Deep Learning-Based Human Action Recognition with Key-Frames Sampling Using Ranking Methods

Nusrat Tasnim and Joong-Hwan Baek *

School of Electronics and Information Engineering, Korea Aerospace University, Goyang 10540, Korea; tasnim.nishu70@kau.kr
* Correspondence: jhbaek@kau.ac.kr; Tel.: +82-2-300-0125

Abstract: Nowadays, the demand for human–machine or object interaction is growing tremendously owing to its diverse applications. The massive advancement in modern technology has greatly influenced researchers to adopt deep learning models in the fields of computer vision and image-processing, particularly human action recognition. Many methods have been developed to recognize human activity, which is limited to effectiveness, efficiency, and use of data modalities. Very few methods have used depth sequences in which they have introduced different encoding techniques to represent an action sequence into the spatial format called dynamic image. Then, they have used a 2D convolutional neural network (CNN) or traditional machine learning algorithms for action recognition. These methods are completely dependent on the effectiveness of the spatial representation. In this article, we propose a novel ranking-based approach to select key frames and adopt a 3D-CNN model for action classification. We directly use the raw sequence instead of generating the dynamic image. We investigate the recognition results with various levels of sampling to show the competency and robustness of the proposed system. We also examine the universality of the proposed method on three benchmark human action datasets: DHA (depth-included human action), MSR-Action3D (Microsoft Action 3D), and UTD-MHAD (University of Texas at Dallas Multimodal Human Action Dataset). The proposed method secures better performance than state-of-the-art techniques using depth sequences.

Keywords: human–machine or object interaction; human action recognition; deep learning; key frames sampling; ranking method

1. Introduction

The rapid development of electronic devices such as smartphones, televisions, notebooks, and personal computers plays an important role in our daily life. The ways of interacting with these devices have also improved dramatically over the past years. To provide easy, smart, and comfortable ways of communication, several devices and applications have been invented ranging from wired keyboards to wireless vision-based communication. Recently, vision-based communication has become very popular owing to time management, cost-effectiveness, and the pandemic. Several real-world vision-based applications have already been introduced, such as human–machine interaction [1], video surveillance systems [2], data retrieval [3], augmented reality [4,5], virtual reality [5], medical care [6], autonomous driving systems [7], and gaming control [8]. Many methods have also been developed for pose estimation [9], facial expression recognition [10], behavior analysis [11], hand gesture recognition [12–14], and action recognition [15–48] to make vision-based decision or communication.

Hand gesture recognition [12–14] is one of the most popular forms of vision-based interaction. It is limited to the classification of actions that are performed by only hands. Thus, it is necessary to develop a system that can understand the actions accomplished by different parts of the human body. Human action recognition focuses on the discrimination
of human actions performed by hands (‘Clap’, ‘Catch’, and ‘SwipeLeft’), leg (‘Kick’, ‘LegCurl’),
or whole-body (‘Jogging’, ‘Run’, ‘Walk’). The advancement in modern data capturing devices
allows researchers to capture multi-modal human action datasets such as inertial, skeleton,
RGB, and depth [15].

The inertial dataset [16] has not gained much attention among researchers because
the user needs to wear multiple inertial sensors on their body to collect inertial data.
Sometimes, it becomes very uneasy and uncomfortable, such as when driving cars and
for robotic operations in medical care systems. The inertial sensor is also sensitive to
sensor location and angle on the body. Compared to the inertial dataset, it is easy and
comfortable to capture skeleton information without intimate contact with sensors. The
skeleton dataset [17–20] contains only a limited amount of joint information about the
human body, which is very little information for perfectly classifying an action. The
shortage of information in the skeleton dataset sometimes degrades the performance
of recognition systems. However, the RGB dataset [21–23] covers full information of the
human body, but it is time-consuming, computationally expensive, and complex in practical
usage. These data are also sensitive to illumination changes, camera calibration, lighting
conditions, and background. By studying the drawbacks of the inertial, skeleton, and
RGB data modalities, we use depth sequences for the proposed method that includes
enough information including the whole body. The depth dataset is also not sensitive to
illumination changes, lighting conditions, and background.

In this article, we propose a novel approach for human action recognition with key-
frames sampling. The key-frames are sampled using ranking metrics. By considering the
objectives of the proposed method, we introduce two well-known similarity comparison
metrics between images, namely, structural similarity index measure (SSIM) and corre-
lation coefficient measure (CCM). The major contributions of the proposed method are
summarized as follows:

1. We introduce a novel ranking-based approach for human action recognition using
3D-CNN with raw depth sequence.
2. First, we use SSIM or CCM ranking metrics to select $k$-ranked frames that contain
more spatial and temporal changes. This allows us to discard the redundant frames
having the same or very similar information.
3. Then, we use transfer learning to perform the recognition of human action. It helps
maintain the knowledge of previous hand gesture datasets and applies them to human
action datasets.
4. We also adopt three different publicly available benchmark human action recognition
datasets to emphasize the robustness of the proposed method.

The remaining parts of this article are organized as follows. We illustrate the re-
lated works in Section 2 by summarizing the key ideas. The detailed methodology of
the proposed method is described in Section 3. Section 4 provides experimental setups,
performance evaluation, and comparisons. We discuss the drawbacks, benefits, and future
works of the prior and proposed methods in Section 5. Finally, we summarize the proposed
method in Section 6.

2. Related Works

The drawbacks of the inertial, skeleton, and RGB sequences lead to the adoption of
depth datasets to effectively recognize human actions. Depth sequences are introduced in
the form of one-channel data, known as grayscale. Many researchers have converted
the depth sequences into several other formats such as skeleton joints [24–27], point
clouds [28–30], and spatial–temporal images [31–44] and then extracted discriminative
features for classification. Figure 1 shows three examples of depth sequence to the skeleton,
point cloud, and dynamic images representation, respectively. Very few methods have
conducted depth sequences directly to discriminate human activities [45–48].
The skeleton data represent the 3D coordinate values of joints in the human body. Very few methods have been published in which skeleton joint information is extracted from depth videos for action classification. Yang et al. [24] extracted joint information from depth sequences and computed the Eigen-joints feature, which combined static posture, motion, and offset of action. They conducted a naïve Bayes nearest neighbor (NB-NN) classifier for recognizing actions. Xia et al. [25] also estimated the joint information from depth sequences and computed the Eigen-joints feature, which combined static posture, motion, and offset of action. They conducted a naïve Bayes nearest neighbor (NB-NN) classifier for recognizing actions.

