International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Multi-Class Human Activity Prediction using Deep Learning Algorithm

Jitha Janardhanan*[1],Dr, S. Umamaheswari[2]

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: The most common technique for understanding human behavior is action recognition based on videos. Videos provide significantly more information than image-based action recognition. Reducing action ambiguity as well as several dataset-focused studies, innovative models, and learning strategies over the past ten years have elevated video action identification to a higher level. However, there are difficulties and unresolved issues, particularly in real-time CCTV analytics where data gathering and labeling are more complex, necessitating the annotation of data. Additionally, the actions could happen very quickly, making it challenging to distinguish between them. This paper presented a video based multiple human activity recognition for real-time CCTV camera videos. To propose a HAR to introduce an Enhanced TimeSformer Model with Multi-Layer Perceptron (ETMLP) Neural Networks classification algorithm applies self-attention over the patches and human regions. By interacting with uniform classification tokens in this manner and enhancing them with contextualized human activity data, visual human pose estimation areas are able to estimate human pose.

Keywords: *Human activity, EfficientNET, Pose estimation, TimeSformer, Optical flow*

1. Introduction

There is a huge demand for video understanding, which includes things like understanding human behaviour, tracking objects, spotting anomalous behaviour, and content-based video retrieval. The amount of videos is growing quickly, and this demand is driving the development of new technologies. Numerous applications in daily life, such as monitoring systems, are possible as a result of the advancement of video understanding technology. Video understanding is a fundamental module for video analysis, and action recognition is at its core. In order to increase the recognition accuracy, researchers have labelled several videos [1], [2]–[7], and proposed a number of outstanding models [8]–[9]. The popular datasets, however, such as ActivityNet [5] and Kinetics-400 [10], only take into account the activities to engage in on a daily basis, such as walking, leaping, and riding bikes. Additionally, need to tag videos that concentrate on different human actions, such as running, walking, jumping, driving, etc., in order to reach the goal of fine-grained sports action detection. Additionally, in order to create fine-grained annotations, it is typically necessary to have class labels and domain expertise.

Video transformers have recently been recommended as efficient models for video understanding [13], spurred on by the success of transforms in language [11] and vision [12]. Each video frame in these models is split into patches, and self-attention architecture creates a contextualized representation for each patch. The explicit depiction of things is absent from this methodology, nevertheless. The most important finding is that self-attention may be used to increase both object and spatiotemporal representations simultaneously, providing a simple and elegant method for doing so.

Since numerous applications, such as video retrieval [14], video captioning [15], video quality assurance [16], etc., can advantage from better action recognition modelling, action recognition has attracted a lot of attention from the research community. In terms of improvement, datasets are one factor. Millions of videos retrieved from the Internet have been added to video databases, which have grown from a few hundred videos shot in controlled settings to millions of videos now. The range of videos has expanded along with their availability. For instance, the human activities are covered by video recognition datasets have progressed from straightforward actions like walking, standing and waving hands to more intricate ones that occur in present daily lives.

The complexity of modelling architecture has evolved concurrently with the complexity of data and class distribution [14]. Among these designs, Transformer-based techniques have lately shown cutting-edge performance on a number of benchmarks [17]. Action recognition's training and tuning process is simple and exhibits excellent empirical findings, but it might be too constrained for creating models that can be used for a variety of different actions. Action recognition datasets like Kinetics [18] cover fewer subjects than datasets like ImageNet [19], which covers a wide range of object recognition classes. Kinetics, for instance, concentrates on activities like "cliff diving" and "ice climbing". Therefore, it is likely that applying an action recognition that has been optimized on Kinetics may lead to bad performance. Variations in objects and video backgrounds

