2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) | 979-8-3503-9648-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/InCACCT57535.2023.10141761

Effectiveness of Vision Transformers in Human Activity Recognition from Videos

Rahul Kumar
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
rahuldtucs@gmail.com

Shailender Kumar
Department of Computer Science & Engineering
Delhi Technological University
Delhi, India
shailenderkumar@dce.ac.in

***Abstract*— Human Action Recognition (HAR) has got the attention of computer vision domain researchers due to its wide variety of applications like surveillance, behavior detection, sports action monitoring, and elderly monitoring. Due to the huge amount of data, the Deep Learning-based method is widely used in HAR compared to the Machine Learning-based approach. This study explored the various Deep Learning and pre-trained Deep Learning models in HAR. In the pre-trained model, we do not require to train the model from scratch, which is already trained on huge data. This study explored the recent pre-trained Deep Learning model to classify action accurately. This study helps the researcher to evaluate the benefit of the latest Vision Transformer model in the domain of HAR.UCF 50 action dataset is used in this study to examine the effectiveness of the Vision Transformer model in HAR. On UCF 50 action dataset, we have achieved 94.70 % accuracy using the Vision Transformer model variant.**

Keywords—Transformer; Deep Learning; Human Action Recognition; Machine Learning

I. INTRODUCTION

HAR is an active hot research area due to its broad variety of applications, such as video retrieval, security, behavior monitoring, patient activity, elderly monitoring and defense. HAR has several steps: data pre-processing, feature extraction (pose-based, shape-based), learning algorithm and action classification. Real-time Action Recognition (AR) is a hot research domain in HAR for security reasons. In real-time, first, we have to train the model on less available data; after that, data increase rapidly and requires lots of computation. Machine Learning (ML) based classification method used by the researcher earlier to recognize the action. Due to a large amount of data ML-based method does not perform better. Due to the less effectiveness of the ML-based method on huge data, the researcher used Deep Learning (DL) methods to identify action from the video sequences. DL-based methods require more computation power to train the model on huge data. Convolution-Neural-Network (CONET) plays a crucial role in computer vision to identify images and patterns. Other deep learning models like Long Short-Term Memory (LSTM), Recurrent Neural Networks (RNN), and Generative Adversarial Networks (GAN) are used by the researcher's community and are constantly improved. The DL-based approach performs better as compared to traditional ML-based methods. HAR can be classified into 4 action levels[1]: Gesture, Interaction, Group Activity and Action.

HAR still faces issues with action classification due to the view-invariant, low resolution and variable length sequences in datasets. Several ML-based algorithms have been introduced, but each has some limitations. Some methods work well on a specific dataset, short-length videos, single view and unrealistic conditions. The well-suited HAR model can handle challenges and perform better in each circumstance. The DL methods can handle a large amount of data compared to ML methods because they need to perform better when data is large. Researchers are constantly improving methods to recognize action accurately. Some multimodal approaches are also introduced to learn specific actions. In the multimodal approach, we fed various modalities to the model, for example, RGB, Depth data, and Skelton data. Fusion techniques (early fusion, slow fusion, late fusion) are used in a multimodal approach to combine features we received from various modalities. Common steps in the HAR process are Data Pre-Processing, Object Segmentation, Feature Extraction, Training of data and classification. The DL-based method uses an automated feature extraction method and the ML-based approach uses a handcrafted feature approach (pose-based, shape based). The HAR process is depicted in Fig. 1.

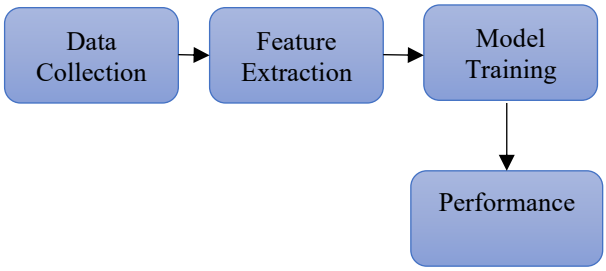

Fig. 1 HAR Complete Process.