Li et al. [28] used an action graph to model the dynamics of actions and collected a bag of 3D points that corresponded to the nodes of the action graph for action discrimination. Rahmani et al. [29] proposed a point cloud-based human action recognition system in which they introduced a descriptor and a key point detection algorithm. The descriptor was calculated based on a histogram of oriented principal components. In [30], Li et al. extracted human behavioral features called historical point cloud track features, including depth and skeleton information for action recognition.

Megavannan et al. [31] calculated motion information from depth difference and average depth images by dividing the silhouette bounding box hierarchically and encoded them into motion history images (MHI). They used translation, scale, and orientation invariant Hu moments to encode the MHIs. Xia et al. [32] applied a filtering method for determining the spatio-temporal interest points from the depth (DSTIP) sequence and used the depth cuboid similarity feature (DCSF) to describe the local depth cuboid around DSTIP.

Eum et al. [33] designed a feature map called depth histogram of oriented gradient (MHI-
HOG) and modeled actions through k-means clustering. They evaluated the proposed method using HMM model. Bulbul et al. [34] described a new feature descriptor based on depth motion map (DMM), contourlet transform (CT), and HOG. They applied the CT on DMM and computed HOG features for classification. Liu et al. [35] extracted salient depth map (SDM) and binary shape map (BSM) features from depth sequences and formed a bag of map words for action classification. Bulbul et al. [36] encoded the depth sequences into three DMMs along the front, side, and top projections. From the three DMMs, they computed contourlet-based HOG (CL-HOG) and local binary pattern (LBP) features. They investigated the performance using majority voting (MV) and logarithmic opinion pool (LOGP) fusion strategies. Likewise, Chen et al. [37] projected the depth maps into three orthogonal Cartesian planes and generated three DMMs along the front, side, and top views. They used an l2-regularized classifier with a distance-weighted Tikhonov matrix for action discrimination. Jin et al. [38] partitioned a whole depth sequence into sub-sequences with uniform length, called vague division (VD), and projected them onto three orthogonal planes. They also computed DMMs by determining the difference between adjacent frames of the projected views. Azad et al. [39] sampled the entire depth sequences based on the motion energy of key-frames and produced a weighted depth motion map (WDMM). From the WDMM, they extracted HOG and LBP features to perform the classification. Li et al. [40] extended the DMMs-based human action recognition in which they extracted local ternary pattern features to filter the DMMs and conducted a CNN model for classification. Liang et al. [41] segmented a depth sequence to sub-actions called energy-guided sub-actions (EGSA) and time-guided sub-action (TGS) to explore the sub-actions relationship for action recognition. Li et al. [42] used a multi-scale feature map called Laplacian pyramid depth motion images (LP-DMI) obtained from DMI. Finally, they conducted an extreme learning machine (ELM) for action recognition using the features extracted with HOG and visual geometry group (VGG) descriptors. Similarly, Bulbul et al. [43] split a depth sequence into two different sizes of sub-sequences and computed several DMMs. From the obtained DMMs, they extracted auto-correlation of gradient features (ACG) and classified them using an l2-regularized collaborative representation classifier (CRC). Pareek et al. [44] proposed a new algorithm called self-adaptive differential evolution with a knowledge-based control parameter-extreme learning machine. They calculated the LBP features from DMM and evaluated the recognition results using CRC, probability CRC, and kernel ELM with LBP features.

The performance of the above methods is vastly dependent on the effective representation of depth sequence into the skeleton, point cloud, and dynamic image. When the number of frames and position of the camera or human varies, the spatial and the temporal information of the same action can differ and degrade the overall performance. Moreover, it is very cumbersome and time-consuming to convert the depth sequence into the skeleton, point cloud, and dynamic image and extract discriminative features from them. Therefore, some methods have used raw depth sequence to recognize human action. Wang et al. [45] designed a view correlation discovery network (VCDN) to generate high-level information by fusing multi-modality information. Caballero et al. [46] proposed a 3D-CNN model to automatically extract spatial and temporal information from the raw depth sequence. In [47], Liu et al. introduced the view-correlation adaptation (VCA) technique to handle multi-modality datasets for action recognition. They integrated a view encoder which extracted the view representation features from the depth and RGB sequences. They evaluated the proposed method with different features representation, such as VCA-Entropy and semi-supervised feature augmentation (SeMix). Bai et al. [48] proposed a new method for human action recognition with raw depth sequences. They suggested a collaborative attention mechanism (CAM) to solve the multi-view problems. To obtain the multi-view collaborative process, they conducted an extended LSTM model.

As described earlier [24–44], the encoding of depth sequences into other formats such as skeleton, point cloud, and dynamic required additional effort and time. As a result, it is a good idea to use the raw data directly to recognize the action. The methods
suggested in [12–14] considered raw depth sequences to hand gestures recognition. They used whole sequences to detect and classify hand gestures based on a fixed number of frames. The methods [46,48] also used raw depth sequences for activity recognition using 3D-FCNN and LSTM. Even though the methods in [45–48] performed action classification with raw depth sequences, they achieved very low performance. Therefore, we introduce a new method to select \( k \)-ranked frames which contain the almost full spatial and temporal information of action. The proposed method achieved higher recognition performance over the state-of-the-art methods. The detail of the proposed method is illustrated in the next section.

3. Proposed Methodology

The explanation of the proposed method including motivation, key-frames sampling, and action classification is demonstrated in this section. First, we provide the inspiration for the research on human action recognition. Then, we illustrate the main concepts by providing the overall architectural workflow of the proposed method.

3.1. Motivation

The demand for easy, smart, and effective methods of interacting with electronic devices has reached its peak owing to the massive improvements in modern technology. Vision-based methods, particularly human action recognition, have become very popular nowadays. Most of the methods described in Section 2 encoded the depth sequences into dynamic images to capture spatial and temporal changes in action. However, there can be several neighboring frames that are very similar in terms of spatial and motion information. These frames have less significance in the overall structure and texture information of action. Thus, it is important to define a method that can effectively sample more significant frames by keeping the spatial and temporal information unaltered. At the same time, most of the papers used hand-crafted features to recognize the actions by using traditional machine learning techniques such as NB-NN, MV, CRC, and ELM. By considering the above drawbacks, we suggest a novel approach to sample key-frames using the ranking method. In addition, we directly use the sampled frames to recognize the action using the 3D-CNN model.