[1] *Research Scholar, Dept. of Computer Science, Dr.G.R.Damodaran College of Science, Coimbatore, India*
ORCID ID: 0000-0001-9154-609X
[2] *Associate Professor, Dept. of Computer Science, Dr.G.. R. Damodaran College of Science, Coimbatore, India*
ORCID ID:: 0000-0003-0244-8338
** Corresponding Author Email: jithajanardhanan@gmail.com*

between datasets make it more difficult to learn. The prior research [14] suggests that to attain high performance, extensive data augmentation and regularization are still needed. This result might be a sign that the model soon overfits on the target dataset, which limits its ability to generalize to other action detection tasks.
Building a training method for a general-purpose multiple human activity recognition models is goal in this work. The authors suggested using video data to jointly train numerous human activity recognition models, drawing inspiration from earlier efforts in vision and language that show a Transformer model may be extended to many downstream tasks [20]. Two key findings lend credence to this strategy. First, several video datasets include a variety of activities, and combining them to train a distinct model may produce a model that performs well across a broad range of tasks. Additionally, video is a terrific way to learn about motion, and video frames are excellent for utilizing visual structure.

In this study, the authors concentrated on the recognition of video activity in diverse real-time CCTV videos. The Depth wise Separable Convolution with Bidirectional Long Short-Term Memory (DSC-BLSTM) approach is tailored to the HAR task, according to Jitha Janardhanan and S. Umamaheshwari's 2022 [30] proposal. This system makes use of a powerful feature extraction and classification model to increase the accuracy of activity recognition; however it does not focus much on multiple-human activities that have annotations. The problems that need to be taken into consideration going forward are presented in this research together with the current state of multiple-human activity video detection in various real-time CCTV home apartment videos.

These are the three contributions that have been proposed:

- For the goal of modelling multiple human activity recognition data, to study a real-time CCTV video data that has been trained simultaneously.
- By simultaneously learning across various activities in video datasets, to offer the Enhanced TimeSformer Model with Multi-Layer Perceptron (ETMLP) Neural Networks classification algorithm technique, this learns reliable spatial and temporal representations. To compete with earlier pre-training and fine-tuning paradigms, the learned representations can be deployed right away to a variety of downstream tasks.
- The ETMLP method generated novel State-of-The-Art results on a variety of real-time video datasets.

2. Related Work

Girdhar et al. in 2019 [21] proposed a transformer-style architecture to combine features from the spatiotemporal context surrounding the person whose behaviors they are trying to categories. The authors shoed the model may be trained to watch certain persons and to infer semantic context from other human's behaviors by using high-resolution. Additionally, without any other cues besides boxes and class labels, its attention system learns to priorities hands and faces, which are frequently crucial to differentiating an activity.

A family of effective video networks called X3D was introduced by Feichtenhofer in 2020 [22]. They gradually expand modest 2D image classification architecture along various network axes, including those of space, time, width, and depth. Progressive forward expansion and reverse contraction were used to expand X3D to a particular goal complexity. For the same precision as prior work, X3D provides state-of-the-art performance while using 4.8x and 5.5x less multiply-adds and parameters. The most unexpected result of their research is that networks with great spatiotemporal resolution can function well while requiring very little in the way of network width and parameters.

Bertasius, G., and Torresani, L. presented a method for locating, detecting, and separating object instances in a video sequence in 2020 [23]. Their MaskProp approach, which includes a mask propagation branch, translates the well-known Mask R-CNN to video. Their system can forecast frame-level instance tracks with relation to the object instances that are segmented in the middle frame of the clip. To construct video-level object segmentation and classification, the next step is to aggregate the densely created each frame in the sequence has clip-level instance tracking.

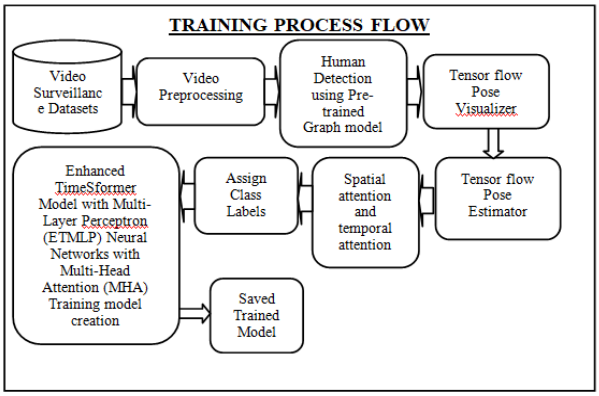
Carion et al. presented a revolutionary method in 2020 [24] discussed about object detection prediction problem. The authors simplifies the detection process and does away with various manually made components, like anchor formations or non-maximum suppression operations, which directly encode the user's prior task knowledge. Transformer encoder-decoder architecture performs unique predictions through bipartite matching are the two main parts of the new framework, also known as DEtection TRansformer or DETR.