DL method succeeded in image classification, object detection and recognizing the action. Deep Learning provides various architectures like CONET employs the best feature extraction approach compared to the handcrafted feature extraction method[2]. Many frameworks depend on CONET to solve the HAR issue. Recurrent Neural Network (RNN) is another framework of DL, whereas Links among nodes create a directed graph along a time sequence, enabling dynamic activity[3]. The main offering of this study is to classify action with the help of the latest Vision Transformer model and compare it with state-of-the-art methods.

This article is structured as follows: Section 2 describes the relevant research in AR. The techniques and data used to assess

ISBN: 979-8-3503-9648-5/23/$31.00 ©2023 IEEE

model performance are described in Section 3. Section 4 addresses the outcome and its comparison to recent approaches.

II. Related work

Deep architecture CONET for feature extraction and sequential pattern learning by LSTM architecture in [4]. Automated Learning of features suggested by ji et al. [5]. They used 3-D CONET for HAR. But this model is unsuitable for variable-length clips because of inflexible architecture frames and optical flow input base dual stream model suggested by Simoyan et al.[6] and input fed into the pre-trained DL model. CONET is employed for FE and classification of action of KTH 1 or KTH 2 dataset in and named two-step approach[7]. Ijjina and Mohan developed a deep hybrid model by mixing homogeneous CONETs, which obtained 99.68% accuracy on the UCF 50 dataset [8] Liang et al.[9] studied a highly unsupervised learning model for human segmentation jobs. The video-context-driven human mask inference and CONET-based segmentation network learning iterated until mutual improvement was no longer possible. PASCAL VOC 2012 passed exhaustive testing with an accuracy of 81.8%.

Safaei M et al. [10] suggested an advanced CONET-based technique for forecasting future action and recognizing the form and position of the image's prominent components. They trained a single framework for every move in a one-versus-all approach and obtained a 76.1% accuracy. In [11] capture, pose-based characteristics from the 3D CONET network are capable of 3D posture, 2D appearance, and motion flow unification. The extraction of joint color features for the 3D CONET will result in considerable complexity; consequently, a 15-channel heatmap is generated, and convolution is conducted in each map. Feichtenhofer et al.[12] present a two-stream convolution system for combining dual temporal and spatial streams, in which RGB info (spatial) and optical flow (motion) are modeled simultaneously and estimates are aggregated in the final layers. Due to optical flow, this network cannot capture long-term motion; another drawback of the spatial CONET stream is that its performance depends on a randomly selected image from the input video. Due to background muddle and viewpoint shifts, complications develop.

Shi et al.[13] offer another two-stream network to improve the performance of skeletal joints-based HAR. The Adaptive Graph Convolutional Network (AGCN) network processes a two-stream containing joint and bone information. The network is made up of a stack of these fundamental components. The soft-max layer has been added to the final result. Ullah et al.[14] use real-time video from a non-stationary camera to perform HAR on the system. CONET, a deep learning approach, automatically extracts frame-level characteristics. Jian et al. describe a method for extracting ROI using a Fully Convolutional Network (FCN)[15]. CONET is utilized to determine each frame's posture probability. Key-frame extraction is carried out using the nearby probability difference of frames. The variation-aware key-frame extraction method considers the frame with the greatest probability of a key pose, as assessed by CONET. If many frames give the same key pose probability value, the Central frame is selected. [16] A DL-based solution for temporal 3D poses identification challenges using a CONET and an LSTM recurrent network. They begin by training the CONET and then tweak the combination (CONET+ LSTM). CONET extracts the relevant information, which the LSTM subsequently uses to classify the target action classes. This sort of innovative training strategy outperforms single-stage training. On small-sized data, the results outperform several state-of-the-art approaches. [17] devised a one-of-a-kind HAR model with three primary stages: pre-processing, background removal, and classification. Their technique greatly enhances the results for five action classes extracted from the INRIA and KTH datasets: running, walking, leaping, standing, and sitting. To capture local Spatio-temporal information, [18] presented a multi-resolution CONET architecture for feature connectivity in the time domain. Such a method is being tested on a current "YouTube 1 million videos dataset" with 487 action sequences of classes. The authors mentioned that the foveated design of CONET sped up training complexity. They raised their action categorization rate for big datasets to 63.9%, but their recognition rate for UCF101 is still 63.3%, which is insufficient for such a critical job as action recognition.

