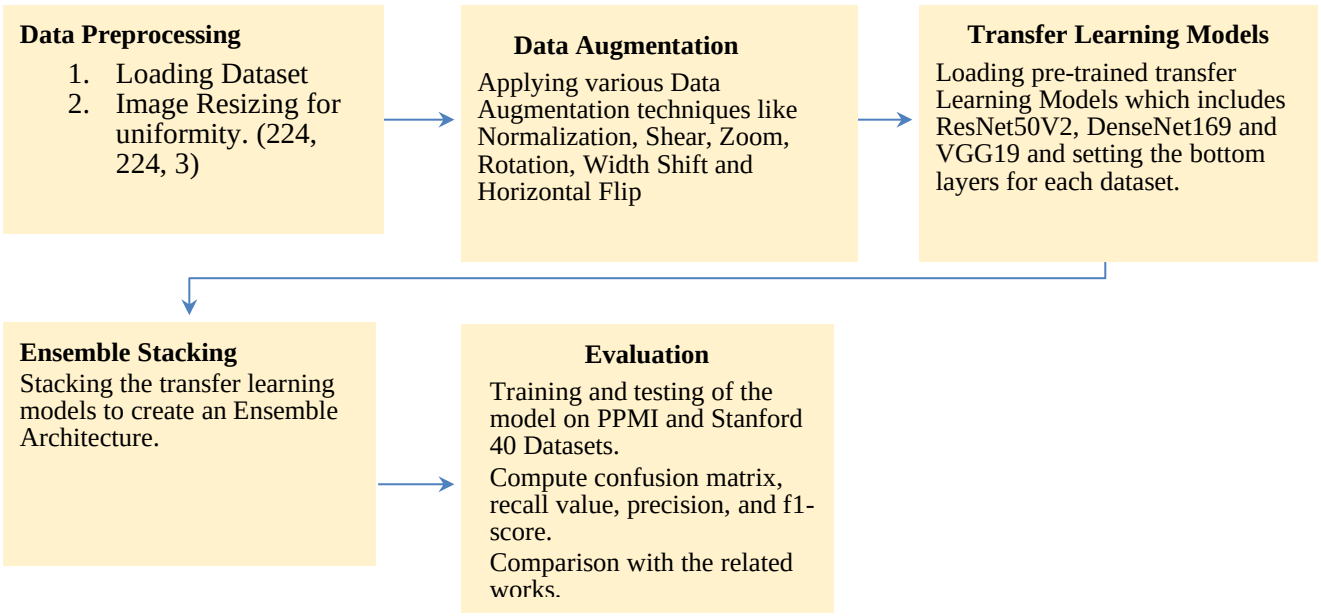

Figure 2. Process flow of the proposed work.

2) Data Augmentation

Data augmentation can help to improve the performance of the any predictive model by providing additional diversity in the input images. By generating diverse training data, data augmentation creates base learners that make different types of errors, which can be beneficial particularly for the ensemble models as it can help to reduce the overall variance and bias of the predictions. The data augmentation techniques used, and their parameter settings are given in table 1.

3) The Stacked Ensemble Model

This paper uses stacked average ensemble model that combines the predictions of multiple individual models to improve the overall predictive performance. Since this method helps to reduce the overall variance and bias of the predictive model, it is particularly useful when the individual models in the ensemble have different types of errors. We have utilized three transfer learning models - VGG19, DenseNet169, and ResNet50V2 - as base learners for our stacked average ensemble model. These models are chosen due to their optimization capability for images of size (224, 224, 3). The predictions from these base learners are then combined and averaged to create the meta learner. The meta learner used in the experiments is shown Figure 3.

