Skeletal keypoint-based transformer model for human action recognition in aerial videos

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 November 2023, accepted 5 January 2024, date of publication 15 January 2024, date of current version 24 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354389

RESEARCH ARTICLE

Skeletal Keypoint-Based Transformer Model for Human Action Recognition in Aerial Videos

SHAHAB UDDIN[1,2], TAHIR NAWAZ[1,2], JAMES FERRYMAN[3], (Member, IEEE), NASIR RASHID[1,2], MD. ASADUZZAMAN[4], (Senior Member, IEEE), AND RAHEEL NAWAZ[4]

[1]College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan
[2]National Centre of Robotics and Automation (NCRA), Islamabad 44000, Pakistan
[3]School of Mathematical, Physical and Computational Sciences, University of Reading, RG6 6AH Reading, U.K.
[4]School of Digital, Technologies and Arts, Staffordshire University, ST4 2DE Stoke-on-Trent, U.K.

Corresponding author: Tahir Nawaz (tahir.nawaz@ceme.nust.edu.pk)

This work was supported by the Higher Education Commission of Pakistan under Grant DF 1009-0031.

ABSTRACT Several efforts have been made to develop effective and robust vision-based solutions for human action recognition in aerial videos. Generally, the existing methods rely on the extraction of either spatial features (patch-based methods) or skeletal key points (pose-based methods) that are fed to a classifier. Unlike the patch-based methods, the pose-based methods are generally regarded to be more robust to background changes and computationally efficient. Moreover, at the classification stage, the use of deep networks has generated significant interest within the community; however, the need remains to develop accurate and computationally effective deep learning-based solutions. To this end, this paper proposes a lightweight Transformer network-based method for human action recognition in aerial videos using the skeletal keypoints extracted using YOLOv8. The effectiveness of the proposed method is shown on a well-known public dataset containing 13 action classes, achieving very encouraging performance in terms of accuracy and computational cost as compared to several existing related methods.

INDEX TERMS Action recognition, transformer network, aerial videos, video surveillance.

I. INTRODUCTION

Human Action Recognition focuses on understanding human behavior and has been an active topic among researchers. Indeed, it has diverse applications including human-machine interface [1], [2], motion tracking [3], video surveillance [4], [5], and crowd monitoring [6], [7].

Traditional methods for action recognition used various sensing modalities, including accelerometers, magnetometers, and gyroscopes, to capture body movements, frequency of motion, angles and orientation of body parts, velocity, and acceleration along with some other advance features [8], [9], [10], [11]. Although these methods are computationally efficient, robust to noise and illumination changes, and easily implementable, they are limited in terms of their scalability, accuracy and adaptability as compared to the computer vision-based methods. With the availability of large image datasets, the use of computer vision has been the trending choice for action recognition [12], [13], [14].

Specifically, vision-based action recognition methods are classified into two main types: patch-based and pose-based. *Patch-based methods* are generally based on the extraction of spatial features at frame level, which are further processed to extract temporal dependencies across the video sequence [12], [13], [14]. A limitation of the patch-based approaches is that they generally have a higher computational cost associated with feature extraction. *Pose-based methods* instead rely on the use of 2D skeleton data, which provides an outline of the human body joints without involving scene context, for action recognition methods [15], [16], [17]. These methods are generally considered to be more robust to background changes and can inherently better represent bodily movements than patch-based methods. Additionally, recent advancements in pose estimation techniques [18], [19],

The associate editor coordinating the review of this manuscript and approving it for publication was Long Xu.

[20] have made it easier to obtain accurate points of human joints, even when they are difficult to distinguish or are obscured. Moreover, processing skeleton data also requires lesser computational resources and has a lower training time as compared to the patch-based methods.

Indeed, there has been a lot of interest among research community in employing deep learning-based models for human action recognition using pose information. Some methods have been proposed that are built on Transformer-based models [21], [22] to solve the problem. Other approaches [23], [24] relied on using Graph Convolutional Network (GCN) for extracting temporal dependencies and demonstrated encouraging performance. However, these methods [21], [22], [23], [24] assumed fixed camera settings and may not be directly applicable for the case of aerial videos (with top-downish view) due to significant viewpoint changes plus the movement of UAVs could cause motion blur.

To this end, this paper proposes an efficient and light-weight deep learning-based model for human action recognition in aerial videos. The proposed method adopts a two-stage approach. The first stage is based on extracting skeletal keypoints using YOLOv8 pose extractor. In the second stage, the extracted keypoints are then fed to the Transformer network to train it on a wide variety of action types. Indeed, the use of the Transformer-based model with skeletal keypoints for aerial videos has been largely unexplored. We evaluated the usefulness of the proposed method on a well-known public dataset that contains a wide variety of action types and assessed the performance and computational complexity with encouraging results as compared to several existing related methods.

The specific contributions of this work are listed below:

1. An efficient and light-weight Transformer based model has been presented for vision-based human aerial action.
2. A two-stage method is adopted in which the first step involves extracting 2D skeletal keypoints using YOLOv8 and the second step performs training and testing on a Transformer network for varying action types. Indeed, the use of Transformer network for aerial action recognition in videos with skeletal keypoints is not well explored.
3. The effectiveness and efficiency of the proposed method demonstrated on a public dataset containing a variety of action types with a superior performance than existing related methods.

II. RELATED WORK

There exist several methods that are based on using traditional machine learning approaches with manual feature crafting for action recognition; however, they have limitations in terms of a trade-off between performance accuracy and computational cost. For example, in [25] the authors extracted skeletal keypoints using Kinect sensor and then used Hidden Markov Models to find the temporal relations for action recognition. The authors in [26] utilized optical flow for the extraction of motion features, which are then fed to SVM to perform classification. Ohn-Bar and Trivedi [27] used skeletal data with Histogram of Oriented Gradients (HOG) for feature description to perform classification of various action types with SVM.

