Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 15, Issue. 2 : 2024
ISSN :**1906-9685**

GOLDENAGER: A SMART ACTIVITY RECOGNITION SYSTEM FOR ELDERLY BASED ON FEATURE FUSION

[1] **KALLURU SOWMYA,** PG SCHOLAR IN DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING IN ANNAMACHARYA INSTITUTE OF TECHNOLOGY AND SCIENCES, KADAPA, A.P. sowmyanagam92@gmail.com**.**
[2] **SHAIK MOHAMMED JABEER,** ASSISTANT PROFESSOR IN DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING IN ANNAMACHARYA INSTITUTE OF TECHNOLOGY AND SCIENCES, KADAPA, A.P. smohdjabeer@gmail.com.

Abstract: Among the many fascinating uses of computer vision, human action detection stands out as a prime research topic. There are a number of practical challenges that contribute to action detection ambiguities, including as camera movement, occlusion, background noise, and the difficulty in precisely identifying the motion of individual body parts. We are motivated to do substantial research in this field since academics have paid little attention to elderly activity identification systems. Despite the many credible approaches presented in the literature, they are still not enough to resolve the problems entirely. Because older people's behaviours differ from younger people's for a number of reasons—the most important of which being health concerns—it would be inappropriate to test a model developed for older people on a dataset consisting of younger people, as the results may change in real-time. The suggested model has both automatically learning and manually designed elements. The model's incorporation of two characteristics acquired via separate methods makes this HAR model more effective. Extensive performance indicators have been used to assess the statistical and qualitative efficacy of these techniques in relation to the suggested model. Public benchmark datasets like the Stanford-40 Dataset have been used to test and verify it with existing models.
Keywords: - Elderly, Elderly activity recognition, HAR, Hand-Engineered Features

1. Introduction

The automated recognition of a number of physical tasks people regularly engage in is known as "human action recognition" (HAR) [1]. The primary objective of HAR is to represent human actions and the interactions between them in an accurate way. These activities are recognized from an unseen dataset [2]. Human activity can be discovered using a range of sensors. Multimodal, vision-based, and sensor-based HAR techniques make up the three categories [3]. An approach based on vision is utilised to record human activities using a single camera. Vision-based systems are the original HAR techniques, which have attracted a lot of attention in the past. The three different categories of sensor-based data are wearable, object-tagged, and device-free data. The wearer of wearable sensors must wear them all the time. However, object tagging uses tags attached to daily routine items to record user activities. Each of these two methods are specific to devices. While adopting the method in which devices are not used, various sensors are positioned near the immediate vicinity [4]. To properly identify human activity, a multimodal approach includes multiple types of sensors [5].

A greater range of applications for 2D and 3D video cameras, including those for consumer surveillance systems, security systems, person monitoring systems, and smart home systems, have been made possible by the growing need for video technology in intelligent and automated environments. Many researchers have focused on 2D and 3D video applications [8].

In order to recognize a particular action, the video is split into number of frame sequences in which every sequence consists of an instance of a human action. The system's goal in such a scenario is to accurately categorize the action into its category. In comparison to image processing, more computing power is required in video data processing. In addition to this, a greater number of input parameters are also required for training purpose. Owing to its capability to improve itself, artificial intelligence is growing in popularity for HAR [9]. Since deep learning was brought to the HAR domain [10], it has gotten simpler to extract relevant characteristics from sensor data. The development of deep learning models, like the combination of CNN and LSTM [17], Inception and ResNets [18], is made possible by the evolution of deep learning. These models include CNN [11], [12], Transfer learning models like Inception [13], [14], ResNets [15], and VGG-16 [16].