The Transformer architecture, which Minderer et al. presented in 2020 [25], is now the de facto standard for tasks involving natural language processing, although its applications to computer vision are still very limited. When applied to convolutional networks in vision, attention is either employed in conjunction with them or is used to replace specific convolutional network elements while maintaining the general structure of the network. The authors demonstrated that a pure transformer applied straight to a sequence of image patches may carry out image classification tasks very well without the need for CNNs.
Sevilla-Lara et al., 2021 [26] tackled the issue head-on by distinguishing action types that require time information to be recognized, referring to these as "temporal classes." The approach would be biased if temporal classes were chosen via a computational method. Instead, they suggested a methodology based on a straightforward and successful trial using human annotation. By switching around the frames of time, they only deleted the temporal information, and then checked to see if the action could still be distinguished. The temporal set includes classes that are not recognizable when the frames are out of order.

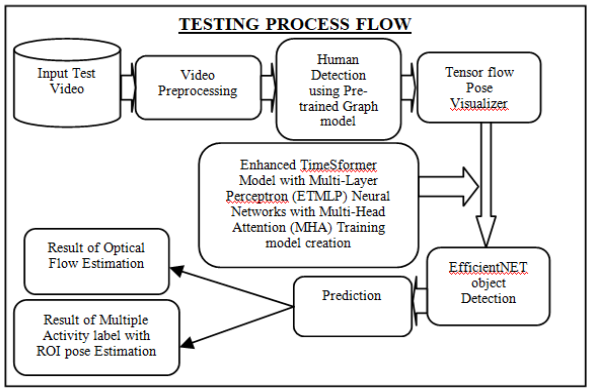
Jitha Janardhanan and S. Umamaheswari introduced a deep learning neural network method employing the Depthwise Separable Convolution (DSC) with Bidirectional Long Short-Term Memory (DSC-BLSTM) algorithm in 2022 [27]. One of the suggested network system's redeeming features is a DSC convolution, which lowers both the amount of parameters that can be learned and the running time of the combined training and testing technique. The positive and negative time directions can be combined using the bidirectional LSTM method.

3. Proposed Methodology

The proposed method presents a human activity recognition based on single and multiple activity models with ROI (Region of Interest) bounding box model. This system performs three stages namely, (i) video frame extraction; (ii) Pose Estimation and Enhanced Bidirectional GRU with LSTM (BGRU-LSTM) method; and (ii) Enhanced TimeSformer Model with Multi-Layer Perceptron (ETMLP) Neural Network classification method. The ETMLP method is an enhanced version of BGRU-LSTM convolution network classification model was already we presented (Jitha Janardhanan, and S. Umamaheshwari, 2022 [30]). Figure 1 shows the overall ETMLP process in detail.



(a)



(b)

Figure 1. Overall ETMLP flow diagram. (a) Training model preparation; (b) Testing Multiple HAR prediction

3.1. Video Frame Extraction

Jitha Janardhanan and S. Umamaheshwari (2022 [27]) already discussed the frame-by-frame evaluation of video frame extraction and how to achieve the same resolution as the original video. By using the automated video frame extraction capability to scan a video, the default frame extraction interval is first identified. Following the calculation of the interval, the system will automatically extract frames based on a proper matching of features in the image frames and add the extracted frames to the appropriate folder. In this phase, choose a selection of all the evenly spaced frames from the videos. More frames were taken into account, but the performance suffered. To keep computation time down, be compatible with the majority of architectures, and provide enough data for accurate classification, a crop size of 224 × 224 centered is extracted from each video. In order to extract the crops, the detection is adjusted to a final zone of 224 by 224 while taking into account the human detector output.