With the advancements in deep learning and the availability of large datasets, most traditional approaches towards action recognition have become less desirable. In [28] and [29], the authors employed two-streamed network that used 2D CNNs on individual frames followed by a 1D module to aggregate the per-frame features. These methods, although effective, are limited in their ability to encode temporal information due to the use of 2D CNNs. Alternatively, the authors of [30] jointly modeled spatial and temporal information by using 3D CNNs. Other modifications of 3D CNNs such as inflating 2D convolution kernels [31] or decomposing 3D convolution kernels [32] were proposed to improve the performance. Sultani and Shah [33] utilized a disjoint multi-task learning approach based on 3D CNNs to address the action recognition task, when there is an availability of a small dataset. They used the game data of GTA and FIFA along with GAN generated aerial data from actual ground data for training, and then the model is tested on real aerial data. Kotecha et al. [34] designed a Faster Motion Feature Modeling (FMFM) based system with Accurate Action Recognition (AAR) modeling. Their proposed system used a cascade of CNN-based models for both FMFM and AAR. Mliki et al. [35] developed a CNN based algorithm that used AlexNet [36] for detection and GoogleNet [37] for activity classification with ten classes. The authors in [38] proposed a model that used VGG16 [39] for CNN-based feature extraction in color and optical flow images and the Lattice LSTM for classification of temporal dependencies. Wang et al. [40] also introduced an action recognition framework named Temporal Segment Network (TSN), which divided videos into equal-length segments. Then, a sequence of snippets is created from these segments, which can be of variable length. A consensus function aggregates the outputs of all the snippets to create the final class hypothesis. In [41], the authors designed an onboard UAV model for ten different gestures, which used YOLOv3-tiny for human detection, OpenPose [18] for pose estimation, and DNN for gesture classification. They used their own data for training and evaluation. In [42], Ahmad et al. used YOLOv5 for object detection in frames and Stochastic Gradient Boosting for action classification with 12 different action types [43]. Ding et al. [44] proposed a lightweight model for action recognition in aerial videos. They employed a TCN based method, which used MobileNetV3 as feature descriptor and attention module for finding temporal relations among the frames. The authors in [45] presented an approach towards action recognition by using semi-supervised and unsupervised domain adaptation. Srivastava et al. [46] proposed a system for violent action detection using Part Affinity Fields [18] for pose estimation and SVM with RBF Kernel for classification. They also created their own data for training and evaluation.

Most of the above-mentioned methods used CNN models for the extraction of spatial features and, in some cases, temporal features as well, and generally have a higher computation complexity and cost; hence, requiring powerful GPUs. This makes them less deployable in real-world applications involving aerial camera settings. Moreover, the use of Transformer networks [47] is growing with encouraging performance in several vision tasks [48], but relatively less explored for solving the human action recognition problem.

III. PROPOSED TRANSFORMER-BASED ACTION RECOGNITION METHOD

The proposed method uses skeletal body keypoints for pose estimation that are extracted using YOLOv8. These keypoints are preprocessed to make them compatible to be fed to the Transformer network for training and testing. The use of the Transformer-based network is inspired from an earlier work [22] that was aimed at ground-based fixed camera setting. The proposed method involved architectural changes including data augmentation and removal of dropout layers to adapt it for the application at hand.

A. POSE ESTIMATION

We employed YOLOv8 pose extractor, which provides 17 keypoints of the whole body. Compared to other pose extractors (OpenPose [18], YOLOv7 [49], EfficientPose [50]), we practically observed that YOLOv8 pose extractor is faster and more accurate. Figure 1 illustrates the extracted keypoints on a sample image in which a person is performing a kicking action. Each input video to the pose extractor has the form of *(T, H, W, C),* where T is the number of frames; H, W and C are the height, width, and number of channels in the video. The pose extractor returns the output in the form of *(T, P)* for each video, where P represents the extracted keypoints, After the keypoints have been extracted, they are preprocessed to be fed to the Transformer model for training and classification.

B. TRANSFORMER NETWORK

The architecture of the Transformer encoder layer is shown in Figure 2. The encoder layer is repeated multiple times, depending upon the requirement and architecture. This model was originally developed for language processing to perform task like Neural Machine Translation. It is very efficient in terms of keeping track of temporal dependencies in long sequences of data. The primary block responsible for memory or temporal relation is the Self Attention block. This block finds the temporal relation of every instance with every other instance. Figure 3 shows different steps of calculating Self Attention. Q, K and V are linearly transformed embedding vectors (or matrices if stacked) of the input instances. Matrix multiplication of Q and K matrices is calculated, which is then scaled as shown in the Figure 3. The scaled values are then passed through a SoftMax layer, whose output is used to calculate the final matrix multiplication with V matrix.

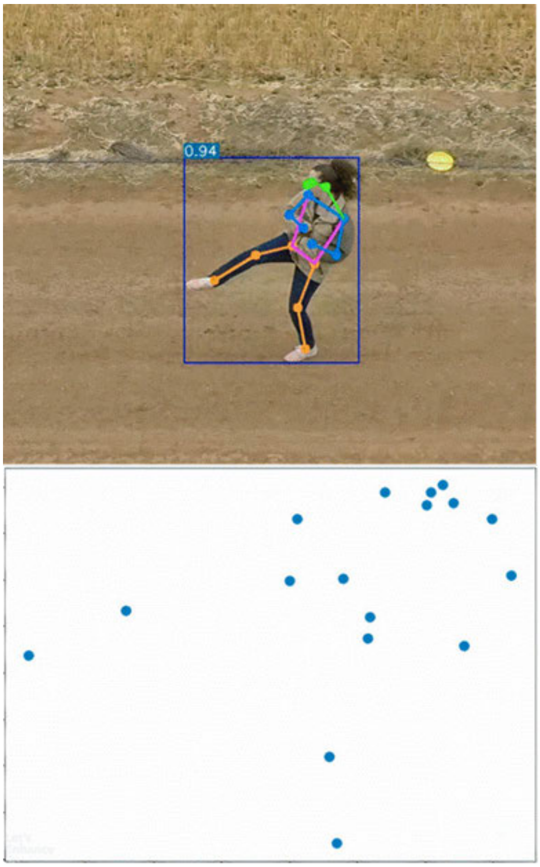


FIGURE 1. Top: Extracted keypoints of the whole body for the 'Kicking' action on a sample image. Bottom: Extracted keypoints shown in the form of a plot.