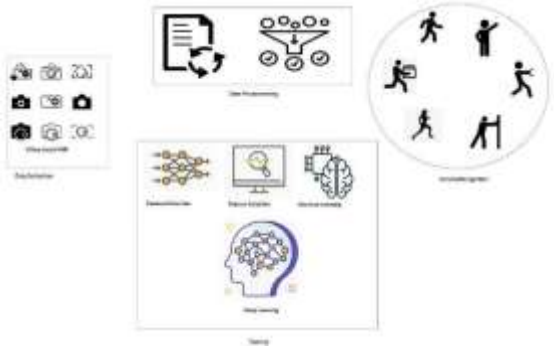

Figure 1. Steps used for Video Based HAR

The way an individual performs an activity is greatly influenced by their habits. The fact that every human being performs actions uniquely makes it extremely challenging to recognize these activities. Another difficult challenge is to develop a strong model to recognize actions using publicly available datasets [2]. A number of benchmark activity recognition datasets [7] have also been made available for further in this field as a result of enormous success of the ImageNet [6] dataset for image processing. However, relatively less research is performed on the

elderly [19]. Therefore, this paper is more focused on elder people daily activity recognition. As older individuals typically follow the same daily routine, caregivers can protect their health by keeping an eye on their activities.

2. Related Study

Nagpal et al. [20] provided a detailed explanation of different human activity recognition methods and research gaps in this field. Additionally, they have explained the methodology to be followed in order to recognize different human actions. The major research gaps addressed by them are data collection and activity recognition methods. The majority of the benchmark datasets that are available were recorded by normal people. Individuals with age 60 or above perform activities in a different manner due to their physical health.

Although there are numerous wearable sensor-based activity recognition methods with remarkable accuracies, as reviewed by Dua et al. [21], most elderly people are reluctant to wear sensors all the time. Methods based on vision have demonstrated astounding precision in identifying a person's daily activities according to Ray et al. [22]. They have emphasized on the context-aware systems that employ vision-based HAR and focused on its diversity using transfer learning methods. Nagpal et al. [23] have performed comparative analysis on the various hybrid deep learning techniques for human actions recognition. It has been analysed that a hybrid deep learning techniques improved the accuracy by optimizing the algorithm.

Over the past ten years, HAR has grown into an active research area for computer vision. Deep learning has dominated most experiments due to its higher performance. However, it is still difficult to extract the crucial features from vision-based data [24]. Complex patterns in an image are extracted using multi-feature extraction techniques [25]. Nagpal et al. [26] developed a multi-feature fusion method for action recognition in which HOG and VGG-16 are used to extract essential features which are then combined to identify the action type.

Wang et al. [28] proposed an abnormal activity detection method for elderly. They have combined CNN-LSTM methods and compared it with the existing methods. Their algorithm performs better as compared to the existing methods in terms of precision, recall and F1-score. Malik et al. [29] discusses the challenges of human action recognition with the conclusion that it is still very difficult to interpret a human action using visual data.

Chakraborty et al. [27] proposed a transfer learning-based action recognition framework on two benchmark datasets and concluded that transfer learning is not sufficient to handle the level of complexity in the HAR classification problem because it is a more general approach.

3. Proposed Model

In machine learning, transfer learning is the process of leveraging knowledge that has already been acquired to address a similar issue [30]. Deep CNNs are particularly useful for transfer learning when the availability of the dataset is limited because CNNs overfits with less amount of data. The enormous amount of annotated data is difficult and expensive to provide, but by increasing the size of the training data, overfitting can be avoided. Transfer learning is helpful in this case to resolve the issue by utilizing the previously trained deep neural network as a foundation [31]. In this paper, VGG-16 architecture is used as a foundation to solve HAR problem for elderly. For feature extraction and image classification, [32] uses an incredibly deep neural network named VGG-16. The RGB-channelled images in the ImageNet database have a fixed size i.e., 224*224. Figure 2 explains the architecture of VGG-16 used in this paper.

The availability of large amounts of data for training deep CNN, such as VGG [32], is one of the most critical aspects in their higher performance. Data augmentation approaches are frequently employed to solve this issue when insufficient data is available to train the network. We have employed a variety of image augmentation techniques to boost the data size because the amount of data available is considerably less in well-known datasets for elderly HAR, such as Stanford40. The benefit of using image augmentation is that numerous images can be produced from a single image to ensure adequate training of the CNN models. Thus, there is less likelihood of over-fitting which results in incorrect classification of target actions. Flips, zooms, and transform are just a few of the often-used augmentation techniques. After pre-processing, the images are passed to VGG-16 architecture for fine tuning. The initial weights of deep neural network are trained on ImageNet dataset and these weights will be passed to CNN for training on Standford40 and Self-made dataset. Finally, SVM classifier is used for classifying the action class.