3.2. Architecture of the Proposed Method

The overall architecture of the proposed system includes two modules, as shown in Figure 2. They are: (1) key-frames sampling using the ranking method and (2) recognition of action using the 3D-CNN model. Initially, the rank values (\( \psi_1, \psi_2, \ldots, \psi_{n-1} \)) are calculated between the neighboring frames from the raw depth sequence having \( n \) frames using rank metrics (SSIM and CCM). The rank values are then sorted in ascending order and selected \( k \)-ranked frames from \( n \) raw frames. Finally, the sampled frames are sent to a 3D-CNN model to produce class labels as outputs.

3.2.1. Key-Frames Sampling Using Ranking Methods

Frame sampling is a procedure in which a fixed number of frames are chosen from a sequence of frames based on a criterion. In this research, we define the frame sampling criterion as the rank metric. The reason behind the fixed number of frames sampling is that there are several redundant frames in an action sequence. The redundant frames in most cases, the adjacent frames contain similar types of information in terms of spatial and temporal changes, as shown in Figure 3. Thus, it is very important to discard the unnecessary frames which have the same motion information. Sometimes, it is crucial to determine the best set of frames having better motion information to effectively recognize actions. We use the rank metric to reorganize the frames based on the rank values and then sample \( k \)-ranked frames as the significant set (\( k = 16, 20, 24 \)). The rank metric can be different forms based on the intended works. As the frame sampling in an action recognition task, we choose SSIM and CCM metrics.
Figure 2. \( k \)-ranked key-frames sampling using ranking methods (the background of each depth frames was removed).

Figure 3. 16-ranked key-frames selection using SSIM ranking method. The SSIM metric considers the structural properties between two adjacent frames to perform the comparisons. Let \( I_t \) and \( I_{t+1} \) be the two adjacent frames at times \( t \) and \( t+1 \) with height and width \( H \) and \( W \), respectively, in an action sequence. Then, the SSIM values \( (\psi) \) between \( I_t \) and \( I_{t+1} \) can be defined as follows:

\[
\psi(I_t, I_{t+1}) = \frac{(2 \times I_t \times I_{t+1} + C_1)(2 \times \text{var}(I_t, I_{t+1}) + C_2)}{(2 \times I_t^2 + I_{t+1}^2 + C_1)(\text{var}(I_t)^2 + \text{var}(I_{t+1})^2 + C_2)}
\] (1)
where \( \text{var}(I_t, I_{t+1}) \) indicates the inter-frame variance; \( \text{var}(I_t) \) and \( \text{var}(I_{t+1}) \) indicate the intra-frame variances. \( c_1 \) and \( c_2 \) are constants.

Figure 3 illustrates how the 16-ranked frames sampling works. We have a total of 23 frames in the 'Bend' action of the DHA dataset. We compute the SSIM value between the current frame and its immediate next frame. As given in Figure 3, we can say that the neighboring frames contain almost the same information. Since the larger SSIM values indicate more similarity, we sort the SSIM values in ascending order and sample the lowest 16-frames for the experiments. From Figure 3, we can assume that the sampled 16-frames also bear the same spatial and temporal information as the original sequences, with a little distortion. Similarly, we also perform the experiments and analysis with \( k = 20 \) and \( 24 \) for showing the generality of the proposed ranking method.

To show the robustness of the proposed system, we also conduct the CCM metric to sample \( k \)-ranked frames. The CCM values (\( \rho \)) between \( I_t \) and \( I_{t+1} \) can be calculated by using Equation (2) as follows:

\[
\rho(I_t, I_{t+1}) = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} (I_t(i,j) - \bar{I}_t)(I_{t+1}(i,j) - \bar{I}_{t+1})}{\sqrt{(I_t(i,j) - \bar{I}_t)^2(I_{t+1}(i,j) - \bar{I}_{t+1})^2}}
\]

(2)

where \( \bar{I} \) indicates mean value of \( I \). The overall procedure is the same as the sampling of frames based on SSIM values.

For sequences with the number of frames less than \( k \), we duplicate every nth frame for making them the same length as \( k \). The overall scenario for balancing the samples in the short sequence is depicted in Figure 4. We use the 'kick' action from the DHA dataset having 13 frames for the illustration of the balancing procedure. Since we need \( k \)-frames (\( k = 16 \)) for training the deep learning, we duplicate the 4th, 8th, and 12th frames for balancing and insert them next to the neighboring frames.

![Figure 4. The procedure of balancing a sequence into 16 frames from 13 frames.](image)

3.2.2. Deep Learning for Human Action Classification

Over the past years, we have observed significant improvement in the field of machine learning, especially deep learning [49–51]. Nowadays, deep learning is being widely applied in every field, including detection, recognition, classification, and segmentation. Owing to the challenges of hand-crafted features in human action datasets, deep learning has also become very successful in human action recognition. However, there are very few methods that directly deal with raw human action sequences to recognize action classes. Even though some methods have been directly applied to raw sequences using RNN and LSTM, the performance was comparably lower. Most of the methods have encoded the
action sequences into spatial formats and then integrated a machine learning or deep learning model to discriminate the actions.

Instead of representing a whole sequence in a single image for action, we consider sampling $k$-ranked frames as the input to the 3D-CNN model. The 3D-CNN can directly extract both spatial and temporal features from raw input frames without encoding them into another domain. We investigate the recognition performance using different deep learning models such as residual network (ResNet) \cite{12}, convolution 3D (C3D) \cite{12}, I3D \cite{52}, R2P1D \cite{53}, X3D \cite{53}, and 3D-FCNN \cite{46}. Figure 5 depicts the average CCM and SSIM results for DHA, MSR-Action3D, and UTD-MHAD datasets. We experimentally find that a 3D ResNet with 101 layers called ResNet101 works well for the proposed method. As a result, we use ResNet101 for the backbone deep learning model for the whole experiment. A large number of convolutional layers in ResNet101 can learn high-level features and useful functions to obtain the hierarchical representation of action information. The average pooling layer helps extract discriminative features containing spatial and temporal directions.

![Comparison of recognition results for different 3D-CNN models on the UTD-MHAD data using SSIM metric having $k = 24$.](image1)

The ResNet101 is one of the most effective 3D-CNN models in which residual information of the previous layer is again connected to the current layer. It extracts deeper features as we consider 101 layers. A residual block in ResNet101 consists of convolution, batch normalization, rectified linear unit, convolution, and batch normalization operations, as shown in Figure 6. The input, $I$, is passed through a residual function, $F(I)$, and generates features, $I'$, by combining the output from $F(I)$ with the input.