Bilen et al. [19]examine the feature maps of a pre-trained model for dynamic image video representation. At the tuning step, they introduced a rank pooling operator and an approximation rank pooling layer, which combine the mappings of all frames into a single dynamic picture representing the whole movie. Deep learning-based algorithms may dependably uncover hidden patterns in visual data due to their extensive feature representation pipeline. On the other hand, it requires a vast quantity of data for training and a high degree of computer processing power. Wu et al. [20] introduced a deep learning-based MPCA-Net technique for classifying human actions. Tensor interaction is used in this strategy to improve the recognition rate. It comprises three layers: projection dictionaries, a projection encoder layer, and a pooling layer. The suggested method is examined on UCF11, different medical imaging datasets, and UCF sports action datasets to establish its effectiveness. Majd et al.[21] suggested a correlational convolutional LSTM (C^2 LSTM) for handling spatial and movement knowledge of surveillance video data. Conv-LSTM is a connectivity combination model for violence detection proposed in [22]. This design employed the CONET network as the fed into the LSTM network for feature analysis and action categorization. Jaouedi et al. proposed a unique method for motion tracking through human monitoring and spatial feature gathering from video sequences. Two techniques were used to build a robust feature vector for classification: Gaussian mixture model (GMM) and Kalman filter (KF) algorithms to detect and extract moving persons, and Gated Recurrent Neural Networks to gather data in each frame and predict human activity. They tested their technique using datasets from UCF Sports, UCF101, and KTH, achieving 89.01%, 89.30%, and 96.30%, respectively.[23]

III. Proposed approach

Deep Learning-based methods play a crucial role in action recognition with their model CONET, LSTM, GAN and Autoencoder and achieve better results than ML-based methods. DL-based methods are capable of dealing with huge amounts of data. But training from scratch for each model is very complex, so various Transfer Learning based methods suggested pre-trained on the huge amount of data. In this study, we evaluated Vision Transformer (VT) for action recognition in video sequences. **Fig. *2*** represents how VT works on frame sequences to detect a class of corresponding frames.

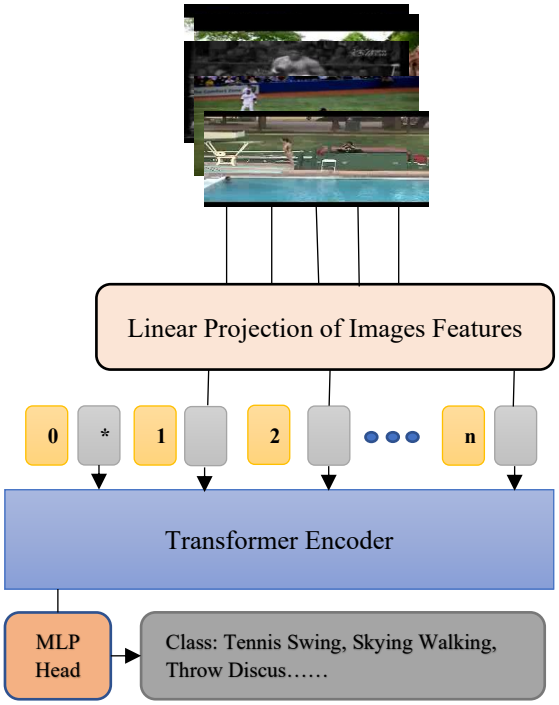

Fig. 2 Action Transformer architecture to recognize action class.

A. Transformer

A 1D sequence of token embeddings is fed into the standard Transformer. In the Transformer, to manage 2-D images reshaping required images $x \epsilon \mathbb{R}^{H \times W \times C}$ into an order of flattened two-dimensional spots $x_p \epsilon \mathbb{R}^{N \times P^2 \times C}$, images resolution is represented by (H, W), count of the channel with the help of C, Each image's resolution shows from (P, P) and the total number of patches represented by $N = HW/P^2$, which also acts as the essential input order span for the Transformer. Because the Transformer uses a constant latent vector size D throughout its stages, they use a trainable linear projection to straighten the areas and transfer to D dimensions. The VT's first layer sequentially reflects the smoothed regions into a shorter realm[24]. The characteristics resemble appropriate basis functions for a low-dimensional representation of the fine structure contained inside each patch. After the projection, the patch representations are then improved with a learned position embedding.

