applied sciences

MDPI

Article

Enhanced Spatial Stream of Two-Stream Network Using Optical Flow for Human Action Recognition

Shahbaz Khan [1], Ali Hassan [1,*], Farhan Hussain [1], Aqib Perwaiz [1], Farhan Riaz [2], Maazen Alsabaan [3] and Wadood Abdul [3]

[1] Department of Computer and Software Engineering, College of Electrical and Mechanical Engineering, National University of Sciences and Technology, Islamabad 46000, Pakistan

[2] School of Computer Science, University of Lincoln, Lincoln LN6 7TS, UK

[3] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia

* Correspondence: alihassan@ceme.nust.edu.pk

Abstract: *Introduction:* Convolutional neural networks (CNNs) have maintained their dominance in deep learning methods for human action recognition (HAR) and other computer vision tasks. However, the need for a large amount of training data always restricts the performance of CNNs. *Method*: This paper is inspired by the two-stream network, where a CNN is deployed to train the network by using the spatial and temporal aspects of an activity, thus exploiting the strengths of both networks to achieve better accuracy. *Contributions*: Our contribution is twofold: first, we deploy an enhanced spatial stream, and it is demonstrated that models pre-trained on a larger dataset, when used in the spatial stream, yield good performance instead of training the entire model from scratch. Second, a dataset augmentation technique is presented to minimize overfitting of CNNs, where we increase the dataset size by performing various transformations on the images such as rotation and flipping, etc. *Results*: UCF101 is a standard benchmark dataset for action videos, and our architecture has been trained and validated on it. Compared with the other two-stream networks, our results outperformed them in terms of accuracy.

Keywords: deep learning; human action recognition; overfitting; two-stream network

check for updates

Citation: Khan, S.; Hassan, A.; Hussain, F.; Perwaiz, A.; Riaz, F.; Alsabaan, M.; Abdul, W. Enhanced Spatial Stream of Two-Stream Network Using Optical Flow for Human Action Recognition. *Appl. Sci.* **2023**, *13*, 8003. https://doi.org/10.3390/app13148003

Academic Editor: João M. F. Rodrigues

Received: 13 February 2023
Revised: 24 June 2023
Accepted: 26 June 2023
Published: 8 July 2023

1. Introduction

Video-based human action recognition (HAR) is one of the most significant study fields in computer vision, with algorithms becoming more effective by the day. HAR has many applications ranging from surveillance, video tagging, activity/event detection, etc. [1,2]. The main goal in human action recognition is to automatically detect the sort of actions that are being performed in the video, e.g., archery, basketball shooting, horse riding, etc. It is a very difficult task owing to the many problems it entails, such as camera motion, varying lighting conditions, backdrop bombardment, various human forms, occlusion, perspective fluctuation, and so on. Such varied changes in the videos make this task challenging and interesting to solve using automated methods. However, the impact of these problems varies depending on the type of action. The four primary kind of action are gestures, interactions, actions, and group activities.

Several researchers have used different modalities for handling HAR, including sensor-based methods [3] and video streams [4]. Due to the vast availability of video data, a major portion of researchers concentrate on video processing for HAR, which is the significant topic of work in this paper as well.

Different strategies have been used to extract the motions inside the pictures and videos. In general, recent research on HAR mainly focuses on spatial and temporal models. Sparse segments are used to simulate long-term temporal structure in temporal segment networks (TSNs) [5]. The utilization of the concept that models learn hierarchical motion patterns within image space has been employed to address diverse temporal modeling challenges in 3D CNN networks [6,7]. Tracking characteristics are employed to improve the effectiveness of temporal modeling [7,8]. As HAR is a challenging domain, the authors in [9] provided a thorough literature review of the available datasets. They identified around 68 different datasets, of which 28 are heterogeneous and 40 are for specific human actions.

In [4], the authors proposed a three-step framework for human activity recognition comprising background subtraction, parameter extraction, and evaluation. In [10], the authors presented a knowledge representation framework to detect the occurrence of specific events defined as targets in the surveillance scenario. In [11], the authors presented high-level video event modeling and recognition based on a Petri net. The results shown are encouraging as the event recognition is fully automated. In [12], the authors presented a framework to jointly learn a view-invariance transfer dictionary, and subsequently a view-invariant classifier. This framework has allowed them to obtain improved performance on the available video datasets.

