

## Article

# Cascading Pose Features with CNN-LSTM for Multiview Human Action Recognition

Najeeb ur Rehman Malik <sup>1,†</sup>, Syed Abdul Rahman Abu-Bakar <sup>1,†</sup>, Usman Ullah Sheikh <sup>1,†</sup>, Asma Channa <sup>2,†</sup>  
and Nirvana Popescu <sup>2,\*</sup>

<sup>1</sup> Computer Vision, Video and Image Processing Lab, ECE Department, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia

<sup>2</sup> Computer Science Department, University POLITEHNICA of Bucharest, 060042 Bucharest, Romania

\* Correspondence: nirvana.popescu@upb.ro

† Current address: DIIES Department, Mediterranean University of Reggio Calabria, 89100 Reggio Calabria, Italy.

‡ These authors contributed equally to this work.

**Abstract:** Human Action Recognition (HAR) is a branch of computer vision that deals with the identification of human actions at various levels including low level, action level, and interaction level. Previously, a number of HAR algorithms have been proposed based on handcrafted methods for action recognition. However, the handcrafted techniques are inefficient in case of recognizing interaction level actions as they involve complex scenarios. Meanwhile, the traditional deep learning-based approaches take the entire image as an input and later extract volumes of features, which greatly increase the complexity of the systems; hence, resulting in significantly higher computational time and utilization of resources. Therefore, this research focuses on the development of an efficient multi-view interaction level action recognition system using 2D skeleton data with higher accuracy while reducing the computation complexity based on deep learning architecture. The proposed system extracts 2D skeleton data from the dataset using the OpenPose technique. Later, the extracted 2D skeleton features are given as an input directly to the Convolutional Neural Networks and Long Short-Term Memory (CNN-LSTM) architecture for action recognition. To reduce the complexity, instead of passing the whole image, only extracted features are given to the CNN-LSTM architecture, thus eliminating the need for feature extraction. The proposed method was compared with other existing methods, and the outcomes confirm the potential of the proposed technique. The proposed OpenPose-CNNLSTM achieved an accuracy of 94.4% for MCAD (Multi-camera action dataset) and 91.67% for IXMAS (INRIA Xmas Motion Acquisition Sequences). Our proposed method also significantly decreases the computational complexity by reducing the number of inputs features to 50.

**Keywords:** human action recognition (HAR); deep learning; CNN-LSTM



**Citation:** Malik, N.u.R.; Abu-Bakar, S.A.R.; Sheikh, U.U.; Channa, A.; Popescu, N. Cascading Pose Features with CNN-LSTM for Multiview Human Action Recognition. *Signals* **2023**, *4*, 40–55. <https://doi.org/10.3390/signals4010002>

Academic Editors: Yinsheng Chen, Aili Wang and Haibin Wu

Received: 21 July 2022

Revised: 12 November 2022

Accepted: 16 November 2022

Published: 4 January 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Over the past years, automatic human activity recognition (HAR), using computer vision has raised much awareness for researchers in the whole world because it provides accurate and desired outcomes. HAR is a versatile tool in many applications. Among these the significant applications are: Human Computer Interaction (HCI), intelligent video surveillance, ambient assisted living, human-robot interaction, entertainment, and content-based video search. In HCI, when a user performs a task, that task is observed by the activity recognition systems and through feedback the user is guided to complete it. Video surveillance utilises the activity recognition system in order to automatically spot and indicate a suspicious activity to authorities for immediate action. In the same manner, in entertainment field, all the activities of players in the game are perceived by these systems.

Consisting on the complexity and duration, these activities can be arranged into four categories, i.e., gestures, actions, interactions, and group activities [1]. Basic movement of the human body parts that provide some meaning is referred as a gesture. Some good examples of gestures include ‘Head shaking’, ‘hand waving’, and ‘facial expression’. A type of activity performed by a single person is called action. As a matter of fact, when multiple gestures are merged together, such as ‘walking’, ‘running’, ‘jogging’, and ‘punching’, these are described as human actions. Interaction is defined as a two-actor activity. One of these two actors must be a person, and the other can either be a person (human–human interaction) or an object (human–object interaction). Group activities are the most sophisticated activities that combine gestures, movements, and interactions. It requires one or more objects and more than two people.

For human action recognition, the bottom-up strategy is generally followed by the handcrafted representation-based technique. There are three major phases in general i.e., foreground detection, handcrafted feature extraction and representation, and classification [2]. Handcrafted action representation is the traditional approach for action recognition. This method is widely used in the HAR community, and it has shown impressive results on a variety of public datasets [3]. Traditional methods receive input data, extract the features from the data, and use them to train the classifier for classification purposes. Input of the system can be RGB, RGB-D, skeleton data and multi-modal [2]. Feature extraction is used for extracting the useful information from the input data which can form suitable set of data for classification of actions.

HAR system’s effectiveness is determined by its capacity for extracting, modelling, and expressing significant features [4]. It appears that problems with the extraction and representation of features are still being extensively researched in the fields of machine learning and computer vision [5]. The technique of extracting features, which describes patterns that are significant in the recognition process, from arbitrary input information, such as photos, videos, and text, is known as feature extraction [6]. Numerous feature-extraction techniques make use of both the low and high levels approaches to provide the required results. The recognition algorithm further fuses the cues acquired at these levels to provide qualitative results [7]. The consequences of model complexity must be taken into account because HAR is a real-time system. If the user must wait a long time to receive the result, even though the model is flawless, it will not be advantageous for the application [8].

The number of features determines the complexity of the model; for example, the computational complexity would increase as the number of features increases [6]. Currently, the HAR system using skeleton data is extracting the skeleton using OpenPose, for each detected body, a 2D skeleton with 25 joints is extracted. The 2D skeletal characteristics are then converted to RGB images. The two-dimensional skeleton features are encoded in the three R, G, and B channels, and an action sequence produces an RGB image. Finally, collected RGB images are used to train a classifier based on deep learning for the HAR system [9]. The disadvantage of converting skeleton data to RGB image is an increase in computational complexity; if we have skeleton features that we can use directly for HAR system training, why convert to RGB image and extract features again at the model’s input layer? To overcome the process of converting skeleton data into RGB images which are further given to deep learning, this paper proposes deep learning based HAR system which can be trained directly using skeleton features instead of converting it to RGB image and again extract features for training or testing of the system.