The pre-processed keypoints of each action are divided into S_K sequences, where each sequence has the form of *(T, P)*; where T is set to 30 in our case and P represents the keypoints as follows:

$$A = (S_1, S_2 \ldots\ldots, S_K)\,; \quad where\ S_k = (T_{k,i}, P_{k,i,j}) \tag{1}$$

Here, i is the number of frames in a sequence and j is the number of keypoints in each frame. The Transformer model will extract the temporal features from 30 consecutive frames of each sequence. The keypoints of the frames are first linearly transformed into an Embedding Matrix and are added with an additional Positional Embedding Matrix, which provides positional information of each frame, creating X_{emb}. The dimension of X_{emb} becomes *(T,* d_{model}*)*, where d_{model} is the embedding dimension of each vector (row). The positional Embedding Matrix has learnable parameters. X_{emb} is then used to create Q, K and V matrices as shown below:

$$Q = X_{emb_b} W_Q, \tag{2}$$

$$K = X_{emb_b} W_K, \tag{3}$$

$$V = X_{emb_b} W_V. \tag{4}$$

W_Q, W_K and W_V have learnable parameters and their dimensions are usually the same i.e., $d_q = d_k = d_v$. So, the

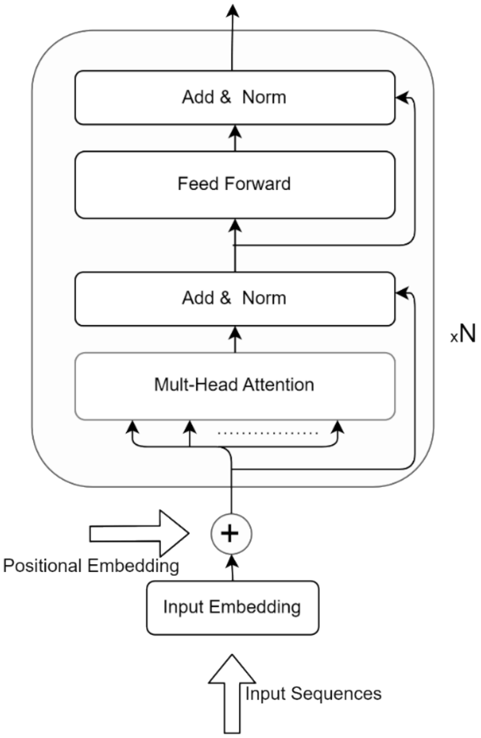


FIGURE 2. Key building blocks of the transformer encoder layer.

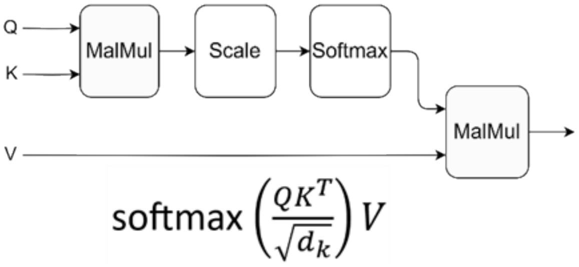


FIGURE 3. Block diagram illustrating different steps required for self-attention.

dimensions of Q, K and V becomes $(T, d_q = d_k = d_v)$, where d is the embedding size of keypoints of each frame. In our case, $d_{model} = d_q = d_k = d_v$, which is 26. Q, K and V matrices are used to perform attention as shown in Figure 3. This process of creating Q, K, V and attention is repeated h times, where h is the number of heads used in the model. Then the results of all the heads are concatenated and are transformed again by another layer through W_0 which has the dimension of (hd_v, d_{model}). So, the output dimension of multi-head attention becomes (T, hd_v). This output is then passed to a feed forward network, which linearly transforms it by the following operations:

$$FF(x) = \max(0, (xW_1 + b_1))\, W_2 + b_2, \tag{5}$$

where W_1 and W_2 has dimensions of (d_{model}, d_{ff}) and (d_{ff}, d_{model}) respectively, and x is the output of the multi heads. We chose $d_{ff} = 4d_{model}$. This whole operation is illustrated in Figure 4. This encoder layer is repeated multiple times.