In [13–15], the authors emphasize using convolutional networks (ConvNets) for this task of HAR. Researchers have shown that temporal filters, such as local spatiotemporal filters, can be applied to spatiotemporal objects such as actions, making it possible to use spatial recognition ideas on temporal objects [14–16]. This difference between time and space is significant, and different techniques have been examined, such as adding optical flow networks (which simulate motion) [17] or modeling time sequences in recurrent structures (which represent patterns in nature) [18–20].

The two-stream architecture is based on a hypothesis that came out of neuroscience research called the two-stream hypothesis [13], which states that the visual cortex has two separate pathways: (i) the ventral pathway, which processes information about the visual attributes of objects such as shape and color, and (ii) the dorsal pathway, that responds to transformations in the object, and to spatial relationships as an object of motion. In a typical implementation of a two-stream network, both streams use a different set of inputs to classify actions. One of the streams is trained on stacked RGB images, hence is referred to as 'the spatial stream'. The second stream, referred as 'the temporal stream', uses motion vector-based images as its input to train, which need to be computed beforehand, and thus this can also be time consuming. As both streams use different types of inputs to train their networks, they can take advantage of different methods of feature representation and extraction, combining the strengths of both. Both streams are merged via different techniques to form 'the two-stream network', and the classification results are calculated. However, the problem of overfitting exists in both streams as deep learning networks need a large amount of data for stable weight training.

As mentioned earlier, the two-stream network approach made a breakthrough in action recognition as some activities are time oriented (temporal) and others are scene oriented (spatial) in a single dataset. Using any single approach would fail to recognize other types of activity. Hence, we used the two-stream network approach.

The contributions of this paper are as follows:

- We demonstrate (using experiments) that using pre-trained models in the spatial stream yields good performance results as compared to training the entire model from scratch and it also saves time. We achieve this by freezing the classification layers of the original two-stream model [13] and attaching the feature extraction layers of different pre-trained models one by one and then training the fully connected layers only to check which model performs best. After selecting the best model, we fine-tune the whole network to see if the results can be improved.
- We present a strategy to deal with the problem of overfitting by using dataset augmentation. The main reason behind overfitting is the limited dataset provided to a deep network to train its model. We incorporate dataset augmentation to increase our dataset size using different augmentation techniques, including horizontal image flipping and rotation.

Our proposed method can avoid training the entire model from scratch, which saves time and avoids the use of high-cost computational resources without compromising the results. The remaining paper is structured as follows: Section 2 gives the relevant state of the art work performed in this domain of video-based HAR, Section 3 gives the details of the proposed methodology, Section 4 gives the implementation details, and this is followed by the experimental results in Section 5 and conclusions in Section 6.

2. State of the Art

In [13], the authors use threefold cross-validation to improve HAR accuracy. Initially, they introduce a concept of two-stream ConvNet that mixes spatial and temporal networks. Secondly, they claim that despite little training data, a ConvNet can be trained on dense optical flow from several consecutive frames and yet achieve outstanding results. They finalize their presentation by demonstrating the potential application of multitask learning, which can effectively amplify the volume of training data and improve performance as a whole. The most important feature is local trajectory pooling, with spatial and temporal tubes that are coordinated across spatiotemporal layers to concentrate on trajectories. Even while the network can detect the optical flow along the trajectories, it ignores trajectories in spatial pooling. However, the designed network still needs to catch up with the current state-of-the-art shallow representation [21] in terms of the achieved accuracy.

In [22], spatiotemporal ResNets were used as a combination of these two methods. To begin, residual connections between the appearance and motion channels of a two-stream architecture were injected to allow for spatiotemporal stream interaction. Then, the learned convolutional filters were applied to adjacent feature maps in time to convert pre-trained image ConvNets to spatiotemporal networks.