![A residual block in ResNet.](image2)
As we have trained the ResNet101 model for the multi-class classification, we have used the cross-entropy loss function $\text{Loss}_{CE}$, defined as follows:

$$\text{Loss}_{CE} = \sum_{i=1}^{C} y_i \log(p_i)$$

(3)

where $C$ is the number of classes and $y_i$ and $p_i$ indicate original and predicted labels, respectively.

4. Experimental Results

This section provides the experimental results by illustrating the environmental settings, performance evaluation, performance comparisons, and complexity analysis.

4.1. Datasets

To show the effectiveness and establish the robustness of the proposed system, we study three publicly available benchmark depth datasets. They are DHA [54], MSR-Action3D [28], and UTD-MHAD [55] datasets.

4.1.1. DHA Dataset

The DHA dataset is an extended version of the Weizmann dataset. It was first introduced by the computer vision exchange lab. They combined 10 classes of actions from the Weizmann dataset along with their own 13 classes of actions dataset to make a total of 23 classes DHA dataset. Each class has 21 sequences which are performed by 21 subjects (12 males and 9 females). The names of the actions are: ‘ArmCurl’, ‘ArmSwing’, ‘Bend’, ‘FrontBox’, ‘FrontClap’, ‘GolfSwing’, ‘Jack’, ‘Jump’, ‘Kick’, ‘LegCurl’, ‘LegKick’, ‘OneHandWave’, ‘Pitch’, ‘Pjump’, ‘RodSwing’, ‘Run’, ‘Side’, ‘SideBox’, ‘SideClip’, ‘Skip’, ‘TaiChi’, ‘TwoHandWave’, ‘Walk’. This dataset contains a total of 483 sequences. Figure 7a shows an example of ‘OneHandWave’ action frames in the DHA dataset.

4.1.2. MSR-Action3D Dataset

The MSR-Action3D dataset had been devised by Wanqing Li and Communication and Collaboration Systems Group at Microsoft Research Red-Mond. A total of 10 subjects repeatedly performed 20 different actions to generate a total of 567 sequences. The action classes are defined as follows: ‘Band’, ‘DrawCircle’, ‘DrawCross’, ‘DrawTick’, ‘ForwardKick’, ‘ForwardPunch’, ‘GolfSwing’, ‘Hammer’, ‘HandCatch’, ‘HandClap’, ‘HighArmWave’, ‘HighThrow’ ‘HorizontalArmWave’, ‘Jogging’, ‘PickUpandThrow’, ‘SideBoxing’, ‘SideKick’, ‘TennisServe’, ‘TennisSwing’, ‘TwoHandWave’. Figure 7b shows an example of ‘TwoHandWave’ action frames in the MSR-Action3D dataset.

4.1.3. UTD-MHAD Dataset

This section provides the experimental results by illustrating the environmental settings, performance evaluation, performance comparisons, and complexity analysis.

4.1. Datasets

To show the effectiveness and establish the robustness of the proposed system, we study three publicly available benchmark depth datasets. They are DHA [54], MSR-Action3D [28], and UTD-MHAD [55] datasets.

4.1.1. DHA Dataset

The DHA dataset is an extended version of the Weizmann dataset. It was first introduced by the computer vision exchange lab. They combined 10 classes of actions from the Weizmann dataset along with their own 13 classes of actions dataset to make a total of 23 classes DHA dataset. Each class has 21 sequences which are performed by 21 subjects (12 males and 9 females). The names of the actions are: ‘ArmCurl’, ‘ArmSwing’, ‘Bend’, ‘FrontBox’, ‘FrontClap’, ‘GolfSwing’, ‘Jack’, ‘Jump’, ‘Kick’, ‘LegCurl’, ‘LegKick’, ‘OneHandWave’, ‘Pitch’, ‘Pjump’, ‘RodSwing’, ‘Run’, ‘Side’, ‘SideBox’, ‘SideClip’, ‘Skip’, ‘TaiChi’, ‘TwoHandWave’, ‘Walk’. This dataset contains a total of 483 sequences. Figure 7a shows an example of ‘OneHandWave’ action frames in the DHA dataset.

4.1.2. MSR-Action3D Dataset

The MSR-Action3D dataset had been devised by Wanqing Li and Communication and Collaboration Systems Group at Microsoft Research Red-Mond. A total of 10 subjects repeatedly performed 20 different actions to generate a total of 567 sequences. The action classes are defined as follows: ‘Band’, ‘DrawCircle’, ‘DrawCross’, ‘DrawTick’, ‘ForwardKick’, ‘ForwardPunch’, ‘GolfSwing’, ‘Hammer’, ‘HandCatch’, ‘HandClap’, ‘HighArmWave’, ‘HighThrow’ ‘HorizontalArmWave’, ‘Jogging’, ‘PickUpandThrow’, ‘SideBoxing’, ‘SideKick’, ‘TennisServe’, ‘TennisSwing’, ‘TwoHandWave’. Figure 7b shows an example of ‘TwoHandWave’ action frames in the MSR-Action3D dataset.
4.1.3. UTD-MHAD Dataset

The UTD-MHAD dataset was captured by the members of the embedded systems and signal processing laboratory at the University of Texas at Dallas. This dataset contains a total of 861 sequences which are performed by 8 different subjects. For better generalization and variability, both male and female subjects are considered while capturing the action sequences. It has a total of 27 classes of actions (‘ArmCross’, ‘ArmCurl’, ‘BaseballSwing’, ‘BasketballShoot’, ‘Bowling’, ‘Boxing’, ‘Catch’, ‘Clap’, ‘DrawCircle (CLW)’, ‘DrawCircle (CCLW)’, ‘DrawTriangle’, ‘DrawX’, ‘Jog’, ‘Knock’, ‘Lunge’, ‘PickUpandThrow’, ‘Push’, ‘SitToStand’, ‘Squat’, ‘StandToSit’, ‘SwipeLeft’, ‘SwipeRight’, ‘TennisServe’, ‘TennisSwing’, ‘Throw’, ‘Walk’, ‘Wave’). Figure 7c shows an example of the ‘BaseballSwing’ action in the UTD-MHAD dataset.