In location embedding resemblance, the model learns to express distance inside the visual, i.e., closer patches have more similar position embeddings. The row-column structure is also obvious; deep characteristics are identical within the same row/column. Because of self-attention, VT can acquire data during the whole visual, even at the lowest levels[24].

B. Dataset

UCF 50 action dataset was used to evaluate the performance of the VT model. This dataset was presented by Reddy et al.[25] in 2012. Action videos are gathered from online sources such as YouTube and have a realistic environment. This dataset has 50 distinct action classes, such as playing tabla, baseball pitch, yo-yo, walking with the dog, and throwing discus. The dataset contains about 100 short videos on every class, with a broad span of camera motion, object look and posture, object scale, perspective, cluttered backdrop, illumination conditions, and so forth. **Fig. *3*** depicts the action sequences from UCF 50 dataset

Fig. 3 Action sequence taken from UCF 50 dataset.

IV. Results & Discussion

Transfer learning is a weight initialization approach where the neural network is not learned using stochastic weights and biases but rather with pre-learned weights and biases from the ImageNet dataset. The network is then permitted to learn and upgrade its parameters on the training data, the HAR vision datasets. In this approach, it just becomes fine-tuned for such a current categorization goal. We used the VT model to classify each action of the UCF50 dataset and compared these with state-of-the-art methods. VT variants are pre-trained deep learning models trained on large-scale dataset ImageNet[26].

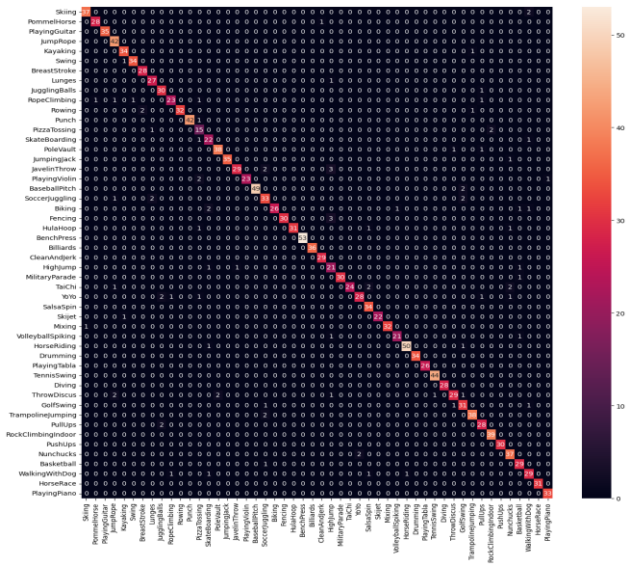

Fig. 4 Confusion Matrix for UCF 50 action dataset classification using VT_B_16 model.

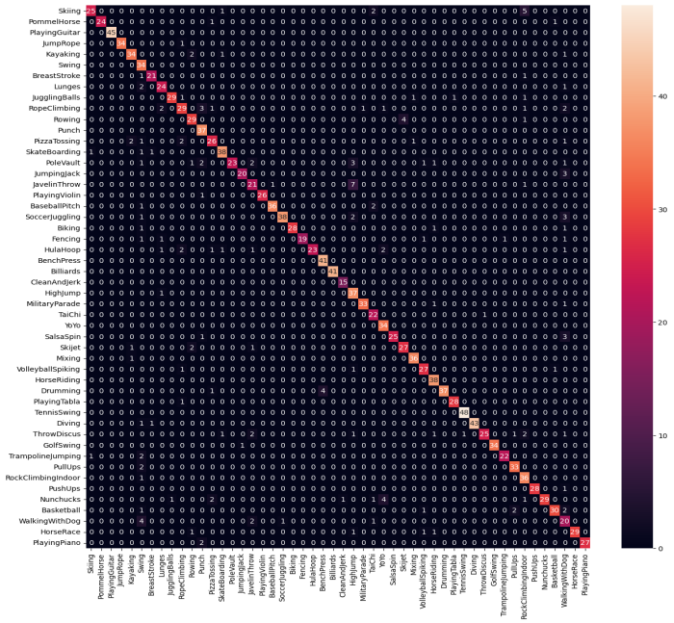

Fig. 5 Classification of action sequence of UCF 50 dataset with VT_l_16 model.