In their subsequent paper [23], the same authors have reduced the parameters by merging the spatial and temporal networks at a convolutional layer, with no performance loss. A spatiotemporal architecture was designed for two-stream networks comprising a novel temporal and a convolutional fusion layers, which were connected to the networks. Regarding performance, the innovative design exceeded the top rank on two common benchmark datasets without significantly increasing the number of parameters. According to the findings, it was found that learning correspondence between ConvNet characteristics that are very abstract in both space and time is extremely important.

In [24], remodeling of the dataset was deployed for initializing model learning by using the augmentation of data, and ResNet101 layer parameters trained on datasets such as ImageNet were used to deal with the overfitting issues caused by having less data. Deeper ConvNet was developed for learning the complexity of action. Using a disorder testing and training method, the model and procedure may provide a substantial boost in action recognition. The experiments showed that the strategy beats current top-ranked methodologies on two advanced datasets, the UCF101 [25] and the KTH action datasets [26]. The temporal network with deeper convolutional networks did not perform well compared to the appearance networks on the UCF101 dataset during the experimental evaluation. The following potential alternative might help to overcome this constraint where it was proposed to capture information on motion with a deep temporal structure by adopting deeper recurrent neural networks (RNNs).

A two-stream adaptive graph convolutional network (2S-AGCN) was designed specifically for action recognition in [27], which uses the skeleton technique. In this technique, graph convolutional networks (GCNs) are used, which model the human body skeleton as spatiotemporal graphs. The backpropagation technique may learn the network architecture either uniformly or individually as it goes along. Making this data-driven technique part of the model increases the model's flexibility for constructing graphs and increases the model's generality for varying data samples. To explain both first-order and second-order information, a two-stream framework was developed, and a significant improvement in recognition accuracy was achieved.

In [28], the authors proposed a motion-attentive transition network for zero-shot video object segmentation. They named this network MATNet; it uses a two-stream encoder network to treat motion and appearance independently in separate streams. The authors tested their network on four challenging public benchmark datasets and showed the effectiveness of their network.

In [29], residual images were used to feed the temporal stream of the network rather than conventional optical flow images. This reduced the computational requirements, due to less data to process, and increased the accuracy in comparison to the state of the art models. Because residual frames offer minimal information on object appearance, they utilized a 2D convolutional network to extract appearance features and combined them with residual frame findings to build a two-path solution, reporting a marked improvement in the speed of execution and accuracy. Table 1 shows the state-of-the-art research work on the UCF101 dataset, where we have summarised the work of researchers who have used optical flow.

Table 1. Summary of performance of state-of-the-art methods on UCF101 dataset.

Author	Method	Optical Flow	Accuracy (%)
[13]	Spatial and temporal two-stream networks	✓	86.9
[22]	Spatiotemporal ResNet	✗	93.4
[23]	Late fusion of two-stream network	✗	93.5
[24]	Deeper two-stream ConvNets	✓	95.1
[28]	Two-stream ResNets with encoder/decoder setup	✓	82.4
[29]	Two-path network	✗	90.6

3. Proposed Methodology

The block diagram of the proposed deep two-stream convolutional network is shown in Figure 1 with its respective spatial and temporal streams. The spatial stream uses each video frame as a single image for feature extraction and processing. It uses a pre-trained ImageNet model [30] as the feature extraction part. For the temporal stream architecture, we use a stack of optical flow fields as input to the architectures. Doing so achieves two goals: firstly, the data being processed reduces as there are less data in optical flow fields as compared to RGB images; secondly, the optical flow will capture the moving regions in an image, thus making it easier to identify the HAR. The details of each stream architecture are given in Sections 3.2 and 3.3.