3.2. Pose Estimation and Enhanced Bidirectional GRU with LSTM (BGRU-LSTM) method

The system first locates the human key points of interest in the input video frame before considering the Pose Estimation (Jitha Janardhanan and S. Umamaheshwari, 2022), which takes into account human position estimation. Many computer vision applications, particularly the understanding of human behavior, rely heavily on this technique. The model of human position estimation is handled by the EfficientNet transfer learning algorithm. EfficientNet has the distinction of being able to attain excellent accuracy with a minimal number of parameters.

On the basis of real-time surveillance video datasets, the Enhanced Bidirectional GRU with LSTM (BGRU-LSTM) technique was already reported (Jitha Janardhanan and S. Umamaheshwari, 2022). It takes human activity into account. The residual nodes are learned by the LSTM using the hidden state as a reference. In many applications, including the recognition of numerous human activities, the Bidirectional GRU with LSTM structure, which has two regular LSTM layers for extracting temporal dynamics from both forward and backward directions, is a significant RNN variant.

3.3. Enhanced TimeSformer Model with Multi-Layer Perceptron (ETMLP) Neural Network classification

The proposed ETMLP method takes as input N number Human pose estimation RGB frames with the dimension of 224 × 224. Every frame is then divided into patches of dimension Patch×Patch (where Patch is a hyper parameter). A learnable embedding made up of a positional embedding and a learnable matrix is then attached to each patch. Following that, the encoder, which consists of M encoding blocks, receives these embeddings. The query (Qr), K, and V values for the following encoding block are calculated for each of these.

A clip C ∈ $R^{H\times W\times 3\times F}$ consisting of F RGB frames of size Height (H) × Width (W) sampled from the human pose estimation video is used as the input for the proposed ETMLP algorithm. The ETMLP approach divides every frame into M non-overlapping patches, each of size Patch× Patch, so that the M patches cover the full frame, i.e., M = Total_pixels/P^2. This is in accordance with the Vision Transformer [24]. These patches are flattened into vectors $C_{(p,t)} \in R^{3Patch^2}$, where patch = 1, 2,..., M represents spatial positions and t = 1, 2,..., F denotes a frame index.

Each patch $C_{(p,t)}$ is linearly mapped into an embedding vector $eV^{(0)}_{(p,t)} = R^D$ by means of learnable matrix L ∈ $R^{D\times 3P2}$

$$eV^{(0)}_{(p,t)} = LC_{(p,t)} + posEmb^{pos}_{(p,t)} \quad (1)$$

where each patch's spatiotemporal position is encoded by $posEmb^{pos}_{(p,t)} \in R^D$, a learnable positional embedding that has been added. The sequence of embedding vectors that is produced for p = 1, 2,..., M, and t = 1, 2,... F serves as the Transformer's input and functions similarly to the sequences of embedded activities that are provided to text Transformers in NLP (Natural Language Programming).

In the ETMLP Transformer, E encoding blocks are used. From the form $eV_{(p,t)}^{(e-1)}$ encoded by the preceding block, a QR or K or V vector is constructed for each patch at block e:

$$Qr_{(p,t)}^{(e,ma)} = W_{Qr}^{(e,ma)} N\left(eV_{(p,t)}^{(e-1)}\right) \in \mathbb{R}^{D_h} \quad (2)$$

$$K_{(p,t)}^{(e,ma)} = W_{K}^{(e,ma)} N\left(eV_{(p,t)}^{(e-1)}\right) \in \mathbb{R}^{D_h} \quad (3)$$

$$V_{(p,t)}^{(e,ma)} = W_{V}^{(e,ma)} N\left(eV_{(p,t)}^{(e-1)}\right) \in \mathbb{R}^{D_h \mathrm{v}} \quad (4)$$

Where *ma* = 1, 2,..., MA is a list of several attention heads and MA is the amount of attention heads, and N() denotes LayerNorm [29]. Each attention head's latent dimensionality is set to Dh = D/MA.