Table 1. Data augmentation with parameters

Augmentation Technique	Parameters

Normalization	-
Shear	0.2
Zoom	0.2
Rotation	20
Width Shift	0.2
Horizontal Flip	-

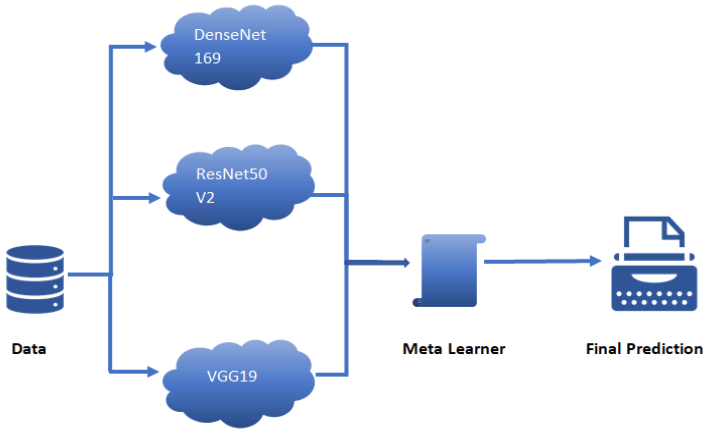

Figure 3. Stacked Ensemble Model

The experiments are carried out on Google Collab Pro platform. In all the experiment, a train-test split of 70%-30% is used, i.e., 70% of the data is used for training and 30% of the data is used for testing purposes. The performance of the stacked ensemble is reported in terms of accuracy, sensitivity, specificity F-score etc.

IV. Experiments and Results

1) Experiments Conducted

A total of 6 experiments are conducted to measure the performance of the proposed approach out of which three experiments are conducted on *Stanford 40 Actions* dataset on *Body Action Dataset, Non Body Action* Dataset and the whole dataset including all the 40 classes. The remaining three experiments pertain to PPMI. The Stacked Ensemble method includes keras implementation of the three transfer learning models pre-trained on 'imagenet' dataset, i.e., VGG19, DenseNet169 and ResNet50V2. These models were selected due to homogeneity in their input layers, i.e., they all accept images of the height and width of 224 pixels with 3 channels which denote RGB. The hyperparameters used in the experiments are given in Table 2.

Table 2. Hyperparameters used in experiments

Hyperparameters	**Value**
Optimization Algorithm	SGD
Learning Rate	0.001
Momentum	0.9
Batch Size	32
Number of Epochs	50

2) Experimental Results

Number This section focuses on the results obtained for each Dataset. The performance of our approach is evaluated in terms of precision, accuracy, F-score. The results are shown in Table 3. Two sample accuracy and loss plots are shown in Figure 4.1 and Figure 4.2.

Table 3. Performance of proposed method

Datasets	**Accuracy**	**Recall**	**F1-Score**	**Precision**
Stanford 40 Body Motion	93%	93%	93%	93%
Stanford 40 Non-body Motion	70%	69%	72%	84%
Stanford 40 All Classes	75%	74%	75%	79%
PPMI Play Instrument	93%	93%	93%	94%
PPMI With Instrument	83%	83%	83%	85%
PPMI 24 Classes	81%	81%	81%	83%

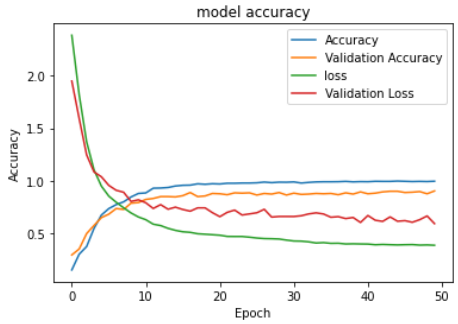

Figure 4.1 Accuracy and Loss plot for PPMI play instrument Dataset

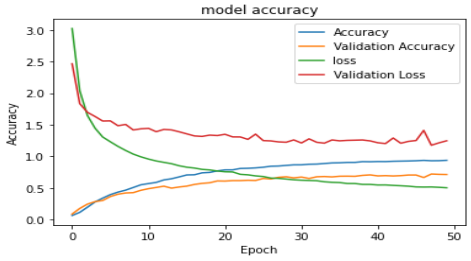

Figure 4.2 Accuracy and loss plot for Stanford 40 Non Body Motion Dataset

The proposed model performed well on the PPMI play instrument dataset, with accuracy, F1 score, recall, and precision of 0.93, 0.93, 0.93, and 0.94, respectively. We found that the class-specific precision is relatively higher for French horn, guitar, harp, violin, and bassoon, and relatively lower for saxophone, flute, recorder, clarinet, and trumpet. When tested on PPMI with *Instrument* dataset, the model achieves an accuracy of 0.83, recall of 0.83, F1 score of 0.83, and precision of 0.85. The relatively higher-performing classes in terms of precision are harp, trumpet, erhu, saxophone, and violin, while the relatively lower precision classes are flute, bassoon, guitar, cello, and French horn. On *Stanford 40 Body Motion* Dataset, the proposed method achieves 0.93 accuracy, 0.93 recall, 0.93 F1 score and 0.94 precision. The relatively higher performing classes in terms of precision are climbing, riding a bike, riding a horse and shooting an arrow, while the classes with relatively precision are jumping, throwing Frisby, running, walking the dog, waving hands and rowing a boat. The model obtains an accuracy of 0.70, recall of 0.69, F1 score of 0.72, and precision of 0.84 on the Stanford 40 *Non Body Motion Dataset*. The relatively higher performing classes in terms of precision are blowing bubbles, brushing teeth, playing guitar, watching TV, and fixing a car, while the classes with relatively lower precision are fixing a bike, writing on a book, smoking, reading, and phoning.