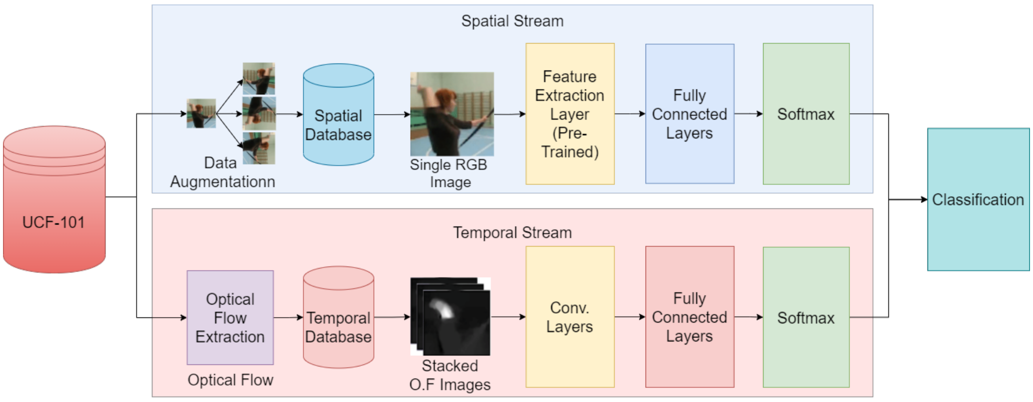

Figure 1. Proposed two-stream network's architecture.

3.1. Dataset

The UCF101 dataset contains a total of 13,320 videos from 101 human action classes extracted from videos in the wild. It has the widest variety in terms of actions, and it is the most complex dataset to date in this domain, with substantial differences in camera motion, object look and position, object size, perspective, cluttered backdrop, light conditions, and so on. This is a supplement to UCF50 dataset, that contains 50 activities. Because most accessible action recognition datasets are unrealistic and produced by actors, UCF101 intends to inspire future action recognition research by learning and exploring new realistic action categories. The videos are divided into 25 action groups, where each group may contain 4–7 clips. The details of the dataset are given in Table 2.

Table 2. Summary of UCF101 dataset.

Activity	Details
Actions	101
Clips	13,320
Groups per Action	25
Clips per Group	4–7
Total Duration	1600 min
Min. Clip Length	1.06 s
Resolution	320 × 240
Max. Clip Length	71.04 s
Frame Rate	25 fps

3.2. Spatial Stream

For some actions, a single frame from the whole video can be enough to recognize the actions correctly. This can be true mostly for actions that involve human–object interactions such as playing a guitar, discus throwing, hammering, or basketball. As in these action videos, recognizing an object correctly in a frame can lead to recognizing the associated action successfully. Keeping this hypothesis in mind, the spatial stream can simply be called a modified version of the image classification stream, which takes a single RGB image as input for action recognition.

Enhanced Spatial Stream

As discussed earlier, the spatial stream is in fact an image recognition architecture. Keeping this analysis in mind, we use advanced pretrained models like ImageNet to our advantage and train the classification part only instead of training the original CNN model from scratch. Then, we fine-tune them on our dataset (UCF101) using a transfer learning technique [31] to form an enhanced spatial stream by adding a fully connected layer at the end and running the training session for a few iterations. The proposed architecture of our enhanced stream is shown in Figure 2, where the feature extraction part of the pre-trained model is merged with the classification part of the original architecture from a two-stream network [13].

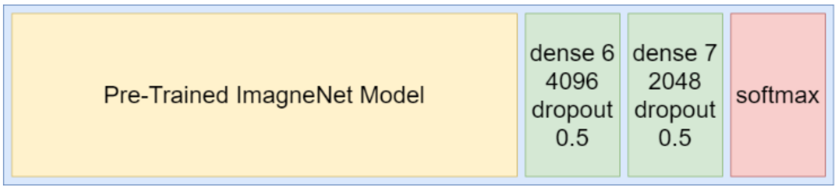

Figure 2. Enhanced model of spatial stream.

ImageNet [30] is a well-known dataset that has been extensively used in computer vision since its publication in 2015. The dataset contains over 100,000 classes and has been made available for educational and non-commercial research purposes. Several research teams, including Google, Nvidia, etc., have trained their models on this dataset and have made their trained models available for research purposes. Table 3 contains different models trained on ImageNet with the details of the number of layers in the model and the number of parameters being used in the specific model.

Table 3. Details of models pre-trained on ImageNet.

