Vision Transformer-based Model for Human Action Recognition in Still Images

Divya Rani R[1], Prabhakar C J[2]

[1]Department of Computer Science, Kuvempu University Shivamogga, Karnataka, India,
Email: divyarsbg@gmail.com
[2]Department of Computer Science, Kuvempu University, Shivamogga, Karnataka, India,
Email: psajjan@yahoo.com

Received: 15.07.2024 Revised: 13.08.2024 Accepted: 12.09.2024

ABSTRACT

Human action recognition is a critical task in computer vision, enabling systems to understand and interpret human actions from images.Action recognition in still images presents a unique challenge,as traditional methods often rely on temporal information that is absent in static images. This study explores the advantage of Vision Transformers (ViTs) for recognizing human actions in still images,exploiting their ability to capture complex patterns and relationships in visual data.We propose a robust method for recognizing human actions in still images, that employs spatial attention mechanisms to effectively highlight relevant features associated with various human poses and contexts and effectively addresses the challenges such as occlusion and varying poses by employing self-attention mechanism that focus on key pose and contextual cues. We conduct extensive experiments using benchmark datasets: Stanford40 and PASCAL VOC 2012 Action, the proposed the model achieved an impressive accuracy of 97.4% for Stanford40 and 94.8% for PASCAL VOC 2012 Action dataset. Experimental results demonstrates that the proposed method achieves SOTA performance on both still image datasets. The high accuracy suggests that the ViT model can generalize well across different action categories, even when the dataset includes variations in human posture, background, and scene complexity.

Keywords: Human action recognition, HAR, Vision transformer, ViT, Transformer encoder

1. INTRODUCTION

Human action recognition (HAR) in still images is to recognize the actions performed in a single image frame. HARhas received maximum attention in the areas of computer vision due to its massive applications, such as intelligent surveillance, robotics, human computer interactions (HCI), image retrieval, annotations and health care. With all these applications, the action recognition problem faces numerous challenges, such as cluttered backgrounds, occlusion, illumination changes, view-point variation, inter-class and intra-class similarities. In addition to these challenges, action recognition in still images is inherently more difficult than in video-based action recognition because spatio-temporal features are crucial for accurately characterizing actions. In still images, temporal information is not available, which is essential for describing actions. Many techniques are proposed in the literature to solve the issues associated with recognizing human actionsusing traditional handcrafted features techniques and deep learning-based techniques. Deep learning-based techniques are widely adopted thanhandcrafted feature-based techniques due to their superior performance in recognizing actions.For vision-based action recognition problems, convolutional neural networks (CNN)are extensively applied[1] [2]. Other deep learning models such as, 3D CNN, two-stream CNN, RNN, LSTM, generative models, Inception models,and hybrid models are employed for HAR [3] [4] [5] [6].One limitationof deep CNNs is that their difficulty in capturing relationship between different regions of an image, which are crucial for action recognition. This failure to find the relationshipscan lead to misclassification.

With the improvement of technology, the techniques used for action recognition have gradually migrated towards Vision transformer (ViT) models. Vision Transformers (ViTs) [7] are gaining more popularity in Computer Vision producing very promising results compared to the results produced by CNN-based models.ViTs were first introduced in Natural Language Processing to handle sequences of text efficiently.The transformer models produced promising results when used for various computer vision tasks, leading to the creation of Vision transformers.They have the capability to capture long-range dependencies and contextual relationships. Transformer model contains self-attention and feed-forward layers.Self-attention mechanism helps to capture both spatial and temporal dependencies. ViTsextract semantic relationships in images and operate on patches extracted from input images rather than processing whole image as in CNN.The presence of attention mechanism makes the transformer model to focus on more significant regions than irrelevant parts. A number of approaches have been proposed using ViTs for recognition tasks such as action recognition, face recognition, biometric recognition, and gesture recognition [7] [8] [9] [10].

In this paper, we have used Vision Transformer model for recognizing human actions in still images.Initially, we used YOLOv8[11] detection model to detect human part in the input image. The human part detected within the bounding box is extracted and is split into equal number of patches. These patches are flattened and converted into a one-dimensional vector by linear embedding. In order to keep the sequential order of patches, position values are embedded with generated vector along with a token class. The position embedded tokens are passed to Vision Encoder system for recognitionof human actions.To analyze the efficacy of the proposed model we used standard still image action datasets such as, Stanford40 [12],and PASCAL VOC 2012 [13] dataset. The proposed model achieves a recognition accuracy of 97.4% for Stanford40 dataset, and 94.8% for PASCAL VOC 2012action dataset, respectively.

The paper is structured as follows, Introduction toHAR and vision transformersare presented in Section 1. The similar works to HAR using ViTsare reviewed in Section 2. The proposed method for HAR using ViT is given in Section 3. The experimentation carried out using still image action datasets and analysis of results are provided in Section 4. In Section 5 discussion about proposed method based on ViT is presented. The last Section 6concludes the paper.