Dot-product is used to calculate self-attention weights. The following equations yield the self-attention weights $sW_{(p,t)}^{(e,ma)} \in \mathbb{R}^{MF+1}$ for the query patch (p, t):

$$sW_{(p,t)}^{(e,ma)^{space}} =$$

$$SA\left(\frac{Qr_{(p,t)}^{(e,ma)}}{\sqrt{D_h}} \cdot \left[K_{(0,0)}^{(e,ma)} \left\{K_{(p',t')}^{(e,ma)}\right\}_{\substack{p'=1,2,\ldots,M \\ t'=1,2,\ldots,F}}\right]\right) \quad (5)$$

SA stands for the softmax activation function in this context. Please take note that the computation is much decreased when attention is only computed in one dimension.

Initially, using the self-attention coefficients from each attention head, the weighted sum of value vectors is calculated with encoding $eV_{(p,t)}^{(e)}$ at block e is achieved. Then, using any remaining connections, an MLP is used to project and send the vectors listed below concatenated from all heads:

$$eV'^{(e)}_{(p,t)} = Weight\begin{bmatrix} sW_{(p,t)}^{(e,1)} \\ \vdots \\ sW_{(p,t)}^{(e,MA)} \end{bmatrix} + eV_{(p,t)}^{(e-1)} \quad (6)$$

$$eV_{(p,t)}^{(e)} = MLP\left(N\left(eV'^{(e)}_{(p,t)}\right)\right) + eV'^{(e)}_{(p,t)} \quad (7)$$

From the last block for the classification token (ct), the final clip embedding is retrieved.

$$ct = N\left(eV_{(0,0)}^{E}\right) \in \mathbb{R}^{D} \quad (8)$$

The final video classes are predicted using this representation with a single hidden layer MLP put on top of it.

By substituting spatial attention for the spatiotemporal attention in Eq. 5 and focusing exclusively on the current frame, ETMLP can minimize the computing cost. The temporal dependencies between frames are not taken into account by such a model. The findings show that, compared to full spatiotemporal attention, strategy has worse classification accuracy, especially on benchmarks that call on robust temporal modeling. The ETMLP suggest "Divided Space-Time Attention" architecture for spatiotemporal attention that is more effective since it applies temporal attention and spatial attention in separate, sequential applications.

Each patch (p, t) is compared to identical spatial location in the other frames; ETMLP determines temporal attention for divided attention inside each block e:

$$sW_{(p,t)}^{(e,ma)^{time}} = SA\left(\frac{Qr_{(p,t)}^{(e,ma)}}{\sqrt{D_h}} \cdot \left[K_{(0,0)}^{(e,ma)} \left\{K_{(p,t')}^{(e,ma)}\right\}_{t'=1,2,\ldots,F}\right]\right) \quad (9)$$

The encoding $eV'^{(e)time}_{(p,t)}$ generated from Eq. 7, the computation of spatial attention using temporal attention is then returned back rather than being passed to the MLP.

Algorithm: ETMLP
Input: Input Video Frames *f*, Class *C*
Output: Multiple activity human prediction result
Preparation:

1. Video frame extraction
2. Human Detection using Graph Network Model (GNM)
3. Detecting Human Prediction Score
4. Feature Extraction using EfficientNET with Heat map method
5. Activity Recognition using Enhanced TimeSformer Model with Multi-Layer Perceptron (ETMLP) Neural Network classification method
6. Compute Evaluation Time

Steps:
While (frames in video)

1. Frame *f* ← Video frame extraction
2. *H* ← Detecting only Humans using GNM method
3. *Hmap* ← Detecting optical flow map estimation between two frames
4. **for** *k* = 1 to *m* **do** // where *m* is collection of frames
    a. *f*(k) ← test video frame.
    b. *H(f)* ← GNM Model using equation 2 // *Human Detection portions*
    c. *ROI(H(f))* ← ROI Marked Humans by EfficientNET Model // *Human Detection with multiple human pose estimation*
    d. *C* ← Prediction Class label with *ROI(H(f))* using (ETMLP)
5. Predicted activity ← Result class label *C*
6. Show the expected activity class in a frame that has a ROI.