These models can classify each action accurately and do not need to train the model from scratch. This study has compared the accuracy of the various variants of VT on the UCF 50 dataset. This study has extracted sequences of images of each action category and trained them with the help of these frames. VT performs better as compared to the various Transfer Learning and DL models. Fig. 4 & Fig. 5 represent the confusion matrix for UCF 50 dataset. We have compared different variant performances of the VT model in Table 1 and plotted a comparison graph in Fig. 6.

Table 1: Comparison of performance between Vision Transformer variants

Model Name	Accuracy	Precision	Recall	F1-Score
Vision Transformer(vit_b-16)	94.70	94.36	94.17	94.11
Vision Transformer(vit_l_16)	89	91.07	89.47	89.69

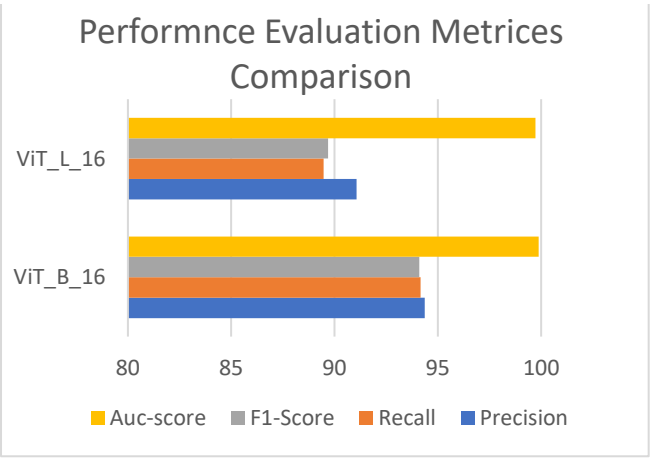

Fig. 6 Comparison Graph of Evaluation Metrics

In this study, UCF 50 action dataset has been employed to examine the effectiveness of the VT model in terms of accuracy, precision and recall. VT models are pre-trained models and can classify each action effectively. We implemented VT models on GPU to evaluate the effectiveness of these models in HAR. This study has used two variants of VT (VT_b_16 & VT_l_16), which means base or large model. They contain 16 X 16 input sizes. The base model has 12 layers with 32 M params and the large model has 24 layers with 307 M params. **Table *2*** compares these models with other state-of-the-art methods in terms of accuracy.

Table 2: Evaluation on the UCF 50 dataset against state-of-the-art methodologies.

Reference Model	**Accuracy (in %)**
Action Bank [27].	76.4
Multi-Channel Descriptor [28]	83.3
Global-Spatio Temporal Feature[29]	70.1
Wang et al[30]	91.7
Vision Transformer (VT_b_16)[24]	**94.70**
Vision Transformer (VT_l_16) [24]	89.00

V. CONCLUSION

This article has explored the VT model in the domain of HAR.VT models are capable of classifying action effectively because of their deep architecture. Using the UCF 50 dataset, we compared the effectiveness of these models to that of state-of-the-art methods. This study compared various evaluation metrics of the model (f1-score, precision, recal,l etc.) to look at effectiveness. The VT model outperforms other suggested methods on the UCF 50 action dataset, with an accuracy of 94.70%. In the future, more complex and Multiview datasets can be tested. These models can be used to classify online action and complex action identification. Numerous variants of Transformer suggested with fewer parameters, requiring less time and computing power. These models are also used in surveillance data for quick response.

VI. REFERENCES

[1] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," in *Visual Computer*, Oct. 2013, vol. 29, no. 10, pp. 983–1009. doi: 10.1007/s00371-012-0752-6.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." [Online]. Available: http://code.google.com/p/cuda-convnet/

[3] A. Graves, A. Mohamed, and G. Hinton, "Speech Recognition with Deep Recurrent Neural Networks," Mar. 2013, [Online]. Available: http://arxiv.org/abs/1303.5778

[4] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action Recognition in Video Sequences using Deep Bi-Directional LSTM with CNN Features," *IEEE Access*, vol. 6, pp. 1155–1166, Nov. 2017, doi: 10.1109/ACCESS.2017.2778011.