2. Related Work

Action recognition in videos is relatively a well-studied research area, whereas still image-based action recognition has received less attention. Action recognition in still images is a challenging task because it does not contain temporal information which is needed to describe an action.Recent advancements in human activity recognition (HAR) in still images have increasingly incorporated Vision Transformers (ViTs), offering new insights into capturing complex human actions from static visuals. Traditional approaches predominantly relied on convolutional neural networks (CNNs), which struggled with spatial relationships and detailed pose variations. ViTs, with their self-attention mechanisms, excel in modeling global contexts and intricate features, thus enhancing recognition capabilities. Several research works for recognizing human actions using ViTs are presented in this section.

2.1 ViT for Image based HAR

Only few research works have been proposed in the literature for recognizing human actions in still images using ViTs. Fan et al. [14] proposed a model for recognizing human actions in images based on Swin transformer called SIFAR (Super Image for Action Recognition). First the 3D video data is converted to video frames and arranged them to form a super image according to pre-defined spatial layout. Action performed in the video is predicted by using image classifier. Chen et al. [15] designed a multi-scale behavioral feature extraction model called Swin-Fusion for recognizing human actions in still images based on Swin-Transformer by fusing multi-scale behavioral features. The Swin-Fusion model combines the Swin-Transformer and Feature Pyramid Network architectures.

Hosseyni et al. [16] proposed a model based on Vision Transformer called ConViT, to recognize human actions in still images. This ConViT comprises of ResNet50 architecture to extract spatial features and two Vision Transformers to extract the relationship between different parts of an image.

2.2 ViT for Video based HAR

Many research works for recognizing human actions in videos using ViTs are proposed in the literature. Anurag et al. [17] proposed a model for HAR using ViT. It works on 3D volumes instead of frames. It considers Tubelets Embedding which preserves contextual time data in the video. It first extracts volumes from the video clips, which includes both frame and time stamp patches. Then, the extracted volumes are flattened to generate tokens. Hussain et al. [18] proposed an approach for HAR, which extracts frame level features from pretrained Vision Transformer. These extracted features are passed to multilayer LSTM to capture long-range dependencies of the action. Gedas et al. [19] extended transformer models to extract spatio-temporal data in videos by incorporating both spatial and temporal attention. To calculate similarity measures for all pairs of tokens, scalable self-attention designs over the space-time volume are used.

A modified transformer for video action recognition, called Action Transformer is proposed by Girdhar et al [20], to classify the action of a person of interest. It is designed to detect all persons and classify all the actions performed all persons in a video. The transformer attends to hand and face regions, which are the efficient features for discriminating an action. Mazzia et al. [21] proposed a pose-based action recognition model, by introducing a model called Action Transformer (AcT), which combines the functionalities of convolutional, recurrent, and attention models.

The thorough review of literature highlights the effectiveness of Vision Transformers in capturing complex spatial relationships and contextual differences that are essential for accurately recognizing human actions in static images.

3. Proposed methodology

This section describes the general framework of our proposed methodology for human action recognitionin still images using Vision Transformers.An overview of the proposed method for HAR using ViT is illustrated in Fig. 1.