4.1.4. Settings of the Training and Testing Dataset

We split the dataset into training and testing datasets. For the DHA, MSR-Action3D, and UTD-MHAD datasets, we follow the training and testing configuration as described in [51]. The total ratio of training and testing data in each dataset is as follows: DHA dataset (253:230), MSR-Action3D dataset (292:275), and UTD-MHAD dataset (431:430).

4.2. Environmental Setup and Evaluation Metrics

We accomplished the overall experiments in the Linux-20.04 environment. The hardware devices including CPU-Intel (R) Core (TM) i7 and GPU-GeForce GTX 1080 are used to perform the experiments. We used Python-3.8 and Matlab-202a as the programming languages. We trained the deep learning model until 100 epochs. The batch size and learning rate were set to 16 and 0.001, respectively. The learning rate drops 10% after every 20 epochs. For optimization, we conducted stochastic gradient descent (SGD) optimizer with a momentum of 0.9. We reported the recognition results by calculating the accuracy, defined as follows:

\[
\text{Accuracy}(\%) = \frac{\text{Correctly Predicted Samples}}{\text{Total Number of Samples}} \times 100
\]  \hspace{1cm} (4)

Owing to the smaller number of video sequences in action recognition datasets, we have used transfer learning in which, first, we have trained the model with the jester dataset [12]. Then, the human action datasets are trained with the pre-trained weight of the ResNet101 model.

4.3. Performance Evaluations and Comparisons

We carefully analyze the dataset and choose \( k \) as 16, 20, and 24, with a difference of 4 that provides better results. We evaluate the performance of the proposed method on the DHA, MSR-Action3D, and UTD-MHAD human action datasets. Table 1 lists the recognition performance for the CCM and SSIM metrics with 16, 20, and 24-ranked frames.

| Metric      | Ranking | DHA Dataset | MSR-Action3D | UTD-MHAD | Average |
|-------------|---------|-------------|--------------|----------|---------|
| CCM (\( \rho \)) | \( k = 16 \) | 91.7% | 92.4% | 93.3% | 92.5% |
|             | \( k = 20 \) | 92.6% | 93.1% | 93.5% | 93.1% |
|             | \( k = 24 \) | 92.2% | 93.8% | 93.5% | 93.2% |
| Average     |         | 92.2% | 93.1% | 93.4% | 92.9% |
| SSIM (\( \psi \)) | \( k = 16 \) | 92.2% | 93.5% | 94.2% | 93.3% |
|             | \( k = 20 \) | 93.0% | 94.2% | 94.4% | 93.9% |
|             | \( k = 24 \) | 93.5% | 94.6% | 94.7% | 94.2% |
| Average     |         | 92.9% | 94.1% | 94.4% | 93.8% |

Table 1. Human action recognition results on the DHA, MSR-Action3D, and UTD-MHAD datasets.
The average recognition results using the CCM metric are approximately 92.2% for the DHA, 93.1% for the MSR-Action3D, and 93.4% for the UTD-MHAD datasets, respectively. For the SSIM metric, the proposed method achieves approximately 92.9% for the DHA, 94.1% for the MSR-Action3D, and 94.4% for the UTD-MHAD datasets. If the value of $k$ is increased from 16 to 20, the average performance improvement is about 0.6% for the CCM and SSIM metrics. On the other hand, if the value of $k$ is increased from 20 to 24, the average performance improvement is about 0.1% for the CCM and 0.3% for the SSIM metrics, which are comparatively much lower. This is because the temporal information remains almost the same even when we increase the number of sampling frames along the temporal direction, such as 16, 20, and 24. However, if we compare the effectiveness of sampling metrics, the SSIM metric can sample more effectively, achieving an average accuracy of 93.8%, than the CCM metric with an average accuracy of 92.9%.

To show the effectiveness of the proposed system on different datasets, we compare the recognition results with several state-of-the-art methods, as provided in Table 2. The first, second, and third columns represent the recognition results of the prior works on the DHA, MSR-Action3D, and UTD-MHAD datasets, respectively. The proposed method can improve by approximately 10%, 5%, and 9% greater average accuracy than the prior works on the DHA, MSR-Action3D, and UTD-MHAD datasets, correspondingly. This is because most of the prior methods encoded the entire sequence into the dynamic images and used 2D-CNN or traditional machine learning techniques to classify the human action. While encoding an action into the spatial format, it cannot capture full temporal changes that reduce the overall performance. Even though few methods conducted 3D-CNN and LSTM, they did not provide better results because of the selection of frames from a whole sequence for training and testing. The network configurations also have a great effect on the performance of the human action recognition systems. The proposed method ensures better performance on the three different datasets because the proposed ranking metric can effectively select $k$-ranked frames that contain meaningful temporal information. The 3D-CNN can extract the discriminative features from the selected frames and provides better results.

| Methods | DHA Dataset | MSR-Action3D | UTD-MHAD |
|---------|-------------|--------------|----------|
| DAAL [27] | - | 92.3% | - |
| DSTIP+DCSF [32] | - | 89.3% | - |
| SDM-BSM [35] | 89.5% | - | - |
| DMM+CL-HOG [36] | - | 89.7% | 83.5% |
| DMM-MV [36] | - | 91.9% | 85.3% |
| DMM-LOGP [36] | - | 84.2% | 88.4% |
| VD-DMM [38] | - | - | 85.1% |
| WDDM-HOG [39] | - | 91.9% | - |
| EGSA [41] | - | 89.5% | 82.8% |
| TGSA [41] | - | 89.8% | 76.2% |
| LP-DMI-HOG [42] | 91.9% | 84.4% | 85.1% |
| LP-DMI-VGG [42] | 84.4% | 91.9% | 81.9% |
| ACG [43] | - | - | 87.7% |
| VCDN [45] | 79.8% | - | - |
| VCA [47] | 80.9% | - | - |
| VCA-Entropy [47] | 82.6% | - | - |
| SeMix [47] | 82.7% | - | - |
| LSTM [48] | 67.7% | - | - |
| **Ours ($\rho$)** | 92.2% | 93.1% | 93.4% |
| **Ours ($\psi$)** | 92.9% | 94.1% | 94.4% |

From the above results, it can be stated that the proposed method with the SSIM metric having $k = 24$ provides the best or similar results. We show the confusion chart for SSIM
TwoHandWave 100%
ArmSwing 20% 80%
GolfSwing 90% 10%
FrontClap 10% 90%
FrontBox 10% 90%
ArmCurl 90% 10%
SideClap 100%
LegKick 70% 30%
LegCurl 100%
SideBox 100%
TaiChi 100%
Pjump 100%
Jump 100%
Walk 10% 90%
Bend 100%
Pitch 100%
Kick 100%
Jack 100%
Skip 100%
Clap 100%
Wrist 100%
Finger 100%
Throw 100%
HandCatch 100%
Claw 100%
FingerL 100%
FingerR 100%
HandL 100%
HandR 100%
TwoHandWave 100%
Walk 100%

Figure 8. Confusion chart of the DHA dataset (best results with SSIM, k = 24).