## 2. Related Work

Handcrafted action representation is the traditional approach for action recognition. Handcrafted methods are further divided into two major groups i.e., space-time-based approach and appearance-based approach. The space time interest point (STIP) detector, feature descriptor, vocabulary builder, and classifier are the four major components of space-time-based technique [10]. STIP detectors are divided into two types: dense and sparse detectors, whereas feature descriptor is further divided into local and global descriptors. Local descriptors work on colour, posture, and texture while global descriptors use speed variations, illumination changes, and phase changes of video [1]. Space-time volumes (STVs) are the features in the space-time domain which are represented as a 3D spatio-temporal cuboids. For action recognition, STVs measure similarity between volumes. Traditional methods, which involve the computation of optical flow, have some limitations that can be avoided by doing action recognition directly in the space-time volume [11–13]. For the purpose of human action recognition, the space-time features-based techniques extract unique features from space-time volumes or space-time trajectories. These aspects are generally local in their nature and contain distinguishing properties of an action. The characteristics of space-time volumes and trajectories can be categorised as either sparse or dense, depending on the nature of the space-time volumes and trajectories [3]. The feature detectors that are considered to be sparse are those that are based on interest point detectors such as Harris3D [14] and Dollar [15]. On the other hand, the feature detectors that are based on optical flow are considered to be dense. These interest point detectors have served as the foundation for the majority of the newly presented algorithms [3]. Based on the interest points that were found with Harris3D [14], the authors of [16] constructed the feature descriptor and classified the data with PCA (principal component analysis)-SVM. A novel local polynomial space-time descriptor based on optical flow was proposed by the authors in [17] for the purpose of action representation. The spatio-temporal volume of a video sequence is modelled as a three-dimensional object using shape-based approaches. The various events that occur in a video each generate their own unique shapes, and the purpose of these kinds of algorithms is to recognise an event by identifying its shape [18]. The shape of an event can be characterised using a variety of ways, such as shape invariants [11,19], which are employed by shape-based methods. Weinland et al. [20] expand this method to motion-history volumes. When the action of interest is carried out in an environment that allows for reliable segmentation, these strategies function in the most optimal manner. In particular, for static scenes, methods such as background reduction can be used to build high-quality spatio-temporal volumes that are suitable to this analysis.

Deep learning has developed as a prominent machine learning direction, outperforming older approaches in many computer vision applications. Deep learning algorithms have the unique capacity to learn features from raw data, eliminating the requirement for constructed feature detectors and descriptors. Deep learning models are classified into two types: unsupervised/generative models and supervised/discriminative models. Convolution neural network (CNN) is one of the most common supervised learning deep learning methods. The majority of existing learning-based representations either directly apply CNN to video frames or use CNN variations for spatio-temporal characteristics. These models have also performed well on challenging human activity recognition datasets [1]. A convolutional neural network is most typically used in deep learning to analyse visual imagery. CNN research is still ongoing and has great room for advancement. The main enhancements in CNN performance are thought to have occurred between 2015 and now. A CNN's representational capability is typically determined by its depth, and an enhanced feature set ranging from simple to complex abstractions can aid in the learning of complicated issues. The vanishing gradient is the main problem that deep architectures have to deal with. At first, researchers tried to solve this problem by connecting intermediate layers to auxiliary learners [21]. In 2015, making new connections to speed up the convergence rate of deep CNN architectures was the most promising area of study. In this area, different ideas have been put forward, such as information gating mechanisms across

multiple layers, skip connections, and cross-layer channel connectivity [22,23]. Modern deep architectures, such as VGG, ResNet, etc., performed well in several experiments, not only for classification issues but also for challenging recognition and localization issues such as semantic and instance-based object segmentation, scene parsing, scene placement, etc. In the same way, many interesting detection algorithms, such as Feature Pyramid Networks, Cascade R-CNN, Libra R-CNN, etc., changed the architectures mentioned above to make them work better [24]. Combining these networks with a recurrent neural network (RNN) made it possible for deep CNN to be used for image captioning. However, the high cost of computation and the need for a lot of memory are two main problems with deep learning architectures. Because of this, it is very hard to use deep CNN models in environments with few resources. Conventional convolution operations require a very large number of multiplications. This makes the inference time longer and limits CNN to applications with limited memory and time. Many real-world applications, such as self-driving cars, robots, healthcare, and mobile apps, do tasks that need to be done on platforms with limited computing power in a timely way. A number of studies have used deep learning architecture for HAR as illustrated in Table 1.

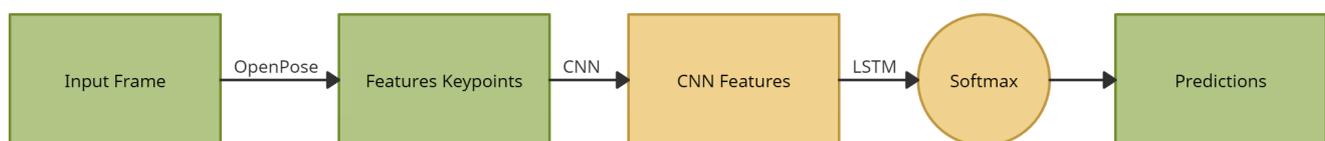
**Table 1.** Review on Deep learning-based techniques for Human Action Recognition.