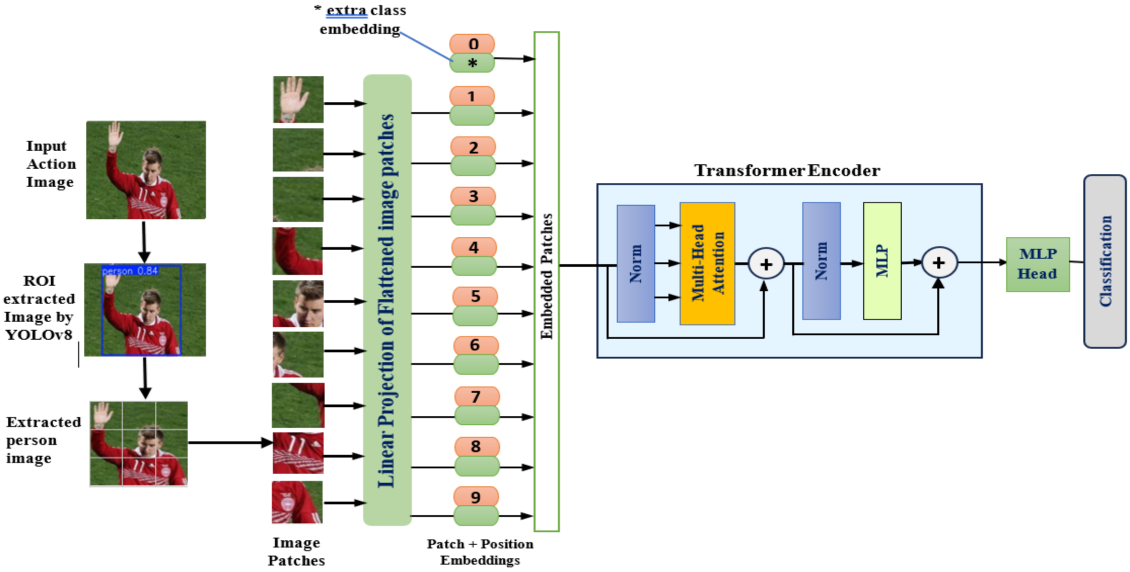


Fig. 1. Block diagram of the proposed method for HAR using ViT

To recognize particular actions performed by a human being, we proposed a method using Vision Transformers. Initially, the input imageI from standard benchmark dataset is resized to a standard size. The input action image may consist of human part, objects, and other backgrounddetails. To detect and extract human part from input action image, we constructed region of interest (ROI) by employing the popular object detection algorithm YOLOv8, which finds the region of interest in the image. YOLOv8 detects humans and objects present in the images by drawing bounding box around them. Thecorresponding class label which the object belongs to is mentioned as annotated text in the bounding box.The ROI extracted images are further considered to extract coordinates of bounding box and class label. The bounding box coordinates are used to extract human part and the class label information is used to check whether the object under consideration is human or not. The sample results obtained by the YOLOv8 model for the Stanford40 action dataset are given in Fig. 2.

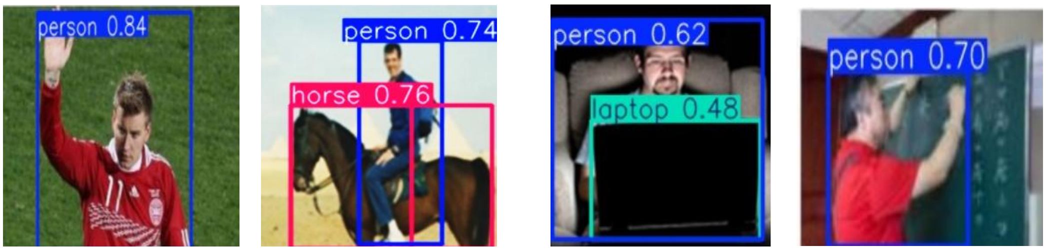


Fig. 2. ROI extracted sample imagesof Stanford action datasetby YOLOv8

In the scenarios like if an action involving objects for example, playing guitar, golfing, horse riding, etc., then only the bounding box specifying human is not enough to describe the action. The solution for this kind of actions is to consider bounding box of both human and object. From the output images obtained by YOLOv8, first we need to consider class labels and object regions. If there exists any object class associated with human class, then we need to merge the bounding boxes of human and objectto generate a single bounding box that encompasses both. The procedure for merging the bounding boxes is explained in detail below:

If human bounding box has the coordinates (hx1, hy1, hx2,hy2) and object bounding boxes are (bx1, by1, bx2,by2) the bounding box corresponding to union regions are obtained by using the following equations:

$$\mathrm{ux1} = \min(\mathrm{hx1}, \mathrm{bx1}) \quad (1)$$

$$\mathrm{uy1} = \min(\mathrm{hy1}, \mathrm{by1}) \quad (2)$$

$$\mathrm{ux2} = \max(\mathrm{hx2}, \mathrm{bx2}) \quad (3)$$

$$\mathrm{uy2} = \max(\mathrm{hy2}, \mathrm{by2}) \quad (4)$$

The above equations specify the bounding box coordinates ofhuman with object. The coordinates corresponding to new bounding box containing both human and object is (ux1,uy1,ux2,uy2).The sample images after merging the bounding boxes of human and object are provided in Fig. 3. The Fig. 3(a) has human and horse, bounding boxesare generated for both objects by YOLOv8, then a single bounding box is generated using equations Eq. 1 to 4, which consist of both human and horse within single bounding box. Similarly, in Fig. 3(b) shows person with laptop, which is identified by a single bounding box and in Fig. 3(c) shows bounding box consisting of person and a book.

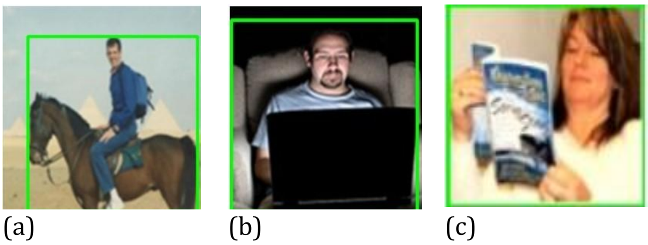


Fig. 3. Sample images of Stanford40 dataset with bounding box containing bothhuman and object

The human part and in some cases human with object part are extracted from the input imageare considered as input for the ViT model.In ViT model, first the input image is divided into number of fixed-size patches. The patches are then flattened and converted to one-dimensional vector by linear projection. In order to keep the sequential order of patches, position values are embedded to each flattened patches along with class token. The position embedded tokens are passed to Vision Encoder system for classification of human actions through fully connected layers. The whole process is explained below.