A GPU is needed for HAR model preparation. The NVIDIA GeForce GTX 1650 with 16GB RAM was used to train this model. The Intel Core i5-10th Gen Processor, clocked at 2.50GHz, was used in the model. The architecture of deep neural networks has been constructed using Keras. ReLU adds non-linearity, hence we employed it as an activation function [33]. If there are any negative values in the feature maps, it will check them and replace them with zero. ReLU aids in the model's nonlinearization.

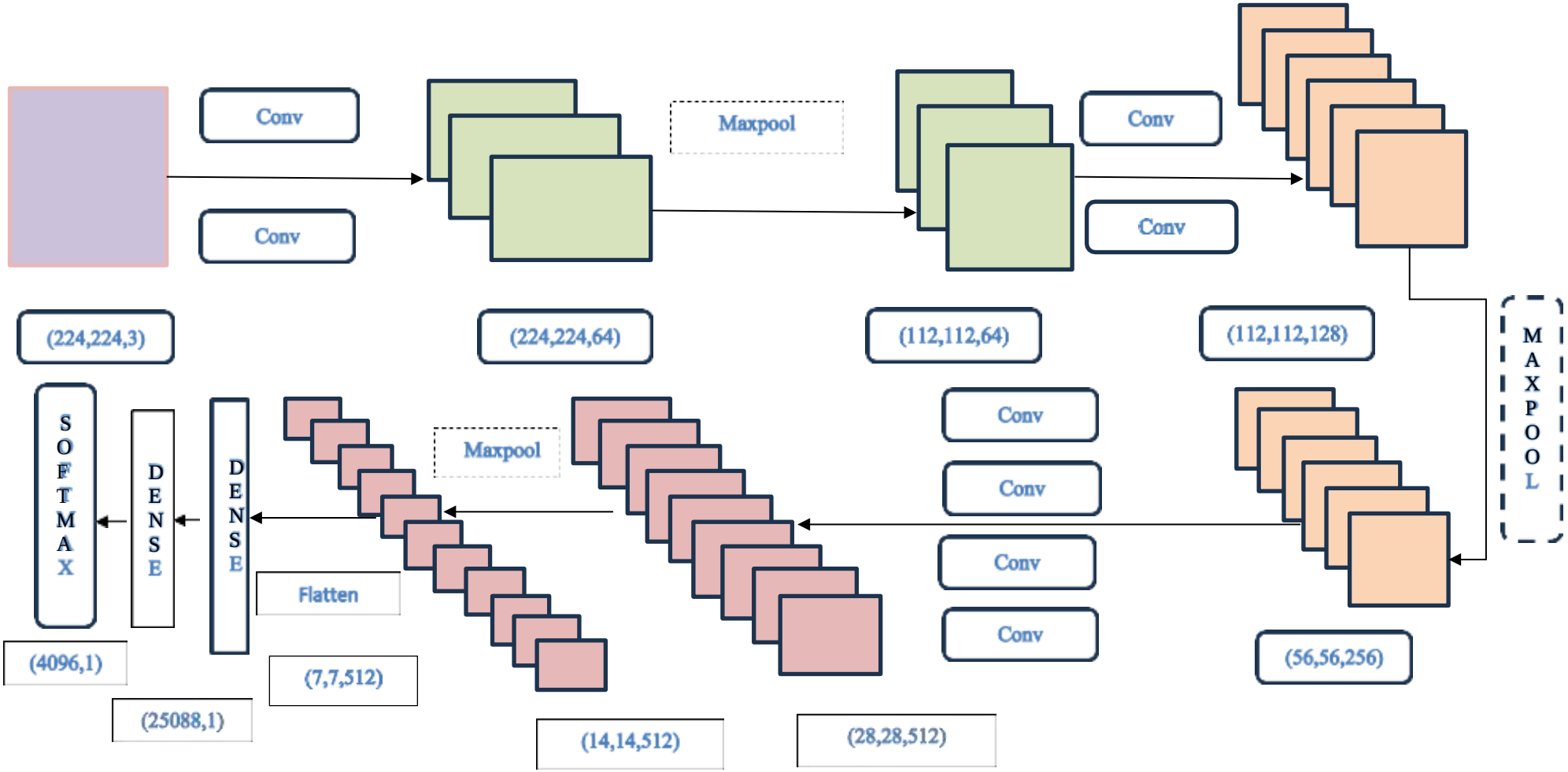

Fig. 1. Architecture of VGG-16

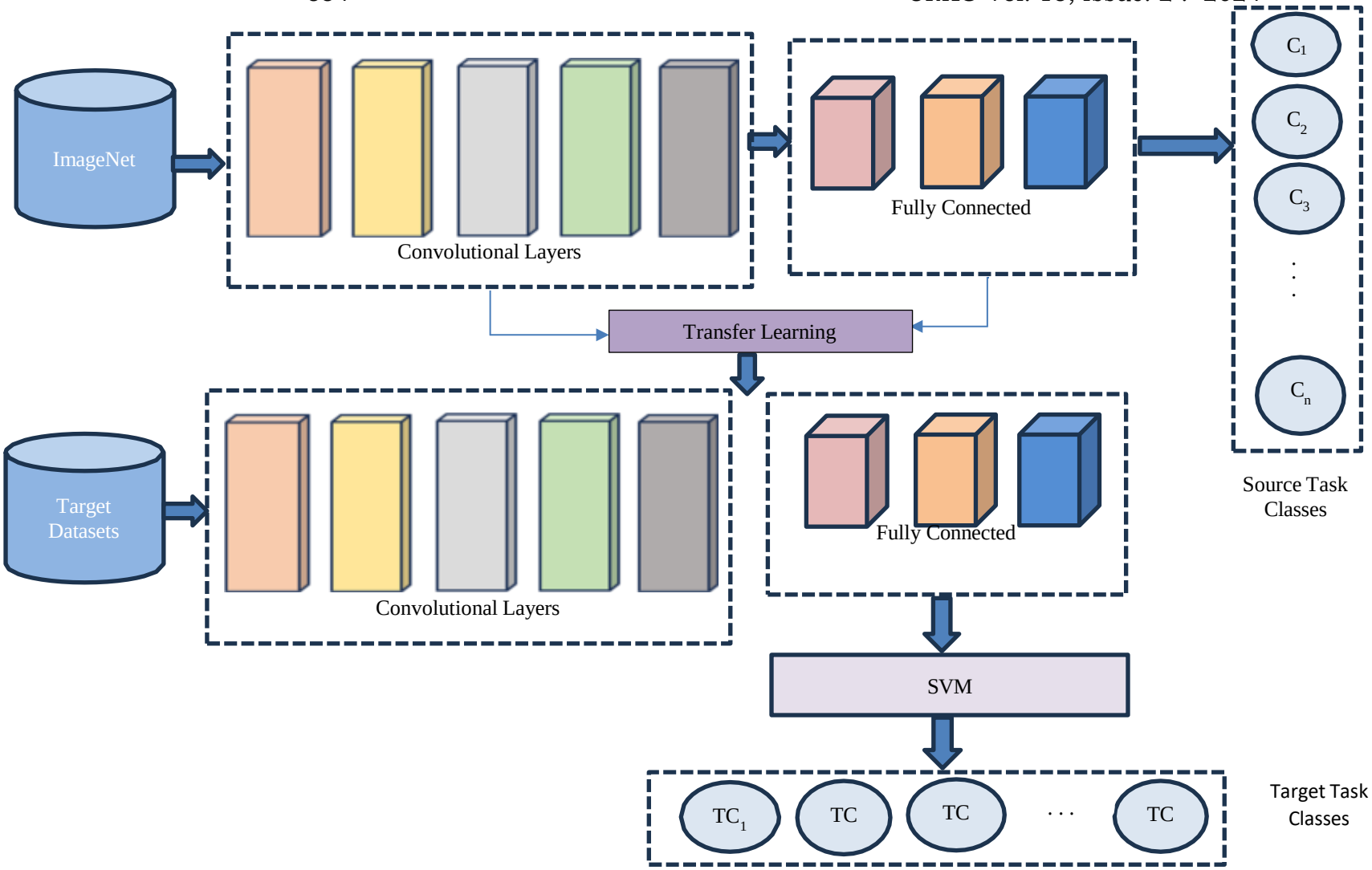

Fig. 2. Proposed Model Architecture

Algorithm: Proposed SVMDNet Algorithm

Input: Image Frames N_f from video V, iteration i=1;
Output: Recognized Activity
Procedure PRE_PROCESSING(V,N_f)
Apply Image Augmentation Techniques
while i < N_f **do**
i) Resize the image i.e., (224,224,3)
ii) Apply geometric transforms
iii) Apply horizontal and vertical flips
iv) Remove noise using Gaussian filter
v) Apply normalization
for i=0 to n-1 **do**

$$I_{(x,y,i)} = \frac{I_{(x,y,i)} - M_{(0,i)}}{SD_{(0,i)}}$$

end
vi) **return** new Image Frames I_f
i+1;
end

Procedure ACTION_RECOGNITION(I_f)
Train the deep neural network on ImageNet Dataset
i) Apply VGG-16 model
ii) To fine-tune, pass the weights through the final dense layers