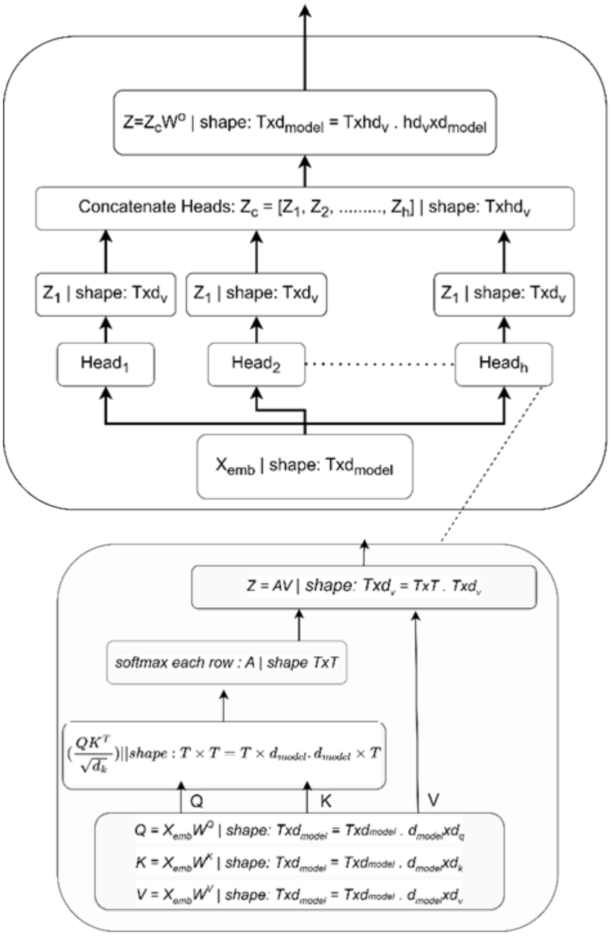

Z=Z$_c$W^O | shape: Txd$_{model}$ = Txhd$_v$. hd$_v$xd$_{model}$
Concatenate Heads: Z$_c$ = [Z$_1$, Z$_2$, ..., Z$_h$] | shape: Txhd$_v$
Z$_1$ | shape: Txd$_v$
Z$_1$ | shape: Txd$_v$
Z$_1$ | shape: Txd$_v$
Head$_1$
Head$_2$
Head$_h$
X$_{emb}$ | shape: Txd$_{model}$
Z = AV | shape: Txd$_v$ = TxT . Txd$_v$
softmax each row : A | shape TxT
$(\frac{QK^T}{\sqrt{d_k}})||shape : T \times T = T \times d_{model}.d_{model} \times T$
Q
K
V
Q = X$_{emb}$W^Q | shape: Txd$_{model}$ = Txd$_{model}$. d$_{model}$xd$_q$
K = X$_{emb}$W^K | shape: Txd$_{model}$ = Txd$_{model}$. d$_{model}$xd$_k$
V = X$_{emb}$W^V | shape: Txd$_{model}$ = Txd$_{model}$. d$_{model}$xd$_v$

FIGURE 4. Data flow and the dimensions of matrices at each step of the encoder layer.

IV. EXPERIMENTAL VALIDATION

This section presents the experimental validation and analysis of the proposed method, including the description of the dataset followed by the analysis of results.

A. DATASET

We used a well-known publicly available dataset for evaluation, the Drone-Action dataset [51]. This dataset contains 13 classes and a total of 240 high resolution (1920 × 1080) videos with 25 frames per second. It is recorded in an outdoor environment with a camera mounted on a low altitude and low speed drone. Also, it has used 10 different actors so as to introduce a level of diversity. The dataset was collected on an unsettled road in the midst of a wheat field from varying top-downish viewpoints. The background wheat field can also pose a challenge (background clutter) to the CNN-based feature extraction approaches. The dataset provides three different splits of training and test datasets, referred to as Split 1, Split 2, and Split 3. Figure 5 shows some

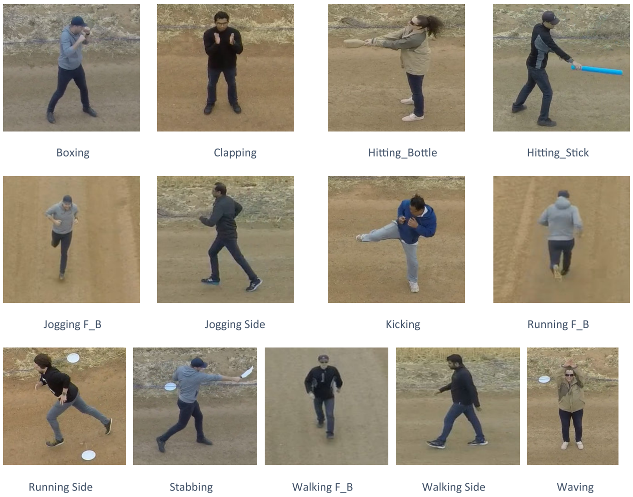


FIGURE 5. Sample images for each of the 13 action classes of Drone-Action dataset, as used in the experimental evaluation.

TABLE 1. Hyperparameters of the proposed Transformer model for training and testing.

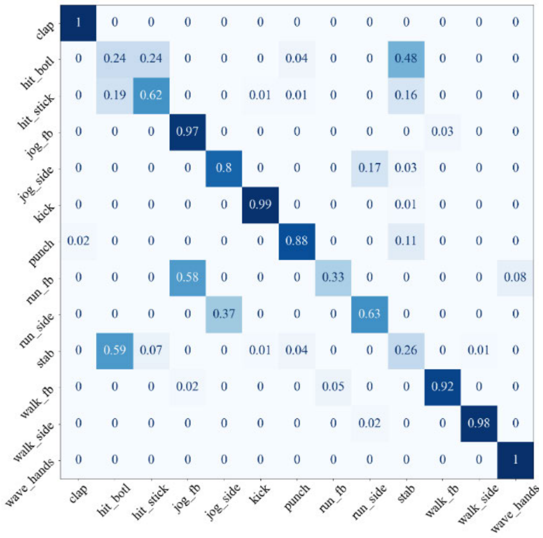
Training epochs	100
Batch size	128
Optimizer	AdamW
Weight decay	0.0001
Dropout	0.3
d_{model}	26
d_{mlp}	128

representative images, showing all action classes used in the evaluation.