The human part$\mathrm{h_b}$extracted from the bounding box is considered as input for the ViT model,$\mathrm{h_b} \in \mathrm{R}^{\mathrm{H \times W \times C}}$and is divided into 'N' number of 2D patches$\mathrm{P_n} \in \mathrm{R}^{\mathrm{n \times (S^2.C)}}$.Here, H, W, and C areheight,width,and number of channels of the human body $\mathrm{h_b}$extracted from bounding box. The component (S.S) is the size of each image patch, and$\mathrm{N} = \mathrm{HW/S^2}$ is the total number of patches. The patch size is generally 16x16 or 32x32, where the small patch size is able to capture longer sequences. In our proposed method, we have considered a patch size of 16x16. The sequence of patches obtained from the image are linearly projected into a 1D vector Vusing a learnable embedded matrix E. Then, these patch embeddings are prepended by aclass token$\mathrm{t_{class}}$. The patch embedded tokens are not in sequential order, as they are in the original human image$\mathrm{h_b}$. To maintain the sequential order of patch embedded tokens, position information $\mathrm{E_{pos}}$ is appended with each patch embedded tokens. Position embeddings maintain the spatial order of the originalimage. The result of the patch embeddings with class token and position embeddings forms a vector, $V_0$and is mathematically represented as in equation (5).

$$\mathrm{V_0} = [\mathrm{t_{class}}\,; \mathrm{P_1 E}, \mathrm{P_2 E}, \ldots \ldots, \mathrm{P_n E}] + \mathrm{E_{pos}}\,, \mathrm{E} \in \mathrm{R}^{(\mathrm{S^2.C}) \times \mathrm{d}}, \mathrm{E} \in \mathrm{R}^{(\mathrm{n+1}) \times \mathrm{d}} (5)$$

The resultant patch embedded vector $V_0$is forwarded to transformer encoder module, which consists of componentslike,Layer Normalization block, Multi-Head Self Attention (MSA), and Multilayer Perceptron (MLP) blocks.The MSA is the central component of the transformer model which finds the most and least important patches and discards the least important patches from the input patch sequences. The mathematical representation of the MSA and MLP is given by the equations6 and 7,

$$\mathrm{v_l^1} = \mathrm{MSA}\big(\mathrm{LN}(\mathrm{v_{l-1}})\big) + \mathrm{v_{l-1}}, \qquad \mathrm{l} = 1{,}2..\mathrm{L} \quad (6)$$

$$\mathrm{v_l} = \mathrm{MSA}\left(\mathrm{LN}(\mathrm{v_l^1})\right) + \mathrm{v_l^1}, \; \mathrm{l} = 1{,}2..\mathrm{L} \quad (7)$$

The MSA layer is composed of linear, self-attention, and concatenation layers to combine the output from multiple heads. The MSA consists of Q(query), K(Key), and V(Value)components which are calculated by

$$\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{V}.\,\mathrm{softmax}(\frac{\mathrm{QK^T}}{\mathrm{d_k}}) \quad (8)$$

where, d_k is the dimensionality of the key vectors and it is used for gradient stability and softmax() is used to find probability distributions.

In the last layer of the encoder, the first tokent_{class} is processed with MLP head to provide predicted class label. The MLP head contains linear layers and RELU non-linearity and dropout layers.The output of encoder is passed to classifier model to classify the action performed in the input image.

4. Experiments and Result Analysis

In this section, we experimentally evaluated the proposed model for action recognition using two standard still image-based action datasets, Stanford40 dataset [13] and PASCAL VOC 2012 [14] datasets. These datasets are widely used for evaluating action recognition models, offering diverse action categories and real-world complexity in visual data. There are different variants of ViT models such as ViT-Base, ViT-Large, and ViT-Huge. According to literature, ViT-Base-16 has given promising results compared to other deep learning models using benchmark datasets.For experimentation, we make use of ViT-Base-16 model with patch size 16 x 16. The performance evaluation metrics used are Accuracy, Precision, and Recall. The results obtained by the experiments on these datasets are analyzed.

4.1 Experimental Setup

In this section, we outline the experimental setup utilized for action recognition in still images, by incorporating Vision Transformers for action recognition and the YOLOv8 model for human detection. This approach allows for an effective pipeline that ensures accurate recognition of actions by first identifying human part within the images.

The Vision Transformer (ViT) model is utilized for the action recognition task. The model consists of transformer encoder layer that process image patches as tokens, capturing long-range dependencies and spatial relationships within the visual input. Pretrained ViT models are fine-tuned on the specific action recognition task using the popular action datasets Stanford40 and PASCAL VOC 2012 action datasets. TheViT model generally requires huge datasets to train the model.The still image action datasets consist of a smaller numberof images, which are not enough to train the ViT models. For smaller datasets, ViT models cause overfitting issues and a decrease in performance. In order to increase the size of the datasets, we employed data augmentation techniques, which artificially increases the training dataset to address the overfitting issue of the model.