Model	Layers	Parameters
InceptionV3	159	23,851,784
VGG-16	23	138,357,544
Xception	126	22,910,480
MobileNet	88	4,253,864
MobileNetV2	88	3,538,984
NASNetMobile	–	5,326,716
DenseNet121	121	8,062,504
DenseNet169	169	14,307,880

In our experiments, we have have found quite a few similarities between the ImageNet and UCF101 datasets. As an example, UCF101 dataset has a class "WalkingWithDog" involving dogs, while the ImageNet also has examples containing dogs, such as "Maltese-Dog". Because of the similarities in the types of task between the two datasets, we used pre-trained models on ImageNet for our problem and tailored them to fit our problem by using transfer learning techniques. The different pre-trained models listed in Table 3 will be evaluated individually by training the classification layers only. The best-performing model will be selected and further fine-tuned to check for any further improvement in the performance.

3.3. Optical Flow Convolutional Networks

Although for certain activities, a single frame can be enough for recognizing object-oriented activities, however, time-oriented activities such as running, jogging, etc., need special input. Here, we describe the input to the ConvNet model for the temporal stream.

Optical flow displacement fields are stacked multiple times to form the input for the temporal stream. Such inputs will explicitly describe the motion between two consecutive frames. As a result, the network is freed of the need to estimate motion and can focus on pattern recognition. We present the details of the proposed optical flow based architecture in Figure 3.

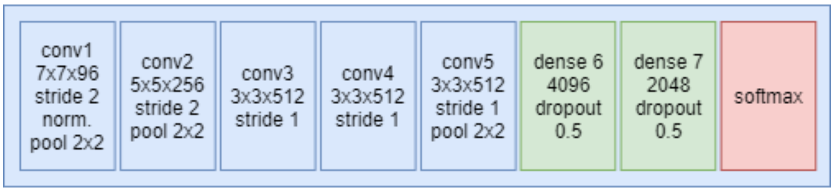

Figure 3. Convolutional model of our proposed architecture.

Stacked Optical Flow

Dense optical flow refers to a collection of displacement vectors, denoted as d_t, which represent the motion between pairs of consecutive frames in a sequence, where d is the displacement vector that represents the movement direction between two consecutive frames in frame t, as in Figure 4a,b. The displacement vector is a point (u, v), at locations u and v in a frame t, which moves to the corresponding point in frame $t+1$, thus, can be denoted by $d_t(u, v)$ for the t^{th} frame. The horizontal component d_t^x and vertical component d_t^y of the displacement vector and their optical flow representation as an image can be seen in Figure 4.

In the corresponding image, only the outlined region is being moved during two consecutive frames (an arm of the person). Figure 4a,b show the movement detection of the arm. Now, Figure 4c of the image shows the x component of the vector d and Figure 4d shows the y component of the vector d. Both the x and y components in the images are represented by white and black colors in the respective images. Motion across the consecutive frames is represented by stacking the vertical and horizontal components' displacement vectors of L continuous frames to form a total of $2L$ input channels. A ConvNet input volume $I_\tau \in \mathbb{R}^{w \times h \times 2L}$ for a video frame of width w and height h, for temporal stream with an arbitrary input τ, can be denoted as in Equations (1) and (2).

$$\begin{aligned} I_\tau(u, v, 2k-1) &= d^x_{\tau+k-1}(u, v) \qquad (1) \\ I_\tau(u, v, 2k-1) &= d^y_{\tau+k-1}(u, v) \qquad (2) \\ u &= [1; w], \\ v &= [1; h], \\ k &= [1 : L]. \end{aligned}$$