B. RESULTS & ANALYSIS

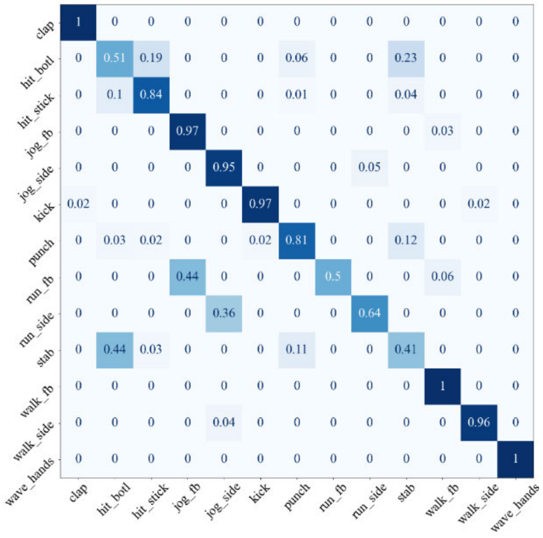
We performed a detailed evaluation of the proposed method on Drone Action dataset. We experimented with varying number of Transformer encoder layers and reported the results accordingly. Also, we have performed data augmentation by flipping frame keypoints horizontally (along y axis) to raise the number of training samples. Table 1 shows the hyperparameters of the Transformer model.

Figures 6, 7, 8 show the confusion matrices for the three splits with four layers encoder architecture based on the experimental evidence given in Table 1, as discussed below in this section. It is clear from the figure that the proposed model architecture shows quite encouraging performance for all classes except for 'Hit_Bottle', 'Hit_Stick', and 'Stab'. This is due to the fact that, in each of these three classes, the actions performed appear quite similar with different object in hand, i.e., bottle, stick and knife. And since the pose estimator extracts keypoints of only body joints, and not the objects being carried, these classes are difficult to distinguish. Also, the model confuses between actions of 'Running_fb' and 'Jogging_fb', which appear quite alike too.



	clap	hit_botl	hit_stick	jog_fb	jog_side	kick	punch	run_fb	run_side	stab	walk_fb	walk_side	wave_hands
clap	1	0	0	0	0	0	0	0	0	0	0	0	0
hit_botl	0	0.24	0.24	0	0	0	0.04	0	0	0.48	0	0	0
hit_stick	0	0.19	0.62	0	0	0.01	0.01	0	0	0.16	0	0	0
jog_fb	0	0	0	0.97	0	0	0	0	0	0	0.03	0	0
jog_side	0	0	0	0	0.8	0	0	0	0.17	0.03	0	0	0
kick	0	0	0	0	0	0.99	0	0	0	0.01	0	0	0
punch	0.02	0	0	0	0	0	0.88	0	0	0.11	0	0	0
run_fb	0	0	0	0.58	0	0	0	0.33	0	0	0	0	0.08
run_side	0	0	0	0	0.37	0	0	0	0.63	0	0	0	0
stab	0	0.59	0.07	0	0	0.01	0.04	0	0	0.26	0	0.01	0
walk_fb	0	0	0	0.02	0	0	0	0.05	0	0	0.92	0	0
walk_side	0	0	0	0	0	0	0	0	0.02	0	0	0.98	0
wave_hands	0	0	0	0	0	0	0	0	0	0	0	0	1

FIGURE 6. Confusion matrix with split 1 (four encoder layers).



	clap	hit_botl	hit_stick	jog_fb	jog_side	kick	punch	run_fb	run_side	stab	walk_fb	walk_side	wave_hands
clap	1	0	0	0	0	0	0	0	0	0	0	0	0
hit_botl	0	0.51	0.19	0	0	0	0.06	0	0	0.23	0	0	0
hit_stick	0	0.1	0.84	0	0	0	0.01	0	0	0.04	0	0	0
jog_fb	0	0	0	0.97	0	0	0	0	0	0	0.03	0	0
jog_side	0	0	0	0	0.95	0	0	0	0.05	0	0	0	0
kick	0.02	0	0	0	0	0.97	0	0	0	0	0	0.02	0
punch	0	0.03	0.02	0	0	0.02	0.81	0	0	0.12	0	0	0
run_fb	0	0	0	0.44	0	0	0	0.5	0	0	0.06	0	0
run_side	0	0	0	0	0.36	0	0	0	0.64	0	0	0	0
stab	0	0.44	0.03	0	0	0	0.11	0	0	0.41	0	0	0
walk_fb	0	0	0	0	0	0	0	0	0	0	1	0	0
walk_side	0	0	0	0	0.04	0	0	0	0	0	0	0.96	0
wave_hands	0	0	0	0	0	0	0	0	0	0	0	0	1

FIGURE 7. Confusion matrix with split 2 (four encoder layers).

The obtained performance by the proposed framework is reported separately for each split using the standard evaluation measures: Precision, Recall, F1-score, and Accuracy (Table 2). The results show that the performance is generally encouraging, considering the number and diversity of action classes under consideration. A point to highlight is that the best performance is mostly obtained with four layers (e.g., see the mean performance scores in Table 2). Overall, the best performance is generally obtained on Split 2.

Table 3 shows the comparison of the proposed method (in the form of the mean performance on all three splits)

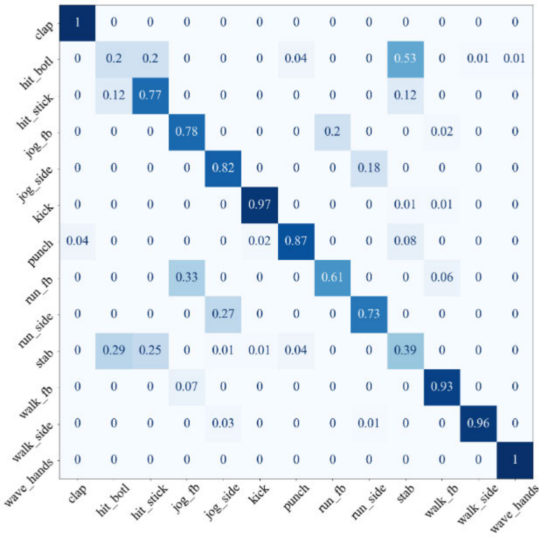


FIGURE 8. Confusion matrix with split 3 (four encoder layers).

TABLE 2. Performance analysis on each split with different number of encoder layers.