The data augmentation techniques adopted in our experiment are rotation, horizontal and vertical flipping, and random cropping of original images. Slight rotations (e.g., ±15 degrees) are applied to make the model invariant to orientation changes. Flipping images horizontally and vertically simulates different viewing angles and helps in recognizing actions from both left and right perspectives. Random cropping of images to different sizes helps the model to focus on various parts of the image. These augmentations are applied in real time during training, resulting in a diverse dataset that increases the model's ability to generalize to unseen data.

Initially the images of both datasetsare resized toa fixed dimension 224 x 224pixels.The resized images are passed to YOLOv8 model, which detects human and other objects present in the action image by generating bounding boxes.The output of the YOLOv8 model includes bounding boxes and confidence scores for detected humans, which serve as inputs for the subsequent action recognition process.Additionally, data augmentation techniques such as rotation, horizontal and vertical flipping, and random cropping are applied to the detected human image by YOLOv8. The generated images are passed as input for ViT model. Some actions like playing guitar, reading, fishing, etc. involves objects along with human part. In such cases, both human and object both are extracted by merging the bounding boxes of both human and object. The image present in the merged bounding box is extracted and passed to the ViT model. In ViT model, the input image is split into number of 16x16 sized patches, each patch is linearly embedded by positional embeddings, and are fed into the transformer encoder. Along with patch and position embeddings, class token is also considered to perform classification task. The MSA block in transformer encoder finds the most important patches and discards the least important patches from the input patch sequences. The MLP performs classification task and assigns corresponding class label for the input image.

4.1.1 Parameters considered

For training, we considered standard cross-entropy loss for classification. The model is optimized using Stochastic Gradient Descent optimizer with initial learning rate of 0.001 and is reduced by a factor of 0.1 for every 10 epochs. The model is trained for a total of 50 epochs, with early stopping applied if the validation accuracy plateaus. The training and test sets are split according to the standard splits provided by each dataset. We make use of ViT-Base-16 model with patch size 16 x 16.The ViT model is fine-tuned on the specific action recognition datasets by training only the last few layers or the entire network, depending on the experimental setup.

4.2 Experimental Results and Analysis

4.2.1 Experimentation on Stanford40 action dataset:

The Stanford40 dataset [12]is used for training the still image based human action recognition models. It contains 40 different daily life human actions such as jumping, running, fishing, cleaning the floor, brushing the teeth, reading book, etc. Sample images of Stanford40 dataset are provided in Fig. 4.In this dataset, there are total 9532 images and these are divided into two sets: 4000 images for training set and 5352 images for testing set. Each class contains 180 to 300 images. The images in the dataset are single-labeled and include a small number of distinct individuals per image. The challenges associated with this dataset are various backgrounds, occlusions, variation in cloth and appearance.

Fig 4. Samples action images of Stanford40 dataset [12]

The ViT model is evaluated on the Stanford40 dataset, and its performance is reported alongside the baseline methods. The ViT model achieved an overall accuracy of 97.4% on the Stanford40 dataset, which indicates its strong performance in recognizing human actions from still images. This result demonstrates the model's ability to effectively capture global context and spatial relationships, crucial for action recognition in diverse settings.To further analyze the performance, a confusion matrix was generated and is illustrated in Fig. 5, which visualizes the model's classification results across the 40 action categories. The confusion matrix provides insight into which action categories were correctly classified and which were misclassified, highlighting areas where the model performed well or struggled.