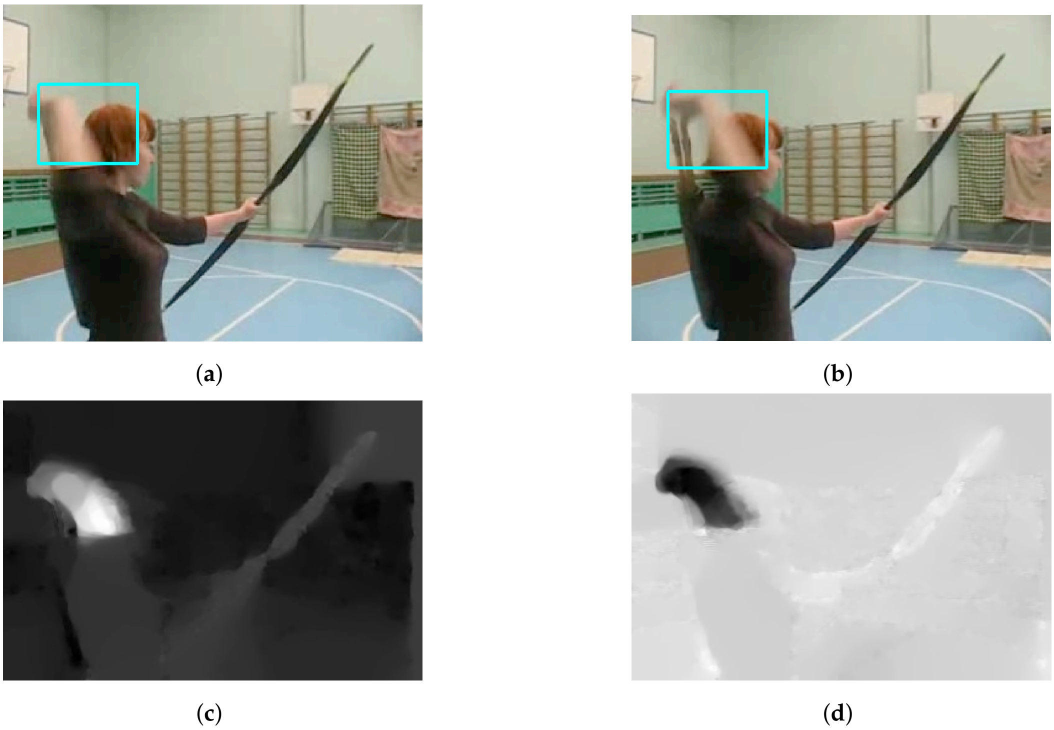

Figure 4. Output of the optical flow; (**a**,**b**) are the two consecutive video frames, (**c**,**d**) corresponding horizontal and vertical components of displacement vector from frames (**a**,**b**).

3.4. Avoiding Overfitting

Overfitting is mainly caused by having a small amount of data available to train a deep network. To tackle this issue, we adapt some helpful techniques to minimize overfitting in the spatial and temporal streams.

3.4.1. Dataset Augmentation for Spatial Stream

Data augmentation is a technique that allows practitioners to substantially enhance the variety of data available for training models without gathering new data. When training large neural networks, data augmentation methods such as trimming, padding, vertical and horizontal flipping, and rotation are frequently employed. However, data augmentation and its strategies that capture data invariances have received less attention than the neural network designs themselves. These techniques were employed during the training phase on the dataset to increase the number of images. Horizontal flip, horizontal shift, vertical shift, and rotation transformations were used to increase the total number of images from 2,482,325 to 2,693,322 images, resulting in an increase of 8.5% in the overall size of the dataset.

3.4.2. Data Variation for Temporal Stream

Instead of clipping the prominent areas of the picture center, as performed in [32], we incorporated a method of data variation in our proposed work to enhance the data variety. All frames are cropped from four corners by randomly choosing the height and width for each set, which was to make use of multiple scale representations. After resizing the clipped areas to 224 × 224 and flipping them horizontally, they were presented to the proposed model as input for training the network. This kind of augmentation method significantly increases the variability in inputs during the training process, which helps to minimize the issue of overfitting.

4. Implementation Details

4.1. Network Configuration

Figure 2 shows the structure of the spatial stream used for training the spatial network. The pre-trained model MobileNet, is used for feature extraction, followed by two dense 6 and dense 7 layers. The two-stream network is formed by combining the spatial and temporal streams. So, after enhancing the spatial stream and checking the results, we combined the two streams.

Figure 3 shows the layer structure of our ConvNet architecture used by the temporal stream. It is comparable to the network of [13] and corresponds to the CNN-M-2048 design of [33]. The rectification (ReLU) activation function is used for all hidden weight layers; max-pooling is performed across 3 × 3 spatial windows with stride 2; and local response normalization is performed using the same parameters as in [32].

Both networks are combined to formulate the proposed two-stream network, as shown in Figure 1, by averaging the softmax scores from both streams.