# Of encoder layers		1	2	3	4	5
Split 1	Precision	0.742	0.755	0.719	0.755	0.806
	Recall	0.714	0.743	0.719	0.741	0.764
	F1 Score	0.720	0.748	0.717	0.742	0.771
	Accuracy (%)	68.55	70.48	70.16	71.61	73.71
Split 2	Precision	0.788	0.768	0.844	0.839	0.818
	Recall	0.752	0.760	0.830	0.812	0.788
	F1 Score	0.759	0.759	0.832	0.817	0.780
	Accuracy (%)	73.43	73.75	79.71	80.35	77.62
Split 3	Precision	0.715	0.711	0.734	0.749	0.744
	Recall	0.718	0.714	0.745	0.771	0.751
	F1 Score	0.716	0.711	0.732	0.756	0.741
	Accuracy (%)	70.69	69.90	71.47	74.29	71.94
Mean	Precision	0.748	0.745	0.766	0.781	0.789
	Recall	0.728	0.739	0.765	0.775	0.767
	F1 Score	0.732	0.739	0.760	0.772	0.764
	Accuracy (%)	70.89	71.38	73.78	75.42	74.42

with the approaches (High-Level Pose Features based method (HLPF) and Pose-based Convolutional Neural Networks (P-CNN)) that are reported in the original dataset paper [51], as well as a recent related method that used YOLOv8 pose extractor in combination with the Long Short-Term Memory (LSTM) network [52] for action recognition. It is evident that the proposed method shows the best performance in terms of Precision, Recall and F1-score, and a comparable performance in terms of Accuracy as compared to existing methods.

TABLE 3. Comparison of the mean performance of the proposed method with the existing methods.

Model	Precision	Recall	F1- score	Accuracy
HLPF	0.66	0.63	0.63	64.36%
P-CNN	0.77	0.77	0.77	75.92%
Pose+LSTM	0.77	0.77	0.76	74.67%
Proposed	0.78	0.77	0.77	75.42%

TABLE 4. Comparison of complexity in terms of inference time as well as accuracy of Transformer model with several existing models.

Model	Accuracy (%))	Inference Time (millisecond/sequence)
3D ResNet	64.00	41.87
ST-GCN	60.45	29.67
ResNet101	63.75	11.45
ResNet18	66.85	2.23
LSTM	68.69	0.71
Transformer	75.42	0.56

TABLE 5. Comparison of the computational complexity of the proposed transformer-based method with recent state-of-the-art action recognition methods in terms of the number of parameters and the floating-point operations (FLOPS).

Model	Parameters (x10^6)	Flops (x10^9)
PoTion [53]	4.75	0.60
PA3D [54]	4.81	0.65
Pose-SlowOnly [15]	2.00	15.9
Pose-X3D-s [15]	0.24	0.60
EfficientGCN-B4 [55]	2.03	15.24
Transformer	0.04	0.000011

For a more holistic performance comparisons, in Table 4, we have provided a comparison of the proposed model with several other deep learning models (3DResNet, ST-GCN, ResNet101, ResNet18, LSTM) in terms of performance accuracy and inference time per sequence of 30 frames. Here, we practically implemented all of these models on Intel Core i3-5005U processors (two physical cores of 2.0 GHz each) and 4GB of RAM. The proposed model took approximately 7 minutes for 1 run of training for 100 epochs. The results show that the proposed model outperforms existing models both in terms of accuracy and inference time.

In Table 5, we also compared the computational complexity of the proposed Transformer-based model with several related state-of-the-art models based on the number of network parameters (in millions) as well as the number of floating-point operations (FLOPS) (in billions). It is evident that the proposed method performs better than all of the existing methods.

V. CONCLUSION

In this paper we presented an effective and efficient Transformer-based model that used the skeletal (target pose) information for human action recognition in aerial videos. We utilized the lightweight attention module for action classification without the use of CNNs in order to reduce the computational cost and complexity. The skeletal keypoints are extracted using YOLOv8 pose estimator, which are fed into the Transformer network. The results show that the proposed method achieved very encouraging performance when compared to existing related methods. The key strength of the proposed method is that the computational complexity is significantly lower as compared to several related methods. This is expected to substantially reduce the computational cost, making it more deployable in real-world applications.

REFERENCES

[1] M. Zhu, T. He, and C. Lee, "Technologies toward next generation human machine interfaces: From machine learning enhanced tactile sensing to neuromorphic sensory systems," *Appl. Phys. Rev.*, vol. 7, no. 3, Sep. 2020, Art. no. 031305, doi: 10.1063/5.0016485.

[2] L. Lo Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognit.*, vol. 53, pp. 130–147, May 2016, doi: 10.1016/J.PATCOG.2015.11.019.

[3] S. N. Paul and Y. J. Singh, "Survey on video analysis of human walking motion," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 7, no. 3, pp. 99–122, Jun. 2014, doi: 10.14257/IJSIP.2014.7.3.10.

[4] J. Barbedo, "A review on the use of unmanned aerial vehicles and imaging sensors for monitoring and assessing plant stresses," *Drones*, vol. 3, no. 2, p. 40, Apr. 2019, doi: 10.3390/DRONES3020040.

[5] L. Li, T. Nawaz, and J. Ferryman, "Performance analysis and formative assessment of visual trackers using PETS critical infrastructure surveillance datasets," *J. Electron. Imag.*, vol. 28, no. 4, p. 1, Jul. 2019, doi: 10.1117/1.JEI.28.4.043004.

[6] A. Al-Kaff, F. M. Moreno, L. J. S. José, F. García, D. Martín, A. de la Escalera, A. Nieva, and J. L. M. Garcéa, "VBII-UAV: Vision-based infrastructure inspection-UAV," in *Proc. World Conf. Inf. Syst. Technologies*, in Advances in Intelligent Systems and Computing, vol. 570, 2017, pp. 221–231, doi: 10.1007/978-3-319-56538-5_24.