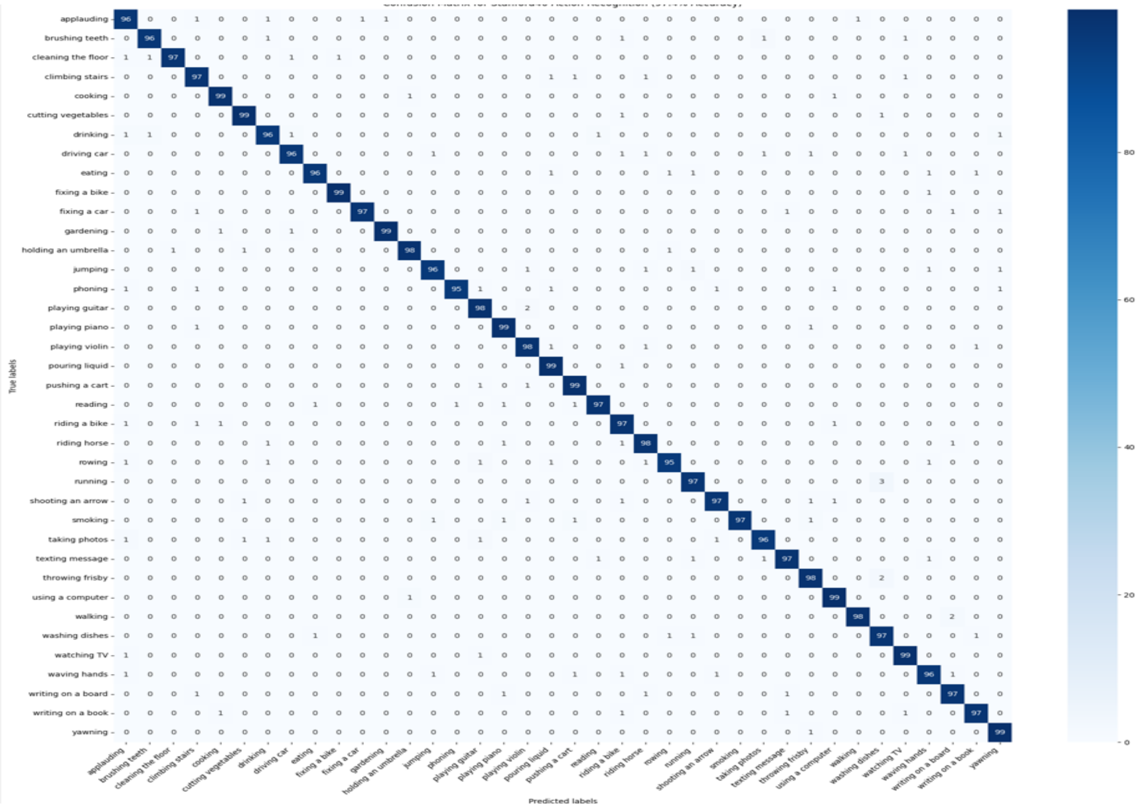


Fig 5. Confusion matrix obtained for the Stanford40 dataset

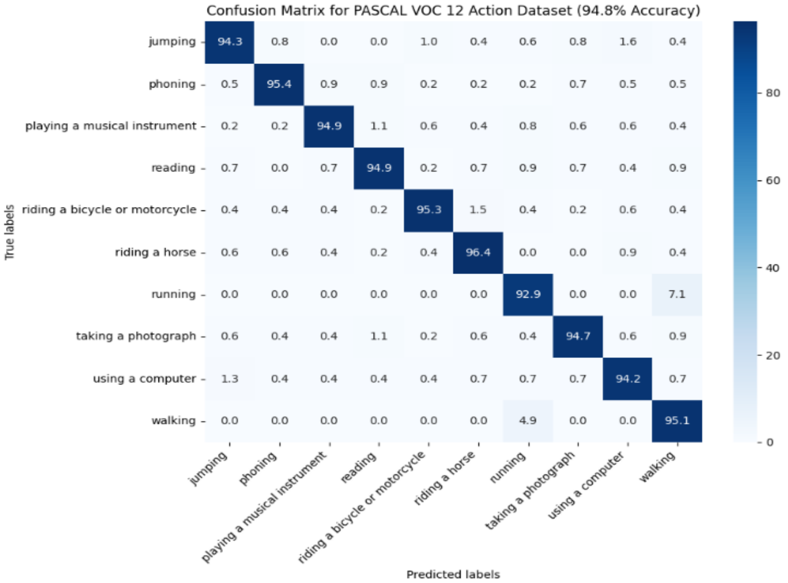
From the confusion matrix it can be observed that, actions like ‘reading’, ‘playing drums’ and ‘pizza tossing’ were correctly classified with high accuracy due to their clear and distinct visual characteristics.The model faced some challenges in distinguishing between actions with visually similar movements, such as ‘running’ versus ‘jogging’ or ‘cricket bowling’ versus ‘javelin’ or ‘playing guitar’ versus ‘playing violin’. These fine-grained distinctions are often difficult to identify in still images without temporal context or additional cues.The confusion matrix also revealed that the model struggled with action categories where occlusion or overlap with other objects occurred, leading to some misclassifications.

4.2.2 Experimentation on PASCAL VOC 2012 action dataset

The PASCAL VOC 2012 Action dataset [13] is although primarily designed for object detection and segmentation tasks, PASCAL VOC 2012 contains annotated images for 20 object classes and is often adapted for action recognition tasks. The dataset consists of 10 different action categories with total 4588 action images. We adapt this dataset for action recognition by selecting relevant images and action-related annotationsSample images of Stanford40 dataset are provided in Fig. 6. The dataset offers real-world complexity with occlusions, cluttered backgrounds, and multiple objects interacting in the scene, making it suitable for testing models in dynamic settings. The dataset is divided into two sets: training set includes 2992 images, and the remaining 1596 are considered for testing.