Authors	Methods	Datasets	Performance
Baccouche et al. [21]	CNN and RNN	KTH	94.39
Ji et al. [22]	3DCNN	KTH	90.02
Grushin et al. [23]	LSTM	KTH	90.7
Sun et al. [24]	Factorised spatio-temporal CNN	HMDB-51	59.1
Simonyan et al. [25]	two stream CNN	HMDB-51	59.4
Wang et al. [26]	CNN	HMDB-51	65.9
Ullah et al. [27]	DB-LSTM	HMDB-51	87.64
Mahasseni et al. [28]	LSTM-CNN	HMDB-51	55.3
Zhang et al. [29]	MV-CNN	UCF101	86.4
Ng et al. [30]	LSTM with 30 frame unroll	UCF101	88.6
Yu et al. [31]	SP-CNN	UCF101	91.6
Fernando et al. [32]	Rank Pooling +CNN	Hollywood2	75.2
Wang et al. [33]	Features (Pose-based)	MSR-action3D	90
Ch et al. [34]	Pose-based CNN	MPII Cooking	71.4
W. Li et al. [35]	Cuboids	MCAD	56.8
M. Faraki et al. [36]	Covariance matrices	MCAD	64.3
W. Li et al. [35]	STIP	MCAD	81.7
H. Wang et al. [37]	IDT	MCAD	84.2
A. Ullah et al. [27]	Conflux LSTM network	MCAD	86.9
Malik et al. [38]	OpenPose + FineKNN	MCAD	86.9

In view of the limited performance of the methods discussed above, this work proposes a simpler strategy on a 2D dataset that can be used to cope with existing real-time systems. To avoid the time-consuming process of converting skeletal joints data to image sequences and then training them using an image-based classifier that extracts the features again, we suggest that the skeletal joints data be used directly for training and testing of the HAR system. Specifically, we propose that a feature vector be used as an input to the CNN-LSTM architecture, which represents a simplification of the work that has been given.

### 3. Methodology

The proposed method as shown in Figure 1 starts with reading frames, extracting skeleton features, preprocessing and finally training CNN-LSTM network. In start the proposed system reads each frame from the input video and extract skeleton features using OpenPose [9]. The OpenPose outputs 25 joints location in the form of X, Y coordinates and confidence map but for the proposed system only X, Y coordinates are used. When we trained our proposed model including confidence maps results were similar, so instead of using X,Y coordinates and confidence map, the proposed research selects only X,Y coordinates for HAR system. Input layer parameters of proposed model are set as  $1 \times 25 \times 2$ , where 1 shows that data for each frame will come and  $25 \times 2$  represents 25 joint-location (X,Y) coordinates. Finally, the X, Y coordinates of extracted skeleton features are fed into the CNN-LSTM system for the classification process. CNN-LSTM architecture as shown in Figure 2 is designed in a way that it can work on skeleton features directly instead of generating heatmaps from skeleton features for the training of deep learning architecture. The proposed CNN-LSTM architecture consists of 12 layers starting with input layer for taking skeleton features as an input. The dimensions of an input layer are  $1 \times 25 \times 2$ , where 1 shows that data are of a single frame and  $25 \times 2$  are the X, Y coordinates of 25 joints locations. A time-distributed CNN layer with 16 filters of size  $3 \times 3$  is used, and for feature extraction, ReLU activation is used on key points of each frame. CNNs are very good at pulling out spatial features that are not affected by scale or rotation. The CNN layer can extract spatial features and angles between the key points in a frame [39]. Batch normalization is used to speed up convergence on the CNN output. The next layer is a dropout layer, which randomly drops some of the weights to avoid overfitting. The CNN output is then flattened and sent to the LSTM layer, which has 20 units and a unit forget bias of 0.5. LSTM is used to see how the features extracted by the CNN layer change over time. This takes advantage of the fact that video input comes in a certain order.



**Figure 1.** Methodology of the proposed skeleton based CNN-LSTM HAR system.

Each frame's output from the LSTM layer is sent to a time-distributed fully connected layer with number of outputs depending on the number of classes and Softmax activation. Each of these eighteen outputs tells you, in terms of cross-entropy, how likely it is that the corresponding action is being performed.

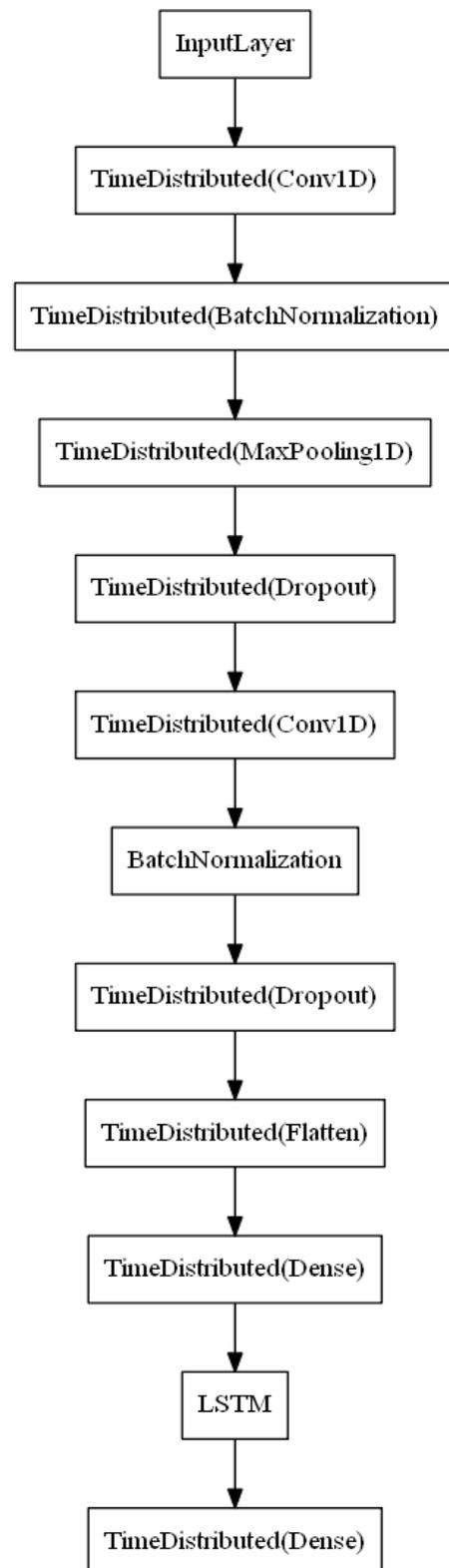


Figure 2. CNN-LSTM Architecture.