[7] J. Boyle, T. Nawaz, and J. Ferryman, "Using deep Siamese networks for trajectory analysis to extract motion patterns in videos," *Electron. Lett.*, vol. 58, no. 9, pp. 356–359, Apr. 2022, doi: 10.1049/ELL2.12460.

[8] A. Mimouna, A. B. Khalifa, and N. E. B. Amara, "Human action recognition using triaxial accelerometer data: Selective approach," in *Proc. 15th Int. Multi-Conf. Syst., Signals Devices (SSD)*, Mar. 2018, pp. 491–496, doi: 10.1109/SSD.2018.8570429.

[9] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, Dec. 2015, doi: 10.3390/S151229858.

[10] M. Al-Amin, W. Tao, D. Doell, R. Lingard, Z. Yin, M. C. Leu, and R. Qin, "Action recognition in manufacturing assembly using multimodal sensor fusion," *Proc. Manuf.*, vol. 39, pp. 158–167, Jan. 2019, doi: 10.1016/J.PROMFG.2020.01.288.

[11] P. Palimkar, V. Bajaj, A. K. Mal, R. N. Shaw, and A. Ghosh, "Unique action identifier by using magnetometer, accelerometer and gyroscope: KNN approach," in *Proc. ICACIT*, vol. 218. Singapore: Springer, 2021, pp. 607–631, doi: 10.1007/978-981-16-2164-2_48.

[12] J. Wu, W. Luo, W. Liu, and C. Zhang, "Global and local discriminative patches exploiting for action recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 3667–3671, doi: 10.1109/ICASSP40776.2020.9054282.

[13] S. H. Park, J. Tack, B. Heo, J. W. Ha, and J. Shin, "K-centered patch sampling for efficient video recognition," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 13695, 2022, pp. 160–176, doi: 10.1007/978-3-031-19833-5_10.

[14] N. Ikizler and P. Duygulu, "Human action recognition using distribution of oriented rectangular patches," in *Proc. Workshop Hum. Motion*, in Lecture Notes in Computer Science: Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 4814, 2007, pp. 271–284, doi: 10.1007/978-3-540-75703-0_19.

[15] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2959–2968, doi: 10.1109/CVPR52688.2022.00298.

[16] S. Sarker, S. Rahman, T. Hossain, S. F. Ahmed, L. Jamal, and M. A. R. Ahad, "Skeleton-based activity recognition: Preprocessing and approaches," *Intell. Syst. Reference Library*, vol. 200, pp. 43–81, Jan. 2021, doi: 10.1007/978-3-030-68590-4_2.

[17] B. Li, C. Tan, J. Wang, R. Qi, P. Qi, and X. Li, "Skeleton-based action recognition with UAV views," in *Proc. 3rd Int. Conf. Video, Signal Image Process.*, Nov. 2021, pp. 16–20, doi: 10.1145/3503961.3503964.

[18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[19] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306, doi: 10.1109/CVPR.2018.00762.

[20] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2353–2362, doi: 10.1109/ICCV.2017.256.

[21] C. Plizzari, M. Cannici, and M. Matteucci, "Skeleton-based action recognition via spatial and temporal transformer networks," *Comput. Vis. Image Understand.*, vols. 208–209, Jul. 2021, Art. no. 103219, doi: 10.1016/J.CVIU.2021.103219.

[22] 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 Recognit.*, vol. 124, Apr. 2022, Art. no. 108487, doi: 10.1016/J.PATCOG.2021.108487.

[23] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1109–1118, doi: 10.1109/CVPR42600.2020.00119.

[24] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027, doi: 10.1109/CVPR.2019.01230.

[25] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27, doi: 10.1109/CVPRW.2012.6239233.

[26] S. Danafar, "Action recognition for surveillance applications using optic flow and SVM," in *Proc. 8th Asian Conf. Comput. Vis.*, vol. 4844. Cham, Switzerland: Springer, 2007, pp. 457–466, doi: 10.1007/978-3-540-76390-1_45.

[27] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 465–470, doi: 10.1109/CVPRW.2013.76.

[28] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941, doi: 10.1109/CVPR.2016.213.

[29] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process Syst.*, Jun. 2014, pp. 568–576.

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.

[31] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733, doi: 10.1109/CVPR.2017.502.

[32] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5534–5542, doi: 10.1109/ICCV.2017.590.

[33] W. Sultani and M. Shah, "Human action recognition in drone videos using a few aerial training examples," *Comput. Vis. Image Understand.*, vol. 206, May 2021, Art. no. 103186, doi: 10.1016/J.CVIU.2021.103186.

[34] K. Kotecha, D. Garg, B. Mishra, P. Narang, and V. K. Mishra, "Background invariant faster motion modeling for drone action recognition," *Drones*, vol. 5, no. 3, p. 87, Aug. 2021, doi: 10.3390/DRONES5030087.

[35] H. Mliki, F. Bouhlel, and M. Hammami, "Human activity recognition from UAV-captured video sequences," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107140, doi: 10.1016/J.PATCOG.2019.107140.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.

[37] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[38] L. Sun, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi, and S. Savarese, "Lattice long short-term memory for human action recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2166–2175, doi: 10.1109/ICCV.2017.236.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, Sep. 2014, pp. 1–14.

[40] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[41] C. Liu and T. Szirányi, "Real-time human detection and gesture recognition for on-board UAV rescue," *Sensors*, vol. 21, no. 6, p. 2180, Mar. 2021, doi: 10.3390/S21062180.

[42] T. Ahmad, M. Cavazza, Y. Matsuo, and H. Prendinger, "Detecting human actions in drone images using YoloV5 and stochastic gradient boosting," *Sensors*, vol. 22, no. 18, p. 7020, Sep. 2022, doi: 10.3390/S22187020.

[43] M. Barekatain, M. Martí, H.-F. Shih, S. Murray, K. Nakayama, Y. Matsuo, and H. Prendinger, "Okutama-action: An aerial view video dataset for concurrent human action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2153–2160, doi: 10.1109/CVPRW.2017.267.