Fig 6. Samples action images of PASCAL VOC 2012Action dataset [13]

In our experimentation with the PASCAL VOC 2012 action dataset, the proposed ViT model for action recognition has achieved an accuracy of 94.8%. To gain a deeper understanding of the model's performance, we generated a confusion matrix as shown in Fig. 7. that visualizes the classification results across the different action categories. The confusion matrix revealed that, the model performed well on action categories that are easily distinguishable, such as "eating," "playing guitar," or "holding an umbrella." These actions often have clear visual cues that the ViT model can capture, making classification more straightforward. But for some actions the model generated misclassifications, particularly in actions like "running" vs. "walking" or "biking" vs. "riding horse." These actions share similar body movements or postures, which made it difficult for the model to differentiate them based solely on the still images, suggesting that temporal context might be beneficial for fine-grained classification.



	jumping	phoning	playing a musical instrument	reading	riding a bicycle or motorcycle	riding a horse	running	taking a photograph	using a computer	walking
jumping	94.3	0.8	0.0	0.0	1.0	0.4	0.6	0.8	1.6	0.4
phoning	0.5	95.4	0.9	0.9	0.2	0.2	0.2	0.7	0.5	0.5
playing a musical instrument	0.2	0.2	94.9	1.1	0.6	0.4	0.8	0.6	0.6	0.4
reading	0.7	0.0	0.7	94.9	0.2	0.7	0.9	0.7	0.4	0.9
riding a bicycle or motorcycle	0.4	0.4	0.4	0.2	95.3	1.5	0.4	0.2	0.6	0.4
riding a horse	0.6	0.6	0.4	0.2	0.4	96.4	0.0	0.0	0.9	0.4
running	0.0	0.0	0.0	0.0	0.0	0.0	92.9	0.0	0.0	7.1
taking a photograph	0.6	0.4	0.4	1.1	0.2	0.6	0.4	94.7	0.6	0.9
using a computer	1.3	0.4	0.4	0.4	0.4	0.7	0.7	0.7	94.2	0.7
walking	0.0	0.0	0.0	0.0	0.0	0.0	4.9	0.0	0.0	95.1

Fig 7. Confusion matrix obtained for the PASCAL VOC 2012 Action dataset

The performance of the ViT model on both the Stanford40 and PASCAL VOC 2012 Action datasets is evaluated based on accuracy. For Stanford40, we expect high accuracy due to the relatively clean and action-focused nature of the dataset, while for PASCAL VOC, the model's ability to handle complex, cluttered scenes with multiple objects and actions will be a key factor.

4.3 Performance Comparison with Existing methods

In order to evaluate the advantage of our proposed method, we made performance comparison of popular image-based HAR techniques using DNNs on still image datasets. These models are typically evaluated on datasets like ImageNet, CUB-200-2011, or Stanford40, which contain labeled images representing different activities or actions.

Table 1: Performance Comparison of Image-Based HAR Techniques Using DNNs on Still Image Datasets

Model	**Accuracy (approx.)**	**Limitations**
2D CNN (e.g., AlexNet, VGGNet)	**~80-90%** for simpler activities (e.g., walking, running)	Struggles with complex actions that need temporal or contextual reasoning
ResNet (Residual Networks)	**~85-95%** for activities requiring deep feature extraction (e.g., sports, complex motions)	High memory consumption, slow training for very deep networks
Inception Networks	**~85-95%** for activities with diverse objects or backgrounds (e.g., sports, complex interactions)	More complex architecture, higher computational overhead
Attention Mechanisms (e.g., SENet, CBAM)	**~90-96%** for recognizing complex activities with varying attention to regions (e.g., dancing, sports)	Additional computation for attention layers, requires large datasets
Vision Transformers (ViT)	**~92-98%** for large-scale, detailed datasets (e.g., ImageNet, Stanford40)	Very high computational cost, needs large labeled data for training
Hybrid CNN + Attention	**~90-98%** for complex, dynamic actions	Increased complexity, requires well-tuned attention mechanisms

5. DISCUSSION

In this research work, we explored the effectiveness of the Vision Transformer (ViT) model for human action recognitionin still images by employing ViT-Base-16 model. The ViT model learns the semantic relationship between differentparts of the image. ViT's ability to capture long-range dependencies through its self-attention mechanism allowed it to outperform traditional CNN-based models, such as ResNet and VGG, especially in recognizing actions amidst varying poses, lighting conditions,and backgrounds. However, despite these strengths, some challenges remained in achieving high accuracy for all action categories. One key limitation was the difficulty in distinguishing between visually similar actions, such as "running" versus "jogging," or "standing" versus "sitting." These types of fine-grained distinctions can be challenging in still images, where the lack of temporal information prevents the model from leveraging motion cues that are often helpful for disambiguating between similar actions. While ViT excels in capturing spatial relationships within a single frame, this still-image approach has inherent limitations when it comes to dynamic actions that involve rapid motion or subtle shifts in body posture.

Another challenge is the need for large datasets to fully exploit the potential of ViT models. While ViT outperformed CNN-based models on some datasets, its performance could be limited by the size and diversity of the training data. In this study, we observed that on smaller datasets, such as Stanford40, the model's performance plateaued, suggesting that more data would be beneficial to help the model learn more robust features for action recognition. Furthermore, the model sometimes struggled with handling occlusions or overlapping objects in complex scenes, a common issue in real-world datasets like PASCAL VOC 2012. While ViT's self-attention mechanism is better equipped to handle some of these challenges compared to CNNs, further improvements could be made to enhance its robustness in such scenarios.