#### 4. Experimental Setup

For evaluation purposes, two multiview human action recognition datasets are used MCAD and IXMAS. The MCAD dataset, which is known for its uncontrolled and multi-view motions, was used in this study to show that the proposed technique worked better. MCAD has 18 action categories and 14,298 action examples. Twenty people perform these actions and five cameras record them [19]. The dataset was split into two parts. Dataset was divided into two parts, 80% for training and remaining 20% for testing purpose. There are 18 actions involved in this experiment as mentioned in Table 2. Class one to nine belongs to single person action category, whereas class ten to eighteen belongs to interaction level actions.

**Table 2.** List of actions from MCAD Dataset.

Class	Action
1	Point
2	Wave
3	Jump
4	Crouch
5	Sneeze
6	SitDown
7	StandUp
8	Walk
9	PersonRun
10	CellToEar
11	UseCellPhone
12	DrinkingWater
13	TakePicture
14	ObjectGet
15	ObjectPut
16	ObjectLeft
17	ObjectCarry
18	ObjectThrow

For further evaluation of the proposed system all experiments are performed on IXMAS dataset also. IXMAS has 12 action categories and 1800 action sample performed by 12 actors and five cameras recorded them. There are 12 actions involved in this experiment as mentioned in Table 3. The IXMAS dataset is also divided into 80% for training and 20% for testing.

**Table 3.** List of actions from IXMAS Dataset.

Class	Action
1	check-watch
2	Cross-arms
3	Get-up
4	kick
5	Pick-up
6	Point
7	Punch
8	Scratch-head
9	Sit-Down
10	Turn-Around
11	Walk
12	Wave

## 5. Result and Discussion

This section describes the results achieved by the proposed system. Skeleton detection from 2D image is shown in Figure 3. There are 25 skeleton joints location that are detected by OpenPose [35] and form a  $25 \times 3$  matrix, contains X,Y coordinates and confidence map. As the proposed system only utilises the X,Y coordinates for the action recognition purpose, the input to the proposed system becomes a  $25 \times 2$  matrix. As mentioned in methodology, this input matrix is given to proposed CNN-LSTM architecture, so that CNN can analyse the skeletal joints location and pass it to LSTM network which classify the performed action.

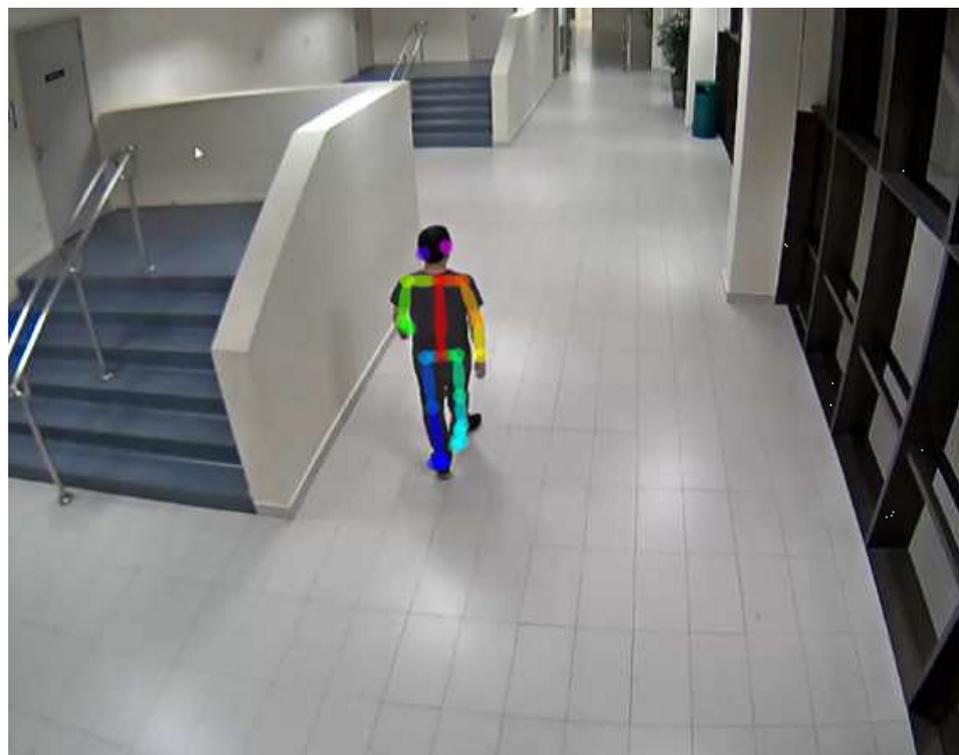
**Figure 3.** Skeleton Detection.

Figure 4 shows the training and validation accuracy of the proposed system. Training accuracy started from 93.3% on first epoch where as the validation accuracy started from 93.2%. As the model training process continues, after 10 epochs training and validation accuracy becomes same around 94.4%. The model was trained on 20 number of epochs, but after 10 epochs training and validation accuracy was constant. Figure 5 shows the loss during training and validation, as the number of epochs increases model loss was decreasing and after 10 number of epochs model loss becomes constant around 0.025.

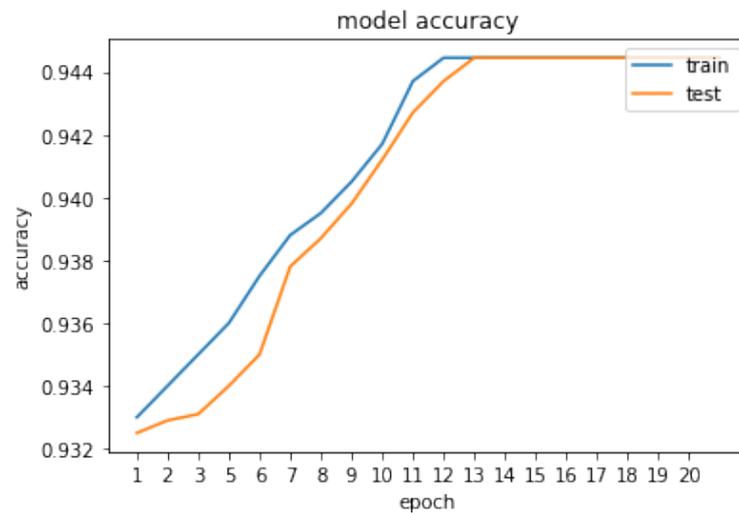


Figure 4. Accuracy of the proposed model during training and testing using MCAD Dataset.