iii) Apply multi-class SVM classifier for activity recognition
iv) **return** action class TC_1……. TC_n

4. Results and Discussions

4.1. Datasets

The proposed model has been trained and tested on well-known Stanford-40 dataset and on self-made dataset. The details of the dataset are given in the following section:

4.1.1. Stanford-40 Dataset*:* The Stanford 40 action recognition dataset created by [34] is a large and challenging dataset. There are 9532 photos and 40 different action classes in it. As a train-test split, we have used 80–20% of the photos in this case. The Stanford 40 is renowned for its wide range of illustrations that capture the spontaneity of human behaviour in daily life. Images often feature people engaged in activities in a variety of stances and obscured situations, making classification a difficult job. 10 action classes are considered from this dataset such as, brushing teeth, cleaning the floor, reading book, climbing, jumping, walking the dog, running, watching TV, walking, wave hands.

4.1.2. Elderly Dataset*:* The dataset used in this study, which consists of video sequences of 10 classes of continuous human movements, was gathered from elderly volunteers over the age of 60. The clips were captured with a static camera over a uniform environment. For each category of activity, videos of 10 seconds were recorded, which were afterwards converted into image frames for the feature extractor's input. Every second, 30 frames were shown. Approximately 2% of the frames had identical information captured with a very tiny change. As a result, we only kept one frame. 15 frames per second were used as the final frame rate. Carry, clap hands, pick up, pull, push, stand up, sit down, throw, walk, and wave hands are among the activities that are addressed. The percentage of samples in each activity are represented out of the 20270 total photos [26].

4.2. Experminental Results

Two datasets are used to test the proposed model. The confusion matrix for elderly dataset is represented in Figure 4. It can be seen from Figure 5 that the model is performing well on elderly dataset with an overall accuracy of 97%. However, in Figure 5, using Standford-40 dataset, the model is getting confused for the activities which involve human-object interaction. The training and validation accuracy is shown in Figure 6.

4.3. Performance Metrics

In the context of the F-measure, Each class's proper classification bears equal weight (F1). It's possible that the model is evaluated more effectively than the precision because it takes each class's recall and accuracy into consideration while calculating the score. The formulas for calculating F1 score, recall, and precision are listed below.