[5] S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition." [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid/

[6] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos."

[7] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential Deep Learning for Human Action Recognition."

[8] E. P. Ijjina and C. Krishna Mohan, "Hybrid deep neural network model for human action recognition," *Applied Soft Computing*

Journal, vol. 46, pp. 936–952, Sep. 2016, doi: 10.1016/j.asoc.2015.08.025.

[9] X. Liang *et al.*, "Learning to Segment Human by Watching YouTube," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 7, pp. 1462–1468, Jul. 2017, doi: 10.1109/TPAMI.2016.2598340.

[10] M. Safaei and H. Foroosh, "Single Image Action Recognition by Predicting Space-Time Saliency," May 2017, [Online]. Available: http://arxiv.org/abs/1705.04641

[11] Y. Huang, S.-H. Lai, and S.-H. Tai, "Human Action Recognition Based on Temporal Pose CNN and Multi-Dimensional Fusion."

[12] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional Two-Stream Network Fusion for Video Action Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Dec. 2016, vol. 2016-December, pp. 1933–1941. doi: 10.1109/CVPR.2016.213.

[13] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition." [Online]. Available: https://github.com/lshiwjx/2s-AGCN

[14] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments," *Future Generation Computer Systems*, vol. 96, pp. 386–397, Jul. 2019, doi: 10.1016/j.future.2019.01.029.

[15] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks." [Online]. Available: https://github.com/daijifeng001/r-fcn.

[16] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, "Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognit*, vol. 76, pp. 80–94, Apr. 2018, doi: 10.1016/j.patcog.2017.10.033.

[17] G. Sulong and A. Mohammedali, "RECOGNITION OF HUMAN ACTIVITIES FROM STILL IMAGE USING NOVEL CLASSIFIER 1," *J Theor Appl Inf Technol*, vol. 10, no. 1, 2015, [Online]. Available: www.jatit.org

[18] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks." [Online]. Available: http://cs.stanford.edu/people/karpathy/deepvideo

[19] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic Image Networks for Action Recognition."

[20] J. Wu, S. Qiu, R. Zeng, Y. Kong, L. Senhadji, and H. Shu, "Multilinear Principal Component Analysis Network for Tensor Object Classification," *IEEE Access*, vol. 5, pp. 3322–3331, 2017, doi: 10.1109/ACCESS.2017.2675478.

[21] M. Majd and R. Safabakhsh, "Correlational Convolutional LSTM for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, Jul. 2020, doi: 10.1016/j.neucom.2018.10.095.

[22] IEEE Signal Processing Society, IEEE Computer Society, and Institute of Electrical and Electronics Engineers, *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) : Aug. 29 2017-Sept. 1 2017.*

[23] N. Jaouedi, N. Boujnah, and M. S. Bouhlel, "A new hybrid deep learning model for human action recognition," *Journal of King Saud University - Computer and Information Sciences*, vol. 32, no. 4, pp. 447–453, May 2020, doi: 10.1016/j.jksuci.2019.09.004.

[24] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11929

[25] K. K. Reddy and M. Shah, "Recognizing 50 Human Action Categories of Web Videos."

[26] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int J Comput Vis*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[27] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 1234–1241. doi: 10.1109/CVPR.2012.6247806.

[28] F. Shi, E. Petriu, and R. Laganiere, "Sampling strategies for real-time action recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2595–2602. doi: 10.1109/CVPR.2013.335.

[29] G. Somasundaram, A. Cherian, V. Morellas, and N. Papanikolopoulos, "Action Recognition Using Global Spatio-Temporal Features Derived from Sparse Representations."

[30] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A Robust and Efficient Video Representation for Action Recognition," *Int J Comput Vis*, vol. 119, no. 3, pp. 219–238, Sep. 2016, doi: 10.1007/s11263-015-0846-5.