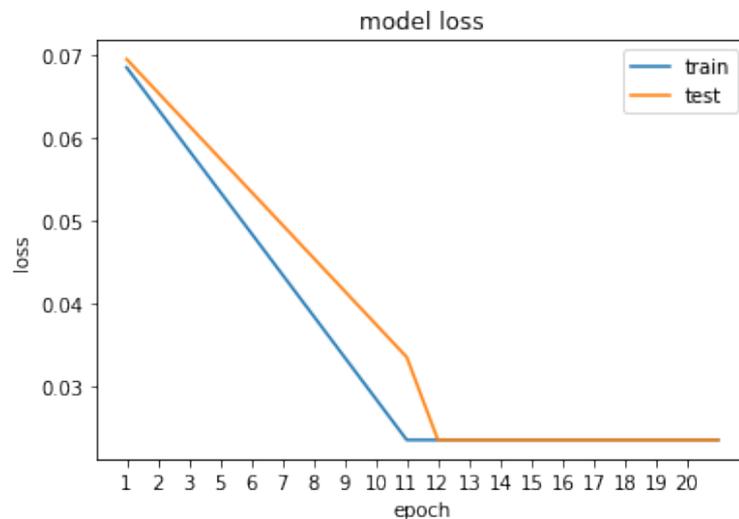


Figure 5. Loss of the proposed model during training and testing using MCAD Dataset.

Figure 6 shows the confusion chart of the proposed system. More statistical results such as Precision, Recall and F1-Score are also added to show the efficacy of the proposed in Table 4. Statistical results also show the performance of model in term of precision, recall and F1-Score, and the proposed model is working well for each class. The proposed model takes 434 s in total during training and approximately 22 s for each epoch.

Table 5 shows the comparative analysis of proposed method with the state of the art work using MCAD dataset. Comparatively our model improved 37.6% accuracy from that achieved previously by Cuboid features [36], 30.1% from the Covariance matrices [37], 12.7% from the STIP features [36], 10.2% from the IDT [40] and 7.5% from the Conflux LSTM network [38]. These results confirmed the superior performance of our method as shown in Table 5.

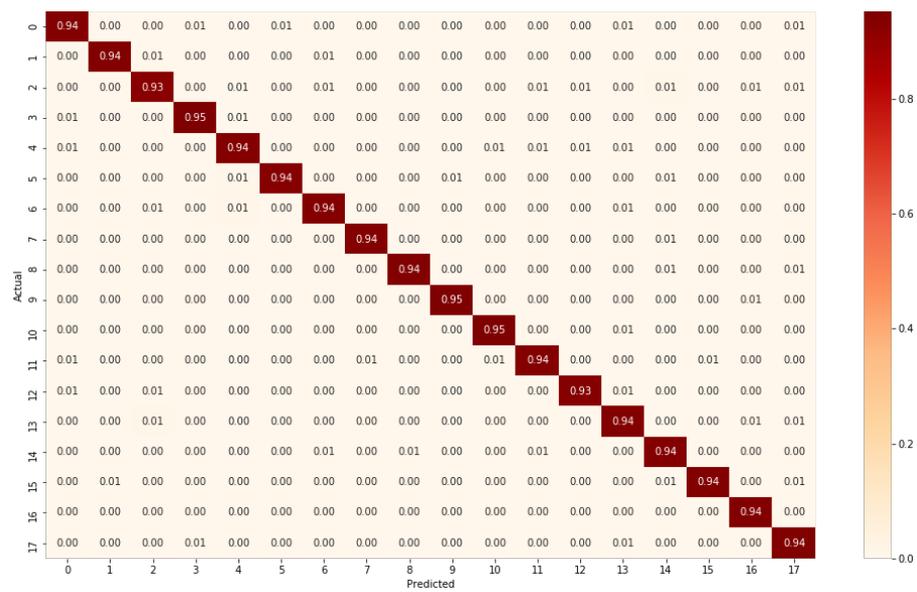


Figure 6. Confusion Matrix using MCAD Dataset.

Table 4. Classification report of proposed deep learning model (Precision, Recall and F1-Score) using MCAD Dataset.

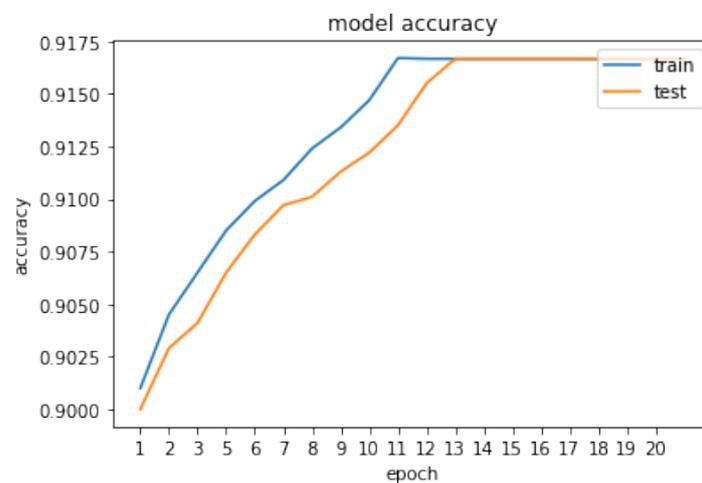
Class	Action	Precision	Recall	F1-Score
1	Point	0.92	0.94	0.93
2	Wave	0.94	0.94	0.94
3	Jump	0.89	0.93	0.91
4	Crouch	0.94	0.95	0.94
5	Sneeze	0.9	0.94	0.92
6	SitDown	0.91	0.94	0.92
7	StandUp	0.9	0.94	0.92
8	Walk	0.97	0.94	0.96
9	PersonRun	0.95	0.94	0.94
10	CellToEar	0.95	0.95	0.95
11	UseCellPhone	0.96	0.95	0.96
12	DrinkingWater	0.95	0.94	0.94
13	TakePicture	0.96	0.93	0.95
14	ObjectGet	0.88	0.94	0.91
15	ObjectPut	0.91	0.94	0.93
16	ObjectLeft	0.94	0.94	0.94
17	ObjectCarry	0.98	0.94	0.96
18	ObjectThrow	0.9	0.94	0.92