Likewise, for the DHA dataset, we provided the confusion chart for the best results of the MSR-Action3D dataset, as shown in Figure 9. The lowest recognition result is about 78.6% for the ‘DrawCross’ action. This action is equally misclassified as ‘ForwardPunch’, ‘HandCatch’, and ‘HorizontalArmWave’ actions, with an accuracy of 7.1%. On the other hand, most of the actions are recognized correctly, with an accuracy of 100% in the MSR-Action3D dataset.

Figure 10 visualizes the confusion chart for the UTD-MHAD dataset in which the minimum accuracy is observed for ‘Throw’, ‘ArmCurl’, and ‘Clap’ actions of approximately 68.8%, 81.3%, and 81.3%, correspondingly. The ‘Throw’ is recognized as ‘Catch’, ‘DrawCircle (CLW)’, ‘Knock’, and ‘SwipeLeft’ actions. On the other hand, the ‘ArmCurl’ action is misclassified as the ‘ArmCross’ action with approximately 18.8% accuracy, which is the highest misclassification rate.

4.4. Ablation Study

We summarize the performance of the proposed methods based on ranking metrics and datasets, as shown in Figure 11. For any dataset, the SSIM rank metric works much better than the CCM metric. Likewise, the recognition results are higher for the UTD-MHAD dataset than for the DHA and MSR-Action3D datasets with the SSIM and CCM ranking metrics, because the action sequences in the UTD-MHAD dataset are more accurately collected.
same, or a slight change happened. This is because only a certain segment in a sequence contains action information, and other frames are static that do not have any effect on the overall recognition performance.

Figure 9. Confusion chart of the MSR-Action3D dataset (best results with SSIM, $k = 24$).

Figure 10. Confusion chart of UTD-MHAD dataset (best results with SSIM, $k = 24$).
To show the effectiveness of the number of key-frames sampling on the recognition performance, we additionally perform experiments with 28, 32, and 36 frames using the UTD-MHAD dataset. Figure 12 depicts the recognition results for different rank values ranging from 16 to 36 with a difference of 4 frames. The recognition accuracy is 94.2% for 16-ranked frames, which improves slightly by about 0.2% while increasing the number of key-frames to the next level 20. From 20 to 36, the classification results remain almost the same, or a slight change happened. This is because only a certain segment in a sequence contains action information, and other frames are static that do not have any effect on the overall recognition performance.

4.5. Complexity Analysis

We determine the network complexity in terms of parameters, floating-point operations (FLOPs), and testing time, as given in Table 3. We only consider the UTD-MHAD dataset with 16, 20, and 24 frames to report the results. The parameters, FLOPs, and testing time are given in million (M), giga (G), and second (s). The testing time for action recognition is calculated by averaging over all testing sequences. The total number of parameters in ResNet101 for UTD-MHAD is 47.58 M, which changes for DHA and MSR-Action3D datasets as 47.57 M and 47.56 M, respectively, as the number of classes varies. We compare the time complexity of the proposed system with some of the state-of-the-art methods. The prior works take too much time to recognize an action using the DHA, MSR-Action3D, and
UTD-MHAD datasets, since the encoding of action requires a greater amount of time than directly extracting discriminative features from raw depth sequences using 3D-CNN in the proposed method. We do not need to spend time in encoding an action that reduces the time complexity significantly.

Table 3. Network complexity analysis.

| Methods          | Datasets          | Parameters (M) | FLOPs (G) | Time (s) |
|------------------|-------------------|----------------|-----------|----------|
| State-of-the-art | WDMM-HOG [38]     | MSR-Action3D   | -         | -        | 0.8      |
| Methods          | LP-DMI-HOG [42]   | DHA            | -         | -        | 0.12     |
| Methods          | LP-DMI-HOG [42]   | MSR-Action3D   | -         | -        | 0.23     |
| Methods          | LP-DMI-HOG [42]   | UTD-MHAD       | -         | -        | 0.17     |
| Proposed         | k = 16            | 47.58          | 9.67      | 0.036    |
| Proposed         | k = 20            | 47.58          | 12.95     | 0.040    |
| Proposed         | k = 24            | 47.58          | 14.55     | 0.042    |

5. Discussion

The widespread applications of modern technology such as human–machine or human–object interaction are significantly increasing. In this article, we propose a new approach for human action recognition using 3D-CNN with the raw depth dataset. Several methods have introduced human action recognition using depth sequences, in which they tried to encode the entire sequence into a spatial format called dynamic image. Some methods also focused on the extraction of the skeleton and point cloud information from depth video. These methods are completely dependent on the effective encoding, skeleton, or point cloud extraction to correctly recognize the human action. Generally, it is very difficult or sometimes impossible to maintain all the temporal information during the encoding process. Very few methods used raw depth sequence to avoid human interference to generate dynamic image, skeleton, or point cloud information. Even though they used raw depth sequence, the recognition performance was much lower.

We analyze the dataset and find that each sequence has a different number of frames. When a sequence contains a large number of frames, most of the frames remain the same, which means there are no spatial and temporal changes. Only a few numbers of frames include action information that changes in the spatial and temporal direction. As a result, we propose a novel ranking-based approach for human action recognition using 3D-CNN with raw depth sequences. We use SSIM and CCM ranking metrics to rank the whole sequence and select $k$-ranked frames that contain more spatial and temporal information. Then, we train a 3D-CNN model with the $k$-ranked frames to recognize specific actions. We do not need to encode the sequence and extract dynamic images for further classification. We consider different levels of ranking and investigate the recognition performance for better understanding. After a certain level of $k$, the recognition performance does not improve further because we must copy the same frame multiple times in most of the sequences.