End for

End While

Table 1. Parameters Description [26]

Parameters	Symbol	Value
Overall Class	C	27 action classes
Sample Duration	Sample_duration	Each iteration has 16 frames.
Sample Size	Sample_size	Pixel wide at 112
Frame Size	*Fs*	224 × 224
Input video frame	*f*	-
tlooping variable	*t*	1 to *n*
n	*n*	Total frames

4. Experimental Results

On a PC running Windows 10 with Python 3.7 simulations, an Intel I5-6500U series CPU operating at 2.71 GHz, along with 8GB of main memory, is used to implement the ETMLP results. Real-time surveillance video data from a Bangalore residence is used to implement this paper. Using ETMLP training data with several classes on real-time video datasets, predict multiple human activities. The following graphs represent the stand and walking activity parameters and the overall ETMLP training accuracy with loss. Figure 2 (a) and (b) shows the multiple human activity recognition results.

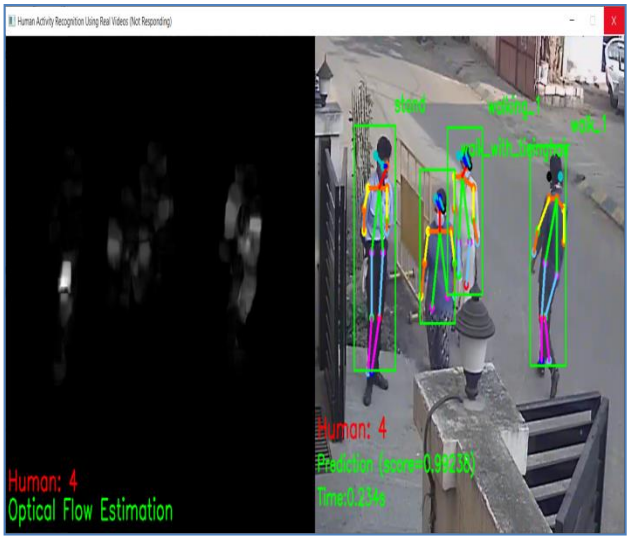


Figure 2. (a) Multiple Activity of HAR result of Stand and walking video with 4 humans of prediction score and Evaluation time

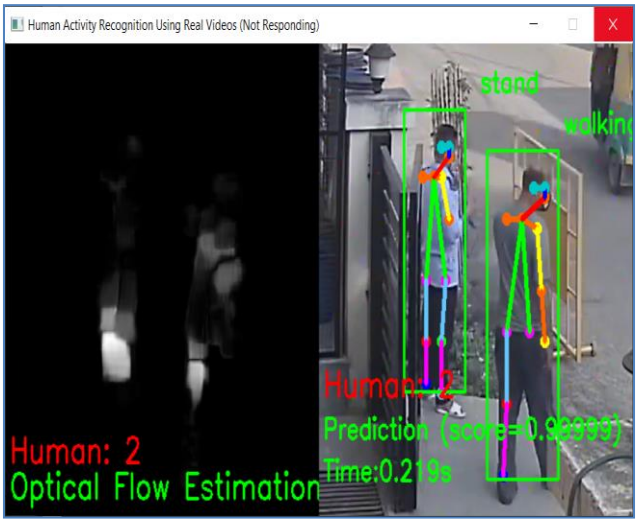


Figure 2. (b) Multiple Activity of HAR results of Stand and walking video with 2 humans of prediction score and Evaluation time