For further validation with the state-of-the-art we executed our proposed method on IXMAS dataset using overall accuracy. Figure 7 shows the accuracy during training and testing, whereas Figure 8 shows loss during training and testing. Figure 9 shows the confusion chart of the proposed system using IXMAS dataset. Statistical results such as Precision, Recall and F1-Score are also added to show the efficacy of the proposed in Table 6. Comparatively as shown in Table 7 our model improved 1.92% accuracy from that achieved previously by Shape Features [41], 11.12% from the LBP [42], 8.64% from

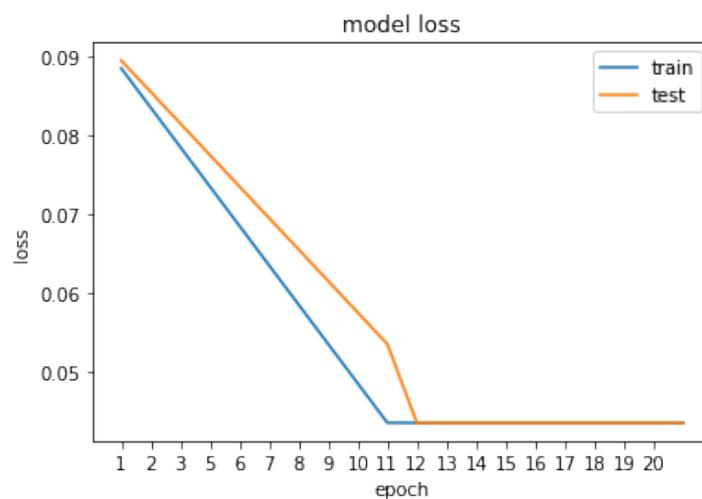
the Motion Features [43], 5.87% from the Shape features [44] and 0.76% from the Shape Features (3D) [45]. These results confirmed the proposed method is suitable for performing multi-view human action recognition.

**Table 5.** Multi-view HAR on MCAD dataset.

Algorithm	Accuracy
Cuboids [35]	56.8
Covariance matrices [36]	64.3
STIP [35]	81.7
IDT [37]	84.2
Conflux LSTM network [27]	86.9
OpenPose+FineKNN [38]	86.9
Proposed method	94.4



**Figure 7.** Accuracy of the proposed model during training and testing using IXMAS Dataset.



**Figure 8.** Loss of the proposed model during training and testing using IXMAS dataset.

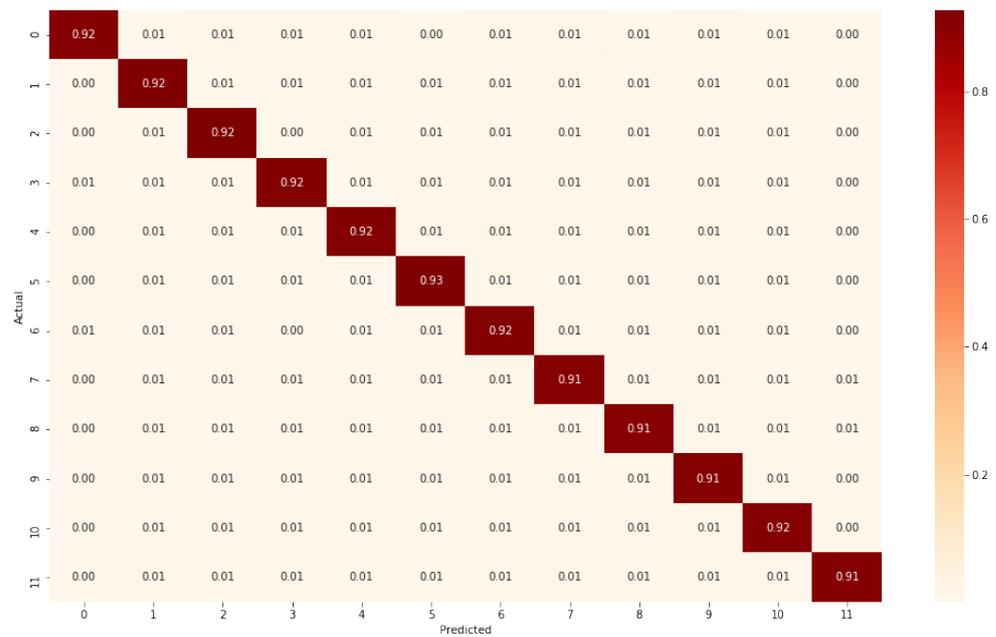


Figure 9. Confusion Matrix using IXMAS Dataset.

Table 6. Classification report of proposed deep learning model(Precision, Recall and F1-Score) using IXMAS dataset.

Class	Action	Precision	Recall	F1-Score
1	Check-watch	0.93	0.92	0.92
2	Cross-arms	0.89	0.92	0.90
3	Get-up	0.90	0.92	0.91
4	Kick	0.91	0.92	0.91
5	Pick-up	0.90	0.94	0.91
6	Point	0.90	0.93	0.91
7	Punch	0.89	0.92	0.91
8	Scratch-head	0.90	0.91	0.91
9	Sit-Down	0.90	0.91	0.91
10	Turn-Around	0.91	0.91	0.91
11	Walk	0.94	0.92	0.93
12	Wave	0.97	0.91	0.94

Table 7. Multi-view HAR on IXMAS dataset.

Algorithm	Accuracy
Shape Features [41]	89.75
LBP [42]	80.55
Motion Features [43]	83.03
Shape features [44]	85.8
Shape Features (3D) [45]	90.91
Proposed method	91.67

We also performed a complexity analysis between our proposed method and other existing methods by leveraging the size of the feature dimension. Table 8 lists the comparison with both the handcrafted and DL methods. The table clearly indicates that by far, the proposed method outperformed other existing methods in terms of the feature dimensions. Our approach uses only 50 features in total which is significantly small compared to other DL approaches.