Table 1. Hyperparameters Used

Hyperparameters	Values
Learning Rate	0.001
Batch Size	20
No. of Epochs	Adaptive
Learning rate	0.1
Output Classes	10 Output Classes brushing teeth, cleaning the floor, reading book, climbing, jumping, walking the dog, running, watching TV, walking, wave hands

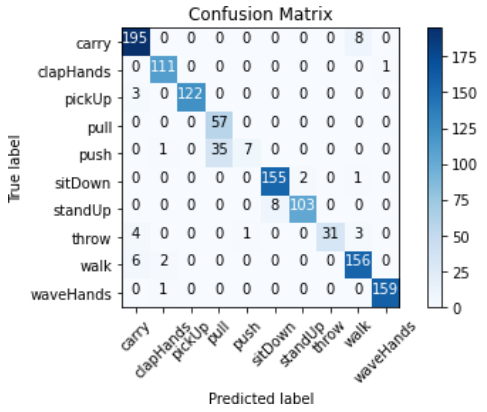

Fig. 4. Confusion Matrix for Elderly Dataset

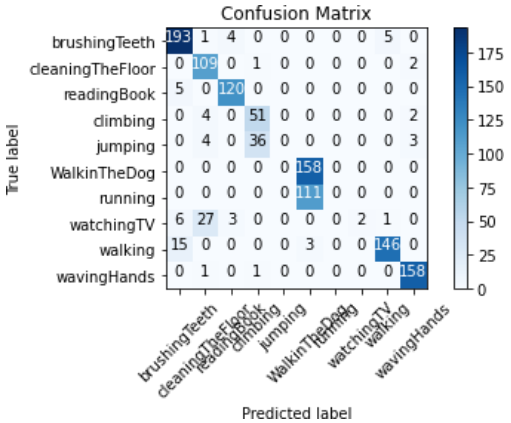

Fig. 5. Confusion Matrix representing human-object interaction using Stanford-40 Dataset

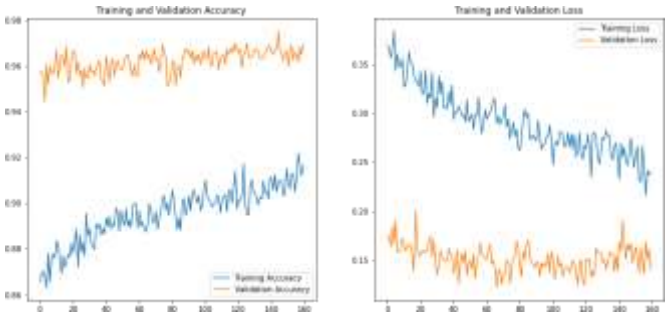

Fig. 6. Training and Validation Accuracy for stanford-40 dataset

Precision:

$$P = \frac{TP}{TP + RP} \quad (1)$$

Recall:

$$R = \frac{TP}{TP + PN} \quad (2)$$

F1-Score:

$$F1 = \sum_i 2 * w_i \frac{Precision.recall_i}{Precision + recall_i} \quad (3)$$

where the associated true and false positive counts are denoted by TP and FP, respectively, and FN is the equivalent count of true negatives. Inequality between classes is addressed by weighting. In case, wi = ni/N, then ni represents the total samples as determined by the sample proportion, and N represents the total number of samples.