3) Performance Comparison with some state-of-the-art methods

The proposed Stacked Ensemble Averaging Method with data augmentation is compared with some similar works on the same datasets in Tables 4 and 5.

In this subsection, the results obtained from the proposed Stacked Averaging Ensemble Method is compared with some recent and benchmark methods on PPMI and Stanford 40 Datasets in terms of Mean Average Precision (mAP) scores. From the comparison presented in Table 4, Table 5, it is evident that the method outperforms the other methods on PPMI Dataset with Mean Average Precision (mAP) scores for PPMI play instrument, PPMI With Instrument and PPMI 24 classes Dataset scores at 0.94, 0.85 and 0.83 respectively. A visual comparison is given in Figures 5 to 10. In case of Stanford 40 Dataset, both in the case of Stanford 40 Classes Dataset and Stanford Body Motion Dataset, our method does not achieve the highest score. However, when considering Stanford 40 Non-Body Motion dataset, our method outperforms the other methods as seen in the Table 5. TSSTN method proves to be the best for Stanford 40 (All Classes) and Stanford 40 (Body Motion Dataset) with 0.86 and 0.97 as mAP scores respectively while our method exceeds all other methods on Non-body Motion Dataset with mAP score at 0.84. Overall, the proposed method is competitive with existing methods for image based HAR.

Table 4. Comparison of the proposed method on PPMI Dataset

Dataset	Reference	Method	Play Instrument	With Instrument	24 Classes
PPMI	[20]	Discriminative Spatial saliency	-	-	0.49
	[21]	Generalized Symmetric pair model	-	-	0.52
	[7]	Color Fusion Deep Learning Model	0.70	0.61	0.66
	[23]	Optimal VLAD	-	-	0.48
	[22]	TuRR	-	-	0.72
	[2]	Transfer Learning based HAR	0.85	0.74	0.74
	The Proposed Method	Stacked Ensemble Averaging Classifier	0.94	0.85	0.83

Table 5. Comparison of the proposed method on Stanford 40 Dataset

Dataset	Reference	Method	Body Motion		Non-Body Motion	All Classes
Stanford 40	[24]	Semantic Pyramids	0.57	0.52		0.53
	[15]	Meta-Learning	_	_		0.63
	[23]	Optimal VLAD	-	-		0.37
	[8]	STCNN	0.94	0.73		0.81
	[25]	TSSTN	**0.97**	0.80		**0.86**
	[2]	Transfer Learning based HAR	0.96	0.82		0.77
	The Proposed Method	Stacked Ensemble Averaging Classifier	0.93	0.84		0.79