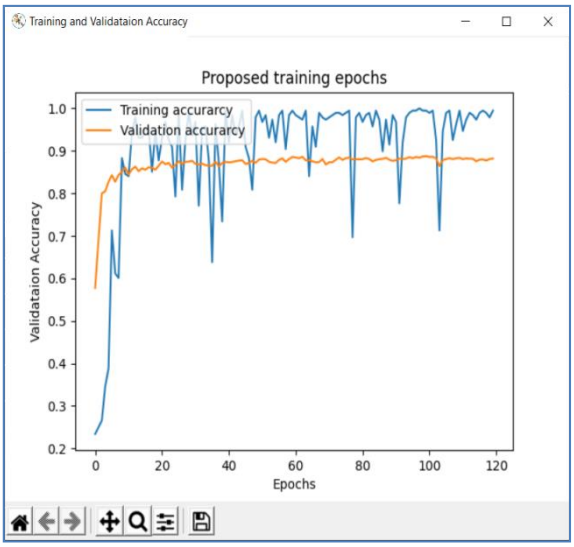


Figure 3. Overall ETMLP training validation accuracy plot result

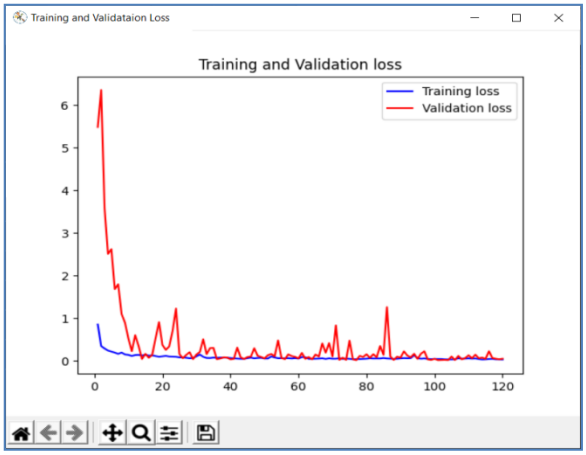


Figure 4. Overall ETMLP training validation Loss plot result

Table 2 shows the overall classification accuracy of existing Bidir-LSTM [28], DSC-BLSTM [27], BGRU-LSTM [30] and Proposed ETMLP classification is shown in Figure 5.

Table 2. Comparison of Real-Time CCTV Video Dataset Accuracy and F1 Score Using Existing DSC-BSLTM, BGRU-LSTM, and Proposed ETMLP Algorithm

Methods	**Bidir-LSTM**	**DSC-BLSTM**	**BGRU-LSTM**	**ETMLP**
Accuracy	91.9	93.8	97.6	**98.9**
F1-Score	91.1	92.637	97.62	**99.2**

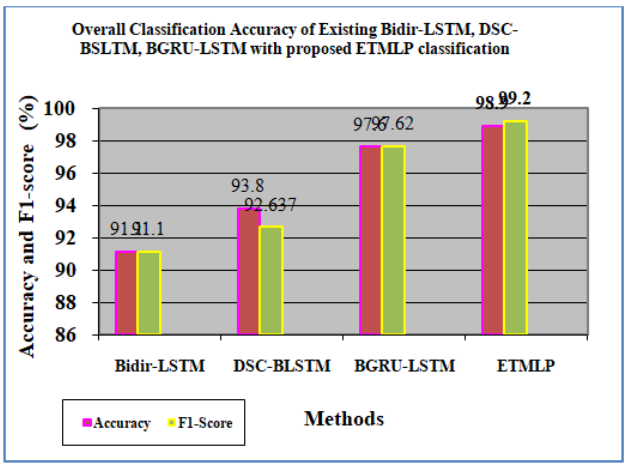


Figure 5. Comparison of Classification accuracy of Bidir-LSTM, DSC-BSLTM, BGRU-LSTM with proposed ETMP classification algorithms

It could be observed from figure 5 that as the multiple human activity real-time datasets, the Accuracy and F1-score of the proposed ETMLP algorithm performs better compared to that of the existing algorithms of Bidri-LSTM, DSC-BLSTM, and BGRU-LSTM. The proposed ETMLP algorithm provides better Accuracy rate for real-time CCTV video datasets. It also shows that with human prediction of features performance gets better in all these datasets. So, it can be concluded that the proposed ETMLP algorithm works well on low and as well as high resolution video data sets.

Figure 6. shows the confusion matrix of proposed ETMLP true label of 27 classes of real-time surveillance video dataset.