6. Conclusions

The extraction of discriminative features from the depth sequence is cumbersome work in the case of action recognition. Thus, it is always demanding to build a system that can directly process the raw sequences and extract the discriminative features for classification. There are several methods suggested for hand gesture recognition with raw depth sequences directly using the 3D-CNN models. However, there are very few methods that have considered raw depth videos to discriminate human activity using RNN and LSTM. Even though they consider raw depth videos, they still have many limitations such as complexity, effectiveness, efficiency, and robustness. In this paper, we derived a
A novel approach to sample key-frames and passed through a 3D-CNN model to perform the classification. Different levels of key-frames sampling were considered to evaluate the robustness of the proposed method. We also applied the proposed system on three different benchmark datasets to show the generalization power. We provided recognition accuracy, performance comparisons, and confusion charts as the experimental results. The proposed method assured better results over the state-of-the-art works.

**Author Contributions:** Conceptualization, analysis, methodology, manuscript preparation, and experiments, N.T.; data curation, writing—review and editing, N.T. and J.-H.B.; supervision, J.-H.B.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This research is supported by the GRRC program of Gyeonggi province (GRRC Aviation 2017-B04, Development of Intelligent Interactive Media and Space Convergence Application System).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We would like to acknowledge Korea Aerospace University with much appreciation for its ongoing support of our research.