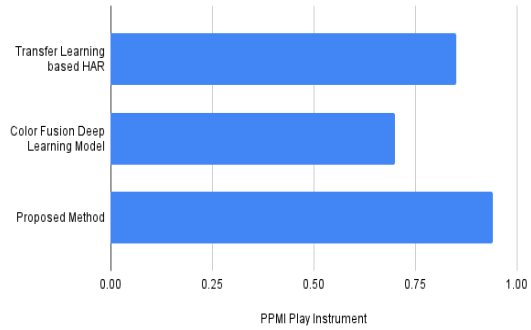

Figure 5. Comparison on PPMI Play Instrument Dataset

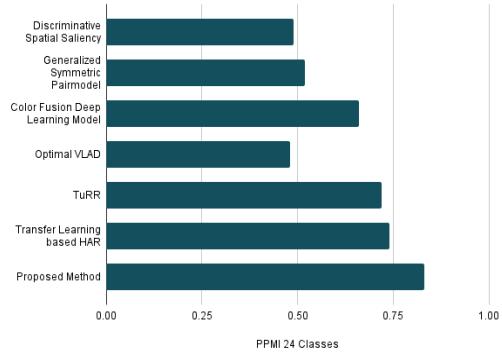

Figure 6 – Comparison on PPMI 24 Classes Dataset

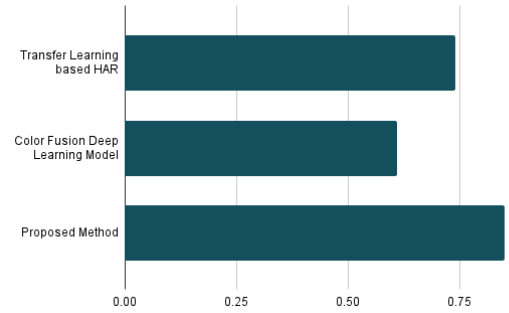

Figure 7. Comparison on PPMI With Instrument dataset

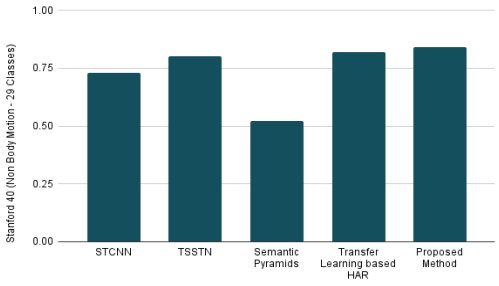

Figure 8. Comparison on Stanford 40 Non Body Motion dataset

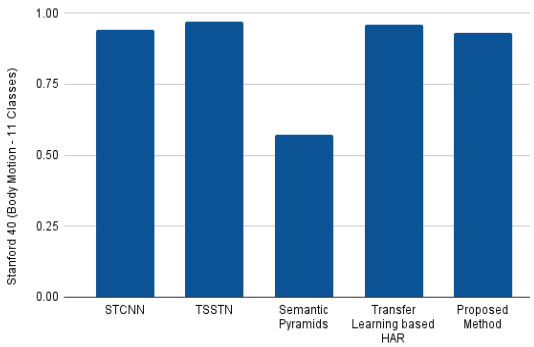

Figure 9. Comparison on Stanford 40 Body Motion Dataset

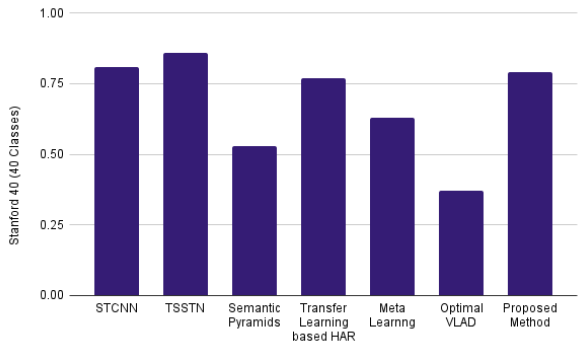

Figure 10. Comparison on Stanford 40 All Classes Dataset

V. Conclusion and Future Scope

In this study, we have worked on Human Activity Recognition based on still images. Image based Human Action Recognition (HAR) is important as it allows for the analysis of actions captured in a single frame, rather than relying on a sequence of frames. This can be useful in various applications such as surveillance, sports analysis, and behavior understanding. Additionally, it is less computationally intensive and can be applied to real-time systems. Here, we have proposed a Stacked Ensemble Transfer Learning approach that is validated on two of the benchmark datasets in the area, i.e., PPMI action dataset and Stanford 40 Actions dataset. We have also used some of the most popular data augmentation techniques to enhance the diversity in the image data. The proposed approach has achieved the highest performance on PPMI actions dataset in all three configurations of dataset, i.e., PPMI play Instrument, PPMI with Instrument and PPMI 24 classes. The suggested method has obtained highest precision score in Stanford 40 Non-Body Motion dataset while it takes the third highest rank on Stanford 40 all classes dataset respectively.

The proposed Ensemble Transfer Learning approach for image-based Human Activity Recognition (HAR) can be further improved. One potential avenue for future research is to incorporate the other state-of-the-art methods, such as attention mechanisms or adversarial training to enhance the performance of the model. Another area of potential exploration is to apply this approach to other image-based recognition tasks, such as object recognition or facial expression recognition. Further, this approach can be adapted to other similar recognition tasks that have challenges of class-imbalance and limited intra-class variability. Lastly, this approach can be extended to real-world applications, such as surveillance systems, robotic assistants, or human-computer interaction, where image-based activity recognition is critical. Overall, the proposed Ensemble Transfer Learning approach has a lot of potential and can be further enhanced and extended to a wide range of practical applications.