Overall, the ViT model showed significant promise for action recognition in still images, and future work should focus on expanding the model's capability to handle motion and leverage more data could further improve its performance and make it a more robust solution for real-world human action recognition tasks.

6. CONCLUSION

In this research work, we have proposed a method forrecognizing human actionsin still images by employing ViT-Base-16 model. The ViT model learns the semantic relationship between different areas of the image, by extracting spatial features from the input image.Its self-attention mechanism allows for a deeper understanding of spatial relationships in human actions, making it particularly effective in complex environments. The proposed model for action recognition demonstrated impressive performance for human action recognition, achieving 97.4% accuracy on the Stanford40 dataset and 94.8% accuracy on the PASCAL VOC 2012 action dataset. These results highlight ViT's capability to effectively capture long-range dependencies and spatial relationships within images, which are essential for recognizing complex human actions. The model outperformed traditional CNNs, particularly in handling diverse action categories and varied environments.However, challenges remain, such as distinguishing between fine-grained action categories and handling dynamic movements in still images. Future improvements could include integrating temporal context, or multimodal inputs to address these limitations, as well as expanding the model’s ability to generalize across diverse datasets. Overall, while ViT shows great promise, further advancements are needed to enhance its robustness and accuracy in real-world action recognition tasks.

REFERENCES

[1] S. Yu, Y. Cheng, S. Su, G. Cai, and S. Li, “Stratified pooling based deep convolutional neural networks for human action recognition,” Multimedia Tools and Applications, vol. 76, pp. 13367-13382, Jul. 2016.

[2] G. Chéron, I. Laptev, and C. Schmid, "P-CNN: Pose-Based CNN Features for Action Recognition," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015, doi: 10.1109/ICCV.2015.368.

[3] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[4] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” Advances in Neural Information Processing Systems, vol. 27, 2014.

[5] G. Varol, I. Laptev, and C. Schmid, “Long-Term temporal convolutions for action recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[6] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, S. W. Baik, “Action recognition in video sequences using deep bi-directional LSTM with CNN features,” IEEE Access, vol. 6, pp. 1155–1166, 2018.

[7] A. Dosovitskyet al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” International Conference on Learning Representations, 2021.

[8] Sun, Zhonglin, and Georgios Tzimiropoulos. "Part-based face recognition with vision transformers." arXiv preprint arXiv:2212.00057, 2022.

[9] R. Garcia-Martin and R. Sanchez-Reillo, "Vision Transformers for Vein Biometric Recognition," IEEE Access, vol. 11, pp. 22060-22080, 2023, doi: 10.1109/ACCESS.2023.3252009.

[10] N. M. Alharthi and S. M. Alzahrani, "Vision Transformers and Transfer Learning Approaches for Arabic Sign Language Recognition," Appl. Sci., vol. 13, no. 21, pp. 11625, 2023. [Online]. Available: https://doi.org/10.3390/app132111625.

[11] G. Wang, Y. Chen, P. An, H. Hong, J. Hu, and T. Huang, "UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios," Sensors (Basel), vol. 23, no. 16, p. 7190, Aug. 2023, doi: 10.3390/s23167190.

[12] B. Yao, X. Jiang, A. Khosla, A.L. Lin, L.J. Guibas, and L. Fei-Fei. Human Action Recognition by Learning Bases of Action Attributes and Parts. Internation Conference on Computer Vision (ICCV), Barcelona, Spain. November 6-13, 2011.

[13] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," Int. J. Comput. Vis., vol. 111, no. 1, pp. 98-136, 2015.

[14] Fan, C. Chen, and R. Panda, "Can an image classifier suffice for action recognition?" International Conference on Learning Representations 2021.

[15] T. Chen and L. Mo, "Swin-Fusion: Swin-Transformer with feature fusion for human action recognition," Neural Process. Lett., vol. 55, pp. 11109–11130, 2023, doi: 10.1007/s11063-023-11367-1.

[16] S. R. Hosseyni, S. Seyedin, and H. Taheri, “Human Action Recognition in Still Images Using ConViT,” 2024 32nd International Conference on Electrical Engineering (ICEE), pp. 1–7, May 2024, doi: 10.1109/icee63041.2024.10668316.

[17] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lu, and Cordelia Schmid. Vivit: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6836–6846, 2021.

[18] A. Hussain, T. Hussain, W. Ullah, and S. W. Baik, "Vision Transformer and Deep Sequence Learning for Human Activity Recognition in Surveillance Videos," Computational Intelligence and Neuroscience, vol. 2022, Article 3454167, 2022. doi: 10.1155/2022/3454167.

[19] GedasBertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In Proceedings of International Conference on Machine Learning (ICML), vol. 2, p. 4, 2021.

[20] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network,". In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 244–253, 2019.

[21] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," Pattern Recognition., vol. 124, p. 108487, 2022.