**Table 8.** Comparison of feature dimension.

Algorithm	Year	Short Descriptions	Data Used	Feature Dimension
HON4D [46]	CVPR 2013	Handcrafted (global descriptor)	Depth	[17,880, 151,200]
HDG [47]	WACV 2014	Handcrafted (local + global descriptor)	Depth+ Skeleton	[1662, 1819]
P-LSTM [48]	CVPR 2016	Deep learning (LSTM)	Skeleton	No. of joints $\times 3 \times 8$
HPM+TM [49]	CVPR 2016	Deep learning (CNN)	Depth	4096
Clips+CNN+ MTLN [50]	CVPR 2017	Deep learning (pre-trained VGG19,MTLN)	Skeleton	7168
RNN [51]	CVPR 2018	Deep learning (RNN)	Skeleton	512
ST-GCN [52]	AAAI 2018	Deep learning (Graph ConvNet)	Skeleton	256
Proposed	2022	OpenPose + CNNLSTM	RGB	$25 \times 2$

## 6. Conclusions

The present research in HAR has been aimed at addressing the issues of complexity and this can be witnessed through the endeavor of the previous works. In this paper, we proposed a method that extracts the 2D skeleton data from the 2D RGB data using the OpenPose technique, and classifies the given action using a proposed deep learning based CNN-LSTM model. As a result, our proposed method significantly decreases the complexity of computation by reducing the feature dimension. In this context, the proposed method was compared with state-of-the-art methods, and the outcomes confirm the potential of our technique. In future, the proposed approach will be used in different applications such as ambient assisted living, patients with neurocognitive disorder i.e., Parkinson's Disease to monitor their daily life activities (DLAs) or predict the falls in elderly. This study is limited by the fact that it uses multi-modality datasets collected in uncontrolled environments where limited number of actions are performed. The same approach can be used on large and challenging datasets for recognizing every person's actions in crowded environments.

**Author Contributions:** Conceptualization, N.u.R.M. and S.A.R.A.-B.; methodology, N.u.R.M.; software, N.u.R.M.; validation, S.A.R.A.-B. and U.U.S.; formal analysis, N.P.; investigation, A.C. and N.u.R.M.; resources, N.u.R.M.; data curation, N.u.R.M. and U.U.S.; writing—original draft preparation, N.u.R.M.; writing—review and editing, A.C.; visualization, N.P.; supervision, S.A.R.A.-B. and U.U.S.; project administration, S.A.R.A.-B.; funding acquisition, A.C. and N.P. All authors have read and agreed to the published version of the manuscript

**Funding:** This research was funded by the European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska Curie grant agreement No. 813278 (A-WEAR: A network for dynamic wearable applications with privacy constraints).

**Acknowledgments:** The authors would like to thank the Ministry of Higher Education Malaysia (KPT) and Universiti Teknologi Malaysia (UTM) for their support under the UTM Fundamental Research Grant (UTMFR), grant number Q.J130000.3823.22H29. The authors also gratefully acknowledge funding from European Union's Horizon 2020 Research and Innovation program under the Marie Skłodowska Curie grant agreement No. 813278 (A-WEAR: A network for dynamic wearable applications with privacy constraints, <http://www.a-wear.eu/> (accessed on 15 July 2022)) and a grant from the Romanian National Authority for Scientific Research and Innovation, UEFISCDI project PN-III-P3-3.6-H2020-2020-0124.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

HAR	Human Action Recognition
HCI	Human Computer Interaction
STIP	Space Time Interest Point
BOW	bag-of-words
STVs	Space Time Volumes
PCA	Principal Component Analysis
CNN	Convolution neural network
LSTM	Long Short Term Memory
RNN	Recurrent Neural Network

## References

1. Aggarwal, J.; Ryoo, M. Human activity analysis: A review. *Acm Comput. Surv. Csur* **2011**, *43*, 1–43. [CrossRef]
2. Abu-Bakar, S. Advances in human action recognition: An updated survey. *IET Image Process.* **2019**, *13*, 2381–2394. [CrossRef]
3. Sargano, A.; Angelov, P.; Habib, Z. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Appl. Sci.* **2017**, *7*, 110. [CrossRef]
4. Wu, D.; Zheng, S.; Zhang, X.; Yuan, C.; Cheng, F.; Zhao, Y.; Lin, Y.; Zhao, Z.; Jiang, Y.; Huang, D. Deep Learning-Based Methods for Person Re-identification: A Comprehensive Review. *Neurocomputing* **2019**, *337*, 354–371. Available online: <https://www.sciencedirect.com/science/article/pii/S0925231219301225> (accessed on 22 May 2022). [CrossRef]
5. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
6. Wang, L.; Xu, Y.; Cheng, J.; Xia, H.; Yin, J.; Wu, J. Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks. *IEEE Access* **2018**, *6*, 17913–17922. [CrossRef]
7. Liu, Y.; Gevers, T.; Li, X. Color Constancy by Combining Low-Mid-High Level Image Cues. *Comput. Vis. Image Underst.* **2015**, *140*, 1–8. Available online: <https://www.sciencedirect.com/science/article/pii/S1077314215001241> (accessed on 22 April 2022).
8. Angerbauer, S.; Palmanshofer, A.; Selinger, S.; Kurz, M. Comparing Human Activity Recognition Models Based on Complexity and Resource Usage. *Appl. Sci.* **2021**, *11*, 8473. [CrossRef]
9. Cao, Z.; Simon, T.; Wei, S.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
10. Das Dawn, D.; Shaikh, S. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. *Vis. Comput.* **2016**, *32*, 289–306. [CrossRef]
11. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. In Proceedings of the Tenth IEEE International Conference On Computer Vision (ICCV'05) Volume 1, Beijing, China, 17–21 October 2005; Volume 2, pp. 1395–1402.
12. Black, M. Explaining optical flow events with parameterized spatio-temporal models. In Proceedings of the 1999 IEEE Computer Society Conference On Computer Vision And Pattern Recognition (Cat. No PR00149), Fort Collins, CO, USA, 23–25 June 1999; Volume 1, pp. 326–332.
13. Efros, Berg; Mori; Malik. Recognizing action at a distance. In Proceedings of the Ninth IEEE International Conference On Computer Vision, Nice, France, 13–16 October 2003; Volume 2, pp. 726–733.
14. Laptev, I. On space-time interest points. *Int. J. Comput. Vis.* **2005**, *64*, 107–123.
15. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior recognition via sparse spatio-temporal features. In Proceedings of the 2005 IEEE International Workshop On Visual Surveillance And Performance Evaluation Of Tracking And Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.
16. Thi, T.; Zhang, J.; Cheng, L.; Wang, L.; Satoh, S. Human action recognition and localization in video using structured learning of local space-time features. In Proceedings of the 7th IEEE International Conference on Advanced Video And Signal Based Surveillance, Boston, MA, USA, 29 August–1 September 2010; pp. 204–211.
17. Kihl, O.; Picard, D.; Gosselin, P.H. Local polynomial space-time descriptors for action classification. *Mach. Vis. Appl.* **2016**, *27*, 351–361. [CrossRef]
18. Ke, Y.; Sukthankar, R.; Hebert, M. Event detection in crowded videos. In Proceedings of the 2007 IEEE 11th International Conference On Computer Vision, Rio de Janeiro, Brazil, 14–21 October 2007; pp. 1–8.
19. Yilmaz, A.; Shah, M. Actions as objects: A novel action representation. In Proceedings of the CVPR-2005, San Diego, CA, USA, 20–26 June 2005.
20. Weinl, D.; Ronfard, R.; Boyer, E. Free viewpoint action recognition using motion history volumes. *Comput. Vis. Image Underst.* **2006**, *104*, 249–257.
21. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In *International Workshop On Human Behavior Understanding*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 29–39.
22. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]