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

**References**

1. Dawar, N.; Kehtarnavaz, N. Continuous detection and recognition of actions of interest among actions of non-interest using a depth camera. In Proceedings of the IEEE International Conference on Image Processing, Beijing, China, 17–20 September 2017. [CrossRef]
2. Zhu, H.; Vial, R.; Lu, S. Tornado: A spatio-temporal convolutional regression network for video action proposal. In Proceedings of the CVPR, Venice, Italy, 22–29 October 2017. [CrossRef]
3. Wen, R.; Nguyen, B.P.; Chng, C.B.; Chui, C.K. In Situ Spatial AR Surgical Planning Using projector-Kinect System. In Proceedings of the Fourth Symposium on Information and Communication Technology, Da Nang, Vietnam, 5–6 December 2013. [CrossRef]
4. Azuma, R.T. A survey of augmented reality. Presence Teleoperators Virtual Environ. **1997**, *6*, 355–385. [CrossRef]
5. Fangbemi, A.S.; Liu, B.; Yu, N.H. Efficient human action recognition interface for augmented and virtual reality applications based on binary descriptor. In Proceedings of the International Conference on Augmented Reality, Virtual Reality and Computer Graphics, Otranto, Italy, 24–27 June 2018. [CrossRef]
6. Jalal, A.; Kamal, S.; Kim, D. A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor Environments. **Sensors** **2014**, *14*, 11735–11759. [CrossRef] [PubMed]
7. Chen, L.; Ma, N.; Wang, P.; Li, J.; Wang, P.; Pang, G.; Shi, X. Survey of pedestrian action recognition techniques for autonomous driving. *Tsinghua Sci. Technol.* **2020**, *25*, 458–470. [CrossRef]
8. Bloom, V.; Makris, D.; Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012. [CrossRef]
9. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
10. Wang, K.; Peng, X.; Yang, J.; Lu, S.; Qiao, Y. Suppressing uncertainties for large-scale facial expression recognition. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020. [CrossRef]
11. Fu, R.; Wu, T.; Luo, Z.; Duan, F.; Qiao, X.; Guo, P. Learning Behavior Analysis in Classroom Based on Deep Learning. In Proceedings of the Tenth International Conference on Intelligent Control and Information Processing (ICICIP), Marrakesh, Morocco, 14–19 December 2019. [CrossRef]
12. Köpüklü, O.; Gunduz, A.; Kose, N.; Rigoll, G. Real-time hand gesture detection and classification using convolutional neural networks. In Proceedings of the 14th International Conference on Automatic Face & Gesture Recog. (FG), Lille, France, 14–18 May 2019. [CrossRef]
13. Ameur, S.; Kahliifa, A.B.; Bouhlel, M.S. A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion. *Entertain. Comput.* **2020**, *35*, 100373. [CrossRef]
14. D’Eusanio, A.; Simoni, A.; Pini, S.; Borghi, G.; Vezzani, R.; Cucchiara, R. A Transformer-Based Network for Dynamic Hand Gesture Recognition. In Proceedings of the International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020. [CrossRef]
15. Liu, T.; Song, Y.; Gu, Y.; Li, A. Human action recognition based on depth images from Microsoft Kinect. In Proceedings of the Fourth Global Congress on Intelligent Systems, Hong Kong, China, 3–4 December 2013. [CrossRef]
16. Ahmad, Z.; Khan, N. Inertial Sensor Data to Image Encoding for Human Action Recognition. *IEEE Sens. J.* 2021, 9, 10978–10988. [CrossRef]
17. Hou, Y.; Li, Z.; Wang, P.; Li, W. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.* 2016, 28, 807–811. [CrossRef]
18. Tasnim, N.; Islam, M.; Baek, J.H. Deep Learning-Based Action Recognition Using 3D Skeleton Joints Information. *Inventions* 2020, 5, 49. [CrossRef]
19. Li, C.; Hou, Y.; Wang, P.; Li, W. Joint distance maps-based action recognition with convolutional neural networks. *IEEE Signal Process. Lett.* 2017, 24, 624–628. [CrossRef]
20. Tasnim, N.; Islam, M.K.; Baek, J.H. Deep Learning Based Human Activity Recognition Using Spatio-Temporal Image Formation of Skeleton Joints. *Appl. Sci.* 2021, 11, 2675. [CrossRef]
21. Mahjoub, A.B.; Atri, M. Human action recognition using RGB data. In Proceedings of the 11th International Design & Test Symposium (IDT), Tunisia, Hammamet, 18–20 December 2016. [CrossRef]
22. Verma, P.; Sah, A.; Srivastava, R. Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition. *Multimed. Syst.* 2020, 26, 671–685. [CrossRef]
23. Dhiman, C.; Vishwakarma, D.K. View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Trans. Image Proc.* 2020, 29, 3835–3844. [CrossRef] [PubMed]
24. Yang, X.; Tian, Y.L. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Providence, RI, USA, 16–21 June 2012. [CrossRef]
25. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3d joints. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012. [CrossRef]
26. Ji, X.; Cheng, J.; Feng, W.; Tao, D. Skeleton embedded motion body partition for human action recognition using depth sequences. *Signal Process.* 2018, 143, 56–68. [CrossRef]
27. Zhang, C.; Tian, Y.; Guo, X.; Liu, J. DAAL: Deep activation-based attribute learning for action recognition in depth videos. *Comput. Vis. Image Underst.* 2018, 167, 37–49. [CrossRef]
28. Li, W.; Zhang, Z.; Liu, Z. Action recognition based on a bag of 3d points. In Proceedings of the Conference on Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 13–18 June 2010. [CrossRef]
29. Rahmani, H.; Mahmood, A.; Huynh, D.Q.; Mian, A. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition. In Proceedings of the European conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014. [CrossRef]
30. Li, D.; Jahan, H.; Huang, X.; Feng, Z. Human action recognition method based on historical point cloud trajectory characteristics. *Vis. Comput.* 2021, 37, 1–9. [CrossRef]
31. Megavannan, V.; Agarwal, B.; Babu, R.V. Human action recognition using depth maps. In Proceedings of the IEEE International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 22–25 July 2012. [CrossRef]
32. Xia, L.; Aggarwal, J.K. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013. [CrossRef]
33. Eum, H.; Yoon, C.; Lee, H.; Park, M. Continuous human action recognition using depth-MHI-HOG and a spotlight model. *Sensors* 2015, 15, 5197–5227. [CrossRef]
34. Bulbul, M.F.; Jiang, Y.; Ma, J. Human action recognition based on DMMs, HOGs and Contourlet transform. In Proceedings of the International Conference on Multimedia Big Data, Beijing, China, 20–22 April 2015. [CrossRef]
35. Liu, H.; Tian, L.; Liu, M.; Tang, H. Sdm-bsm: A fusing depth scheme for human action recognition. In Proceedings of the International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015. [CrossRef]
36. Bulbul, M.F.; Jiang, Y.; Ma, J. DMMs-based multiple features fusion for human action recognition. *Int. J. Multimed. Data Eng.* Manag. 2015, 6, 23–39. [CrossRef]
37. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *J. Real-Time Image Process.* 2016, 12, 155–163. [CrossRef]
38. Jin, K.; Jiang, M.; Kong, J.; Hoo, H.; Wang, X. Action recognition using vague division DMMs. *J. Eng.* 2017, 4, 77–84. [CrossRef]
39. Azad, R.; Asadi-Aghbolaghi, M.; Kasaei, S.; Escalera, S. Dynamic 3D hand gesture recognition by learning weighted depth motion maps. *IEEE Trans. Circ. Syst. Video Technol.* 2018, 12, 1729–1740. [CrossRef]
40. Li, Z.; Zheng, Z.; Lin, F.; Leung, H.; Li, Q. Action recognition from depth sequence using depth motion maps-based local ternary patterns and CNN. *Multimed. Tools Appl.* 2019, 78, 19587–19601. [CrossRef]
41. Liang, C.; Liu, D.; Qi, L.; Guan, L. Multi-modal human action recognition with sub-action exploiting and class-privacy preserved collaborative representation learning. *IEEE Access* 2020, 8, 39920–39933. [CrossRef]
42. Li, C.; Huang, Q.; Li, X.; Wu, Q. Human Action Recognition Based on Multi-scale Feature Maps from Depth Video Sequences. *arXiv* 2021, arXiv:2101.07618. [CrossRef]
43. Bulbul, M.F.; Tabussum, S.; Ali, H.; Zheng, W.; Lee, M.Y.; Ullah, A. Exploring 3D Human Action Recognition Using STACOG on Multi-View Depth Motion Maps Sequences. *Sensors* 2021, 11, 3642. [CrossRef] [PubMed]
44. Pareek, P.; Thakkar, A. RGB-D based human action recognition using evolutionary self-adaptive extreme learning machine with knowledge-based control parameters. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 1–19. [CrossRef]
45. Wang, L.; Ding, Z.; Tao, Z.; Liu, Y.; Fu, Y. Generative multi-view human action recognition. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019. [CrossRef]
46. Sanchez-Caballero, A.; de López-Diz, S.; Fuentes-Jimenez, D.; Losada-Gutierrez, C.; Marrón-Romera, M.; Casillas-Perez, D.; Sarker, M.I. 3dfcn: Real-time action recognition using 3d deep neural networks with raw depth information. *arXiv* **2020**, arXiv:2006.07743. [CrossRef]
47. Liu, Y.; Wang, L.; Bai, Y.; Qin, C.; Ding, Z.; Fu, Y. Generative View-Correlation Adaptation for Semi-supervised Multi-view Learning. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020. [CrossRef]
48. Bai, Y.; Tao, Z.; Wang, L.; Li, S.; Yin, Y.; Fu, Y. Collaborative Attention Mechanism for Multi-View Action Recognition. *arXiv* **2020**, arXiv:2009.06599.
49. Jamshidi, M.B.; Talla, J.; Peroutka, Z. Deep Learning Techniques for Model Reference Adaptive Control and Identification of Complex Systems. In Proceedings of the 2020 19th International Conference on Mechatronics-Mechatronika (ME), Prague, Czech Republic, 2–4 December 2020. [CrossRef]
50. Khalaj, O.; Jamshidi, M.B.; Saebnoori, E.; Mašek, B.; Štadler, C.; Svoboda, J. Hybrid Machine Learning Techniques and Computational Mechanics: Estimating the Dynamic Behavior of Oxide Precipitation Hardened Steel. *IEEE Access* **2021**, *9*, 156930–156946. [CrossRef]
51. Jamshidi, M.B.; Lalbakhsh, A.; Talla, J.; Peroutka, Z.; Roshani, S.; Matousek, V.; Roshani, S.; Mirmozafari, M.; Malek, Z.; Spada, L.L.; et al. Deep Learning Techniques and COVID-19 Drug Discovery: Fundamentals, State-of-the-Art and Future Directions. In *Emerging Technologies during the Era of COVID-19 Pandemic*; Springer: Cham, Switzerland, 2021. [CrossRef]
52. Carreira, J.; Zisserman, A.; Quo, V. Action recognition? a new model and the kinetics dataset. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017. [CrossRef]
53. Facebook Research. Available online: https://github.com/facebookresearch/pytorchvideo/tree/main/pytorchvideo/models (accessed on 20 March 2022).
54. Lin, Y.C.; Hu, M.C.; Cheng, W.H.; Hsieh, Y.H.; Chen, H.M. Human action recognition and retrieval using sole depth information. In Proceedings of the 20th ACM international conference on Multimedia, New York, NY, USA, 29 October–2 November 2012. [CrossRef]
55. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor. In Proceedings of the IEEE International Conference on Image Processing, Quebec City, QC, Canada, 27–30 September 2015. [CrossRef]