References

[1] Md. Z. Islam, M. S. Hossain, R. ul Islam, and K. Andersson, "Static Hand Gesture Recognition using Convolutional Neural Network with Data Augmentation," in *2019 Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and 2019 3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, Spokane, WA, USA, May 2019, pp. 324–329. doi: 10.1109/ICIEV.2019.8858563.

[2] S. Chakraborty, R. Mondal, P. K. Singh, R. Sarkar, and D. Bhattacharjee, "Transfer learning with fine tuning for human action recognition from still images," *Multimed Tools Appl*, vol. 80, no. 13, pp. 20547–20578, May 2021, doi: 10.1007/s11042-021-10753-y.

[3] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[4] M. A. Khan, Y.-D. Zhang, S. A. Khan, M. Attique, A. Rehman, and S. Seo, "A resource conscious human action recognition framework using 26-layered deep convolutional neural network," *Multimed Tools Appl*, vol. 80, no. 28–29, pp. 35827–35849, Nov. 2021, doi: 10.1007/s11042-020-09408-1.

[5] [5] W. S. Ahmed and A. amir A. Karim, "Motion Classification Using CNN Based on Image Difference," in *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, Sydney, Australia, Nov. 2020, pp. 1–6. doi: 10.1109/CITISIA50690.2020.9371835.

[6] M. Bilal, M. Maqsood, S. Yasmin, N. U. Hasan, and S. Rho, "A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes," *J Supercomput*, vol. 78, no. 2, pp. 2873–2908, Feb. 2022, doi: 10.1007/s11227-021-03957-4.

[7] Y. Lavinia, H. H. Vo, and A. Verma, "Fusion Based Deep CNN for Improved Large-Scale Image Action Recognition," in *2016 IEEE International Symposium on Multimedia (ISM)*, San Jose, CA, USA, Dec. 2016, pp. 609–614. doi: 10.1109/ISM.2016.0131.

[8] M. Safaei and H. Foroosh, "Still Image Action Recognition by Predicting Spatial-Temporal Pixel Evolution," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa Village, HI, USA, Jan. 2019, pp. 111–120. doi: 10.1109/WACV.2019.00019.

[9] A. A. Alani, G. Cosma, A. Taherkhani, and T. M. McGinnity, "Hand gesture recognition using an adapted convolutional neural network with data augmentation," in *2018 4th International Conference on Information Management (ICIM)*, Oxford, May 2018, pp. 5–12. doi: 10.1109/INFOMAN.2018.8392660.

[10] F. Meng, H. Liu, Y. Liang, J. Tu, and M. Liu, "Sample Fusion Network: An End-to-End Data Augmentation Network for Skeleton-Based Human Action Recognition," *IEEE Trans. on Image Process.*, vol. 28, no. 11, pp. 5281–5295, Nov. 2019, doi: 10.1109/TIP.2019.2913544.

[11] S. P. Sahoo, S. Ari, K. Mahapatra, and S. P. Mohanty, "HAR-Depth: A Novel Framework for Human Action Recognition Using Sequential Learning and Depth Estimated History Images," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 5, no. 5, pp. 813–825, Oct. 2021, doi: 10.1109/TETCI.2020.3014367.

[12] T. Ozcan and A. Basturk, "Transfer learning-based convolutional neural networks with heuristic optimization for hand gesture recognition," *Neural Comput & Applic*, vol. 31, no. 12, pp. 8955–8970, Dec. 2019, doi: 10.1007/s00521-019-04427-y.

[13] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A Survey on Deep Transfer Learning," in *Artificial Neural Networks and Machine Learning – ICANN 2018*, vol. 11141, V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis, Eds. Cham: Springer International

Publishing, 2018, pp. 270–279. doi: 10.1007/978-3-030-01424-7_27.

[14] R. Nawaratne, D. Alahakoon, D. De Silva, H. Kumara, and X. Yu, "Hierarchical Two-Stream Growing Self-Organizing Maps With Transience for Human Activity Recognition," *IEEE Trans. Ind. Inf.*, vol. 16, no. 12, pp. 7756–7764, Dec. 2020, doi: 10.1109/TII.2019.2957454.