23. Grushin, A.; Monner, D.; Reggia, J.; Mishra, A. Robust human action recognition via long short-term memory. In Proceedings of the the 2013 International Joint Conference On Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 1–8.
24. Sun, L.; Jia, K.; Yeung, D.; Shi, B. Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE International Conference On Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4597–4605.
25. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *arXiv* **2014**. [[CrossRef](#)]
26. Wang, L.; Qiao, Y.; Tang, X. Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; Volume 6.
27. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S. Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE Access* **2017**, *6*, 1155–1166. [[CrossRef](#)]
28. Mahasseni, B.; Todorovic, S. Regularizing long short term memory with 3D human-skeleton sequences for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3054–3062.
29. Zhang, B.; Wang, L.; Wang, Z.; Qiao, Y.; Wang, H. Real-time action recognition with deeply transferred motion vector cnns. *IEEE Trans. Image Process.* **2018**, *27*, 2326–2339. [[CrossRef](#)]
30. Yue-Hei, Ng, J.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
31. Yu, S.; Cheng, Y.; Su, S.; Cai, G.; Li, S. Stratified pooling based deep convolutional neural networks for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 13367–13382. [[CrossRef](#)]
32. Fern, o B.; Gavves, E.; Oramas, J.; Ghodrati, A.; Tuytelaars, T. Rank pooling for action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 773–787.
33. Wang, C.; Wang, Y.; Yuille, A. An approach to pose-based action recognition. In Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 915–922.
34. Chéron, G.; Laptev, I.; Schmid, C. P-cnn: Pose-based cnn features for action recognition. In Proceedings of the IEEE International Conference On Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3218–3226.
35. Li, W.; Wong, Y.; Liu, A.; Li, Y.; Su, Y.; Kankanhalli, M. Multi-camera action dataset for cross-camera action recognition benchmarking. In Proceedings of the IEEE Winter Conference On Applications Of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 187–196.
36. Faraki, M.; Palhang, M.; Sanderson, C. Log-Euclidean bag of words for human action recognition. *IET Comput. Vis.* **2015**, *9*, 331–339. [[CrossRef](#)]
37. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference On Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558.
38. Malik, N.; Abu Bakar, S.; Sheikh, U. Multiview Human Action Recognition System Based on OpenPose and KNN Classifier. In *Proceedings of the 11th International Conference On Robotics, Vision, Signal Processing And Power Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 890–895.
39. Yadav, S.K.; Singh, A.; Gupta, A.; Raheja, J.L. Real-time Yoga recognition using deep learning. *Neural Comput. Applic* **2019**, *31*, 9349–9361.
40. Ullah, A.; Muhammad, K.; Hussain, T.; Baik, S. Conflux LSTMs network: A novel approach for multi-view action recognition. *Neurocomputing* **2021**, *435*, 321–329. [[CrossRef](#)]
41. Sargano, A.B.; Angelov, P.; Habib, Z. Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines. *Appl. Sci.* **2016**, *6*, 309. [[CrossRef](#)]
42. Baumann, F.; Ehlers, A.; Rosenhahn, B.; Liao, J. Recognizing human actions using novel space-time volume binary patterns. *Neurocomputing* **2016**, *173*, 54–63. [[CrossRef](#)]
43. Chun, S.Y.I.; Lee, C.-S. Human Action Recognition Using Histogram of Motion Intensity and Direction from Multi View. *IET Comput. Vis.* **2016**, *10*. [[CrossRef](#)]
44. Vishwakarma, D.K.; Kapoor, R.; Dhiman, A. A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics. *Robot. Auton. Syst.* **2016**, *77*, 25–38. [[CrossRef](#)]
45. Pehlivan, S.; Duygulu, P. A new pose-based representation for recognizing actions from multiple cameras. *Comput. Vis. Image Underst.* **2011**, *115*, 140–151. [[CrossRef](#)]
46. Oreifej, O.; Liu, Z. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013. [[CrossRef](#)]
47. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A. Real time action recognition using histograms of depth gradients and random decision forests. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, 24–26 March 2014.
48. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
49. Rahmani, H.; Mian, A. 3D Action Recognition from Novel Viewpoints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

50. Ke, Q.; Bennamoun, M.; An S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
51. Li, S.; Li, W.; Cook, C.; Zhu, C.; Gao, Y. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
52. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.