[44] M. Ding, N. Li, Z. Song, R. Zhang, X. Zhang, and H. Zhou, "A lightweight action recognition method for unmanned-aerial-vehicle video," in *Proc. IEEE 3rd Int. Conf. Electron. Commun. Eng. (ICECE)*, Dec. 2020, pp. 181–185, doi: 10.1109/ICECE51594.2020.9353008.

[45] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and semi-supervised domain adaptation for action recognition from drones," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1706–1715, doi: 10.1109/WACV45572.2020.9093511.

[46] A. Srivastava, T. Badal, A. Garg, A. Vidyarthi, and R. Singh, "Recognizing human violent action using drone surveillance within real-time proximity," *J. Real-Time Image Process.*, vol. 18, no. 5, pp. 1851–1863, Oct. 2021, doi: 10.1007/S11554-021-01171-2.

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process Syst.*, vol. 30, 2017, pp. 1–11.

[48] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao, "A survey on vision transformer," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[49] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[50] D. Groos, H. Ramampiaro, and E. A. Ihlen, "EfficientPose: Scalable single-person pose estimation," *Int. J. Speech Technol.*, vol. 51, no. 4, pp. 2518–2533, Apr. 2021, doi: 10.1007/S10489-020-01918-7.

[51] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-action: An outdoor recorded drone video dataset for action recognition," *Drones*, vol. 3, no. 4, p. 82, Nov. 2019, doi: 10.3390/DRONES3040082.

[52] S. M. Saeed, H. Akbar, T. Nawaz, H. Elahi, and U. S. Khan, "Body-pose-guided action recognition with convolutional long short-term memory (LSTM) in aerial videos," *Appl. Sci.*, vol. 13, no. 16, p. 9384, Aug. 2023.

[53] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "PoTion: Pose motion representation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7024–7033, doi: 10.1109/CVPR.2018.00734.

[54] A. Yan, Y. Wang, Z. Li, and Y. Qiao, "PA3D: Pose-action 3D machine for video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7914–7923, doi: 10.1109/CVPR.2019.00811.

[55] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1474–1488, Feb. 2023, doi: 10.1109/TPAMI.2022.3157033.

SHAHAB UDDIN received the Bachelor of Science degree in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan. Currently, he is working as a Research Assistant with the College of Aeronautical Engineering, NUST, Risaplur. He has developed innovative solutions while working on several projects. He participated in many workshops while working with the National Center of Robotics and Automation, National University of Sciences and Technology, Islamabad. His research interests include machine learning and computer vision.

TAHIR NAWAZ received the M.Sc. degree in computer vision and robotics under the Erasmus Mundus Scholarship, a joint master's program from Heriot-Watt University, U.K., the University of Girona, Spain, and the University of Burgundy, France, and the Ph.D. degree with a specialization in computer vision, a joint Doctoral program under the highly prestigious Erasmus Mundus Fellowship from the Queen Mary University of London, U.K., and the Alpen-Adria University of Klagenfurt, Austria. In 2005, he also represented Pakistan as the Team Leader of the Asia–Pacific Broadcasting Union (ABU) Robocon Contest (an international robot competition), Beijing, China. He is currently working as an Associate Professor and the Head of the Department (Research) of the College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Pakistan, where his core interests focus around multi-modal sensing techniques, particularly investigating cutting-edge technologies pertaining to automated video surveillance and autonomous vehicles. He has a strong demonstrated track record of research and development in the areas of computer vision (visible/thermal imagery) and artificial intelligence, with more than 15 years of experience working in academic and industrial sectors across multiple European countries in prestigious organizations. He has published more than 28 articles in prestigious publication venues and has been involved in several international funded projects.

JAMES FERRYMAN (Member, IEEE) is currently a Professor of computational vision with the University of Reading, U.K. His current research interest includes the automatic visual surveillance of wide-area scenes using computational vision. The research has contributed new results in the areas of model-based vision, visual tracking, and surveillance, especially using 3D deformable models.

NASIR RASHID received the B.E. degree (Hons.) in mechanical engineering from the College of Electrical and Mechanical Engineering (EME), Islamabad, Pakistan, in 1993, and the M.S. and Ph.D. degrees in mechatronics engineering from the College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Pakistan. He is currently working as an Associate Professor and the Dean of the College of Electrical and Mechanical Engineering, NUST. His field of specialization is artificial intelligence with applications in biomedical engineering and machine vision. His Ph.D. was in the field of non-invasive brain signal classification (biomedical engineering). He has vast experience as a Professional Engineer in Pakistan. He is a Lifetime Member of the Pakistan Engineering Council.

MD. ASADUZZAMAN (Senior Member, IEEE) received the B.Sc. (Hons.) and M.Sc. degrees in applied statistics from the University of Dhaka, Dhaka, Bangladesh, in 1999 and 2001, respectively, the M.Sc. degree in bioinformatics from the Chalmers University of Technology, Gothenburg, Sweden, in 2007, and the Ph.D. degree in operational research from the University of Westminster, London, U.K., in 2010. He is currently an Associate Professor with the Operational Research, Staffordshire University, Stoke-on-Trent, U.K., where he joined as a Lecturer, in 2014, and was promoted to a Senior Lecturer, in 2017. He has a background in statistics and operational research (OR).

RAHEEL NAWAZ is currently the Pro Vice Chancellor (Digital Transformation) and a Professor with Staffordshire University, U.K. In the past roles, he has led large teaching provisions, research centres, and work-based learning teams. He has extensive experience in establishing international research and teaching collaborations. Prior to his appointment, he spent nearly a decade performing senior roles in the public HE sector (Russell Group and modern universities). Before that, he spent many years in leadership roles in the private further and higher education sector, including top positions in some of the largest private further and higher education organizations in the U.K.