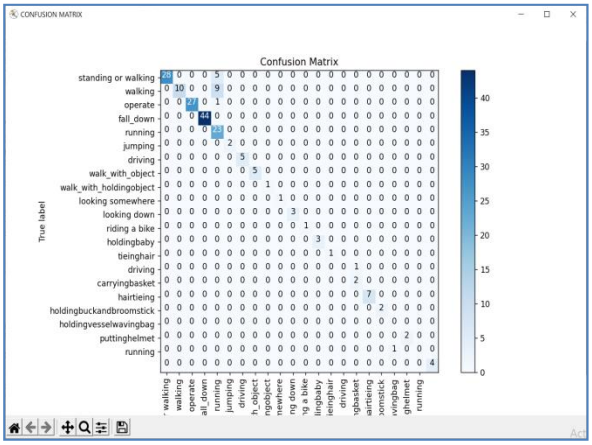


Figure 6. Proposed ETMLP Confusion matrix plot result

5. Conclusion

In this paper, developed Enhanced TimeSformer Model with Multi-Layer Perceptron (ETMLP) Neural Network classification model for video activity recognition as an alternative to the well-known convolution-based video networks paradigm. Incorporating numerous real-time video datasets into a single and multi-task learning paradigm may be advantageous, according to the findings. To maintain solid spatial representations during fine-tuning, to emphasize the significance of continuing to learn on video data. The empirical results show that ETMLP can develop a multi-model that delivers excellent performance across a wide range of real-time video datasets without requiring an additional step of fine-tuning on each downstream application.

References

[1] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012

[2] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in 2011 International conference on computer vision. IEEE, 2011, pp. 2556–2563.

[3] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in Proceedings of the ieee conference on computer vision and pattern recognition, 2015, pp. 961–970.

[4] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," 2016.

[5] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[6] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar et al., "Ava: A video dataset of spatio-temporally localized atomic visual actions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6047–6056.

[7] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag et al., "The "something something" video database for learning and evaluating visual common sense," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 5842–5850.

[8] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lu˘ci´c, and C. Schmid, "Vivit: A video vision transformer," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6836–6846.

[9] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "Tea: Temporal excitation and aggregation for action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 909–918.

[10] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev et al., "The kinetics human action video dataset," 2017.

[11] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? NeurIPS, 2021.

[12] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. ICCV, 2021.

[13] Fan Haoqi, Xiong Bo, Mangalam Karttikeya, Li Yanghao, Yan Zhicheng, Malik Jitendra, and Feichtenhofer Christoph. Multiscale vision transformers, 2021.

[14] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. In NeurIPS, 2020.

[15] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In CVPR, 2017

[16] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William YangWang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. arXiv preprint arXiv:2106.04632, 2021.

[17] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lu˘ci´c, and Cordelia Schmid. Vivit: A video vision transformer. ICCV, 2021.

[18] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017

[19] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In CVPR, 2017

[20] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visuallinguistic representations. ICLR, 2020

[21] Girdhar, R., Carreira, J., Doersch, C., and Zisserman, A. Video action transformer network. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019.

[22] Feichtenhofer, C. X3d: Expanding architectures for efficient video recognition. CVPR, pp. 200–210, 2020.

[23] Bertasius, G. and Torresani, L. Classifying, segmenting, and tracking object instances in video with mask propagation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[24] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In European Conference Computer Vision (ECCV), 2020.

[25] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. CoRR, 2020.

[26] Sevilla-Lara, L., Zha, S., Yan, Z., Goswami, V., Feiszli, M., and Torresani, L. Only time can tell: 24 Discovering temporal data for temporal modeling. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 535–544, January 2021.

[27] Jitha Janardhanan and S. Umamaheswari, "Vision based Human Activity Recognition using Deep Neural Network Framework", International Journal of Advanced Computer Science and Applications(IJACSA), Volume 13 Issue 6, 2022.

[28] Yu Zhao, Rennong Yang, Guillaume Chevalier, Ximeng Xu, and Zhenxing Zhang, “Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors”, Mathematical Problems in Engineering, ,Volume 2018.

[29] Ba, L. J., Kiros, J. R., and Hinton, G. E. Layer normalization. CoRR, 2016.

[30] Jitha Janardhanan and S. Umamaheswari, “Recognizing Multiple Human Activities Using Deep Learning Framework”, Revue d'Intelligence Artificielle,International Inforamtion and Engineering Technology Association (IIETA), ACCEPTED