[15] Y. Jang, H. Lee, S. J. Hwang, and J. Shin, "Learning What and Where to Transfer," in *Proceedings of the 36th International Conference on Machine Learning*, Jun. 2019, vol. 97, pp. 3030–3039. [Online]. Available: https://proceedings.mlr.press/v97/jang19b.html

[16] K. K. Verma and B. Mohan Singh, "Vision based Human Activity Recognition using Deep Transfer Learning and Support Vector Machine," in *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Dehradun, India, Nov. 2021, pp. 1–9. doi: 10.1109/UPCON52273.2021.9667661.

[17] M. Gour and S. Jain, "Automated COVID-19 detection from X-ray and CT images with stacked ensemble convolutional neural network," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 1, pp. 27–41, Jan. 2022, doi: 10.1016/j.bbe.2021.12.001.

[18] K. R. Dhakate and R. Dash, "Distracted Driver Detection using Stacking Ensemble," in *2020 IEEE International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, Bhopal, India, Feb. 2020, pp. 1–5. doi: 10.1109/SCEECS48394.2020.184.

[19] T. Liu, J. Wang, S. Hutchinson, and M. Q.-H. Meng, "Skeleton-Based Human Action Recognition by Pose Specificity and Weighted Voting," *Int J of Soc Robotics*, vol. 11, no. 2, pp. 219–234, Apr. 2019, doi: 10.1007/s12369-018-0498-z.

[20] G. Sharma, F. Jurie, and C. Schmid, "Discriminative spatial saliency for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, Jun. 2012, pp. 3506–3513. doi: 10.1109/CVPR.2012.6248093.

[21] Z. Zhao, H. Ma, and X. Chen, "Generalized symmetric pair model for action classification in still images," *Pattern Recognition*, vol. 64, pp. 347–360, Apr. 2017, doi: 10.1016/j.patcog.2016.10.001.

[22] J. Zhang, Y. Han, and J. Jiang, "Tucker decomposition-based tensor learning for human action recognition," *Multimedia Systems*, vol. 22, no. 3, pp. 343–353, Jun. 2016, doi: 10.1007/s00530-015-0464-7.

[23] L. Zhang, C. Li, P. Peng, X. Xiang, and J. Song, "Towards optimal VLAD for human action recognition from still images," *Image and Vision Computing*, vol. 55, pp. 53–63, Nov. 2016, doi: 10.1016/j.imavis.2016.03.002.

[24] F. S. Khan, J. van de Weijer, R. M. Anwer, M. Felsberg, and C. Gatta, "Semantic Pyramids for Gender and Action Recognition," *IEEE Trans. on Image Process.*, vol. 23, no. 8, pp. 3633–3645, Aug. 2014, doi: 10.1109/TIP.2014.2331759.

[25] M. Safaei, P. Balouchian, and H. Foroosh, "UCF-STAR: A Large Scale Still Image Dataset for Understanding Human Actions," *AAAI*, vol. 34, no. 03, pp. 2677–2684, Apr. 2020, doi: 10.1609/aaai.v34i03.5653.

[26] 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 *2011 International Conference on Computer Vision*, Barcelona, Spain, Nov. 2011, pp. 1331–1338. doi: 10.1109/ICCV.2011.6126386.

[27] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun. 2010, pp. 9–16. doi: 10.1109/CVPR.2010.5540234.

[28] TF Gharib, H Nassar, M Taha, A Abraham, An efficient algorithm for incremental mining of temporal association rules, Data & Knowledge Engineering 69 (8), 800-815, 2010.

[29] S Dasgupta, S Das, A Biswas, A Abraham, On stability and convergence of the population-dynamics in differential evolution, AI Communications, 22 (1), 1-20, 2009.

[30] F. Xhafa, A. Abraham, Metaheuristics for Scheduling in Industrial and Manufacturing Applications, Studies in Computational Intelligence, Vol 128, 2008.

[31] A. Abraham, N.S. Philip, P. Saratchandran, Modeling chaotic behavior of stock indices using intelligent paradigms, arXiv preprint cs/0405018, 2004.

[32] L Dora, S Agrawal, R Panda, A Abraham, Optimal breast cancer classification using Gauss–Newton representation based algorithm, Expert Systems with Applications, 97: 134-145, 2017.

[33] A Rajasekhar, RK Jatoth, A Abraham, Design of intelligent PID/PIλDμ speed controller for chopper fed DC motor drive using opposition based artificial bee colony algorithm, Engineering Applications of Artificial Intelligence, 97: 13-32, 2014.

[34] A Abraham, Intelligent systems: Architectures and perspectives, Recent advances in intelligent paradigms and applications, Springer, 1-35, 2003.