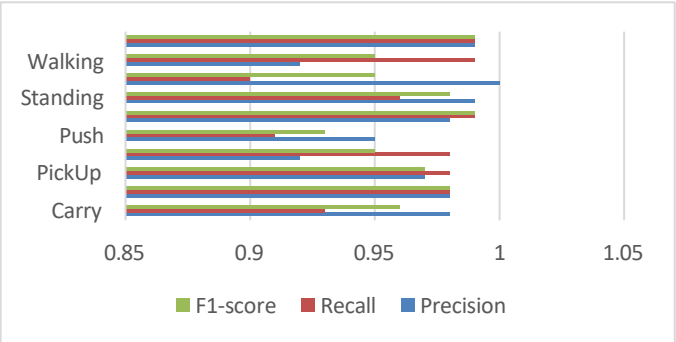

Fig. 7. Performance of the proposed model on Elderly Dataset

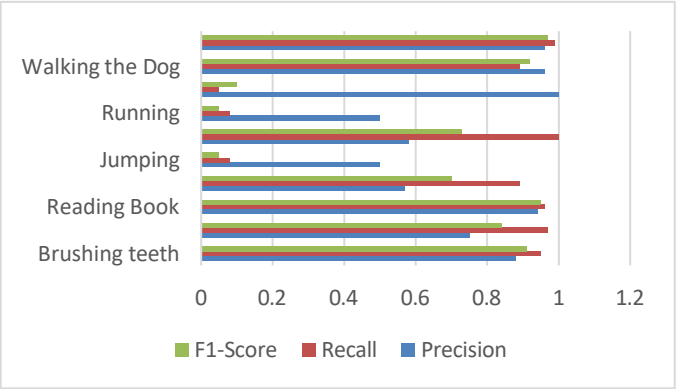

Fig. 8. Performance of the proposed model on Standford-40 Dataset

5. Conclusion

The current study proposes a novel transfer learning-based architecture called SVMDnet. which uses pre trained deep neural network and SVM to classify elderly actions. It has been demonstrated that using transfer learning, we can successfully apply previously learned knowledge to learning a new task with a small training dataset. Transfer learning is very useful when there is not enough data to fully train the deep learning model. Additionally, the proposed method works with RGB images, negating the need for manual feature extraction and handmade representation-based methods. The effectiveness of the suggested strategy was evaluated using the well-known Stanford-40 dataset to identify human-object interactions and on the primary dataset to identify elder people's daily routine actions. SVMDnet attained accuracy rates of 80% and 97%, respectively. We hope to expand on this approach in the future to handle more intricate datasets and interpersonal interactions.

References

[1] Chakraborty, S., Mondal, R., Singh, P. K., Sarkar, R., & Bhattacharjee, D. (2021). Transfer learning with fine tuning for human action recognition from still images. *Multimedia Tools and Applications*, *80*, 20547-20578.

[2] Sharma, V., Gupta, M., Pandey, A. K., Mishra, D., & Kumar, A. (2022). A review of deep learning-based human activity recognition on benchmark video datasets. *Applied Artificial Intelligence*, *36*(1), 2093705.

[3] Putra, P. U., Shima, K., & Shimatani, K. (2022). A deep neural network model for multi-view human activity recognition. PloS one, 17(1), e0262181.

[4] Gupta, S., & Saini, A. K. (2013). Information system security and risk management: Issues and impact on organizations. Global Journal of Enterprise Information System, 5(1), 31-35.

[5] Badawi, A. A., Al-Kabbany, A., & Shaban, H. (2018, December). Multimodal human activity recognition from wearable inertial sensors using machine learning. In 2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES) (pp. 402-407). IEEE.

[6] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[7] Sun, L., Jia, K., Chen, K., Yeung, D. Y., Shi, B. E., & Savarese, S. (2017). Lattice long short-term memory for human action recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2147-2156).

[8] Jindal, S., Sachdeva, M., & Kushwaha, A. K. S. (2022). Deep Learning for Video Based Human Activity Recognition: Review and Recent Developments. In Proceedings of International Conference on Computational Intelligence and Emerging Power System: ICCIPS 2021.

[9] Nguyen, T. H. C., Nebel, J. C., & Florez-Revuelta, F. (2016). Recognition of activities of daily living with egocentric vision: A review. Sensors, 16(1), 72.

[10] Gupta, N., Gupta, S. K., Pathak, R. K., Jain, V., Rashidi, P., & Suri, J. S. (2022). Human activity recognition in artificial intelligence framework: a narrative review. Artificial intelligence review, 1-54.

[11] Gul, M. A., Yousaf, M. H., Nawaz, S., Ur Rehman, Z., & Kim, H. (2020). Patient monitoring by abnormal human activity recognition based on CNN architecture. Electronics, 9(12), 1993.

[12] Khater, S., Hadhoud, M., & Fayek, M. B. (2022). A novel human activity recognition architecture: using residual inception ConvLSTM layer. Journal of Engineering and Applied Science, 69(1), 1-16.

[13] Li, Y., & Wang, L. (2022). Human Activity Recognition Based on Residual Network and BiLSTM. Sensors, 22(2), 635.