4.2. Training

The training method for both streams may be regarded as a modification of the method in reference [32]. The mini-batch stochastic gradient descent with momentum, set to 0.9, is used to learn the network weights. For the temporal stream, we calculate an optical flow volume I for the chosen training frame. A fixed size $224 \times 224 \times 2L$ input is randomly chopped from corners and flipped from that volume, as discussed earlier. In the spatial stream, each batch generates 128 frames cropped down to 224 × 224 by sampling 128 videos (uniformly across all activities). The learning rate is first set to 10^{-2} for both streams and then gradually reduced according to a predetermined schedule that is maintained throughout all training sets. When training both streams from scratch, models are run for 50,000 iterations, whereas the enhanced spatial stream is trained for just 2000 iterations (classification layer training + fine-tuning). In the end, softmax scores from both models are fused by averaging the softmax scores.

4.3. Testing

For the temporal stream, we selected a predetermined number of chunks/segments (5 in our case) from each video with an equal temporal gap between the chunks. We then extracted 10 frames [32] from each chunk and passed them as input for validation. The classification results throughout the whole video are then calculated by averaging the results from all chunks. Spatial stream validation is also carried out in the same manner with the difference that only a single frame from the predetermined number of chunks is passed to the network for validation.

5. Experimental Results

5.1. Evaluation Protocol

We experimented on the UCF101 dataset, which is the benchmark for action videos and is currently the largest dataset available in this field of computer vision. It contains 101 different classes, which can be split into four categories. There are 13,320 videos in the entire dataset. For evaluation, we used k-folds cross-validation, where $k = 3$. The training set contains around 9324 videos, and the test set contains around 3996 videos. The classification accuracy is used as the performance measure, and the reported metric is the mean classification accuracy across three splits. Comparisons are performed with different architectures based on the accuracy of split 1. The mean classification accuracy across three splits of data was compared with the state of the art.

5.2. Temporal Stream

We first evaluated the temporal stream architecture by providing the network with a single and dense optical flow input which was discussed previously. Performance was measured by training the architecture from scratch on UCF101 with different input configurations. First, we used a single optical flow as an input with a dropout rate of 0.5 for better generalization. Single optical flow frame did not provide impressive results, with only 71.6% accuracy, so we then used dense input by stacking five frames and observed an increase of almost 7% in the results. Further increasing the stacking $(L = 10)$ does not help significantly as compared to the previous setting, so we kept it to $L = 5$. Table 4 shows that using dense stacking of optical flow $(L > 1)$ yields good results as compared to using a single frame. This shows the importance of the temporal aspect of an activity. Figure 5 shows the accuracy curve of the temporal stream by plotting the training and testing results, whereas Figure 6 shows the loss curves of the stream.

Table 4. Temporal stream accuracy on UCF101 (split 1) with dropout rate = 0.5.

Input Configuration	Accuracy (%)
Single-frame Optical Flow (L = 1)	71.6
Multiple Optical Flow (L = 5)	78.3
Multiple Optical Flow (L = 10)	80.2

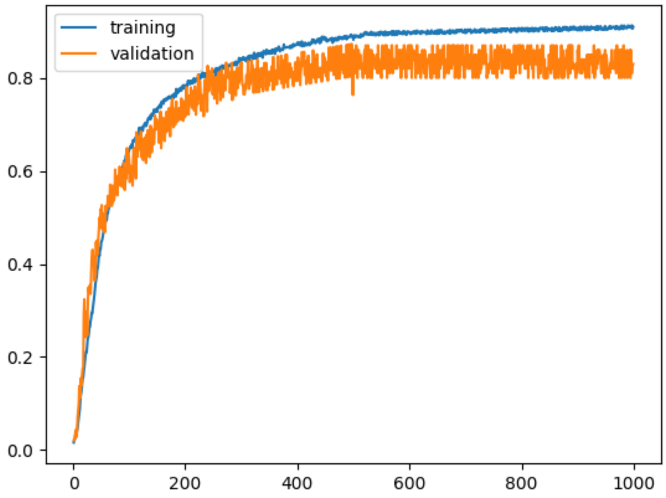

Figure 5. Temporal stream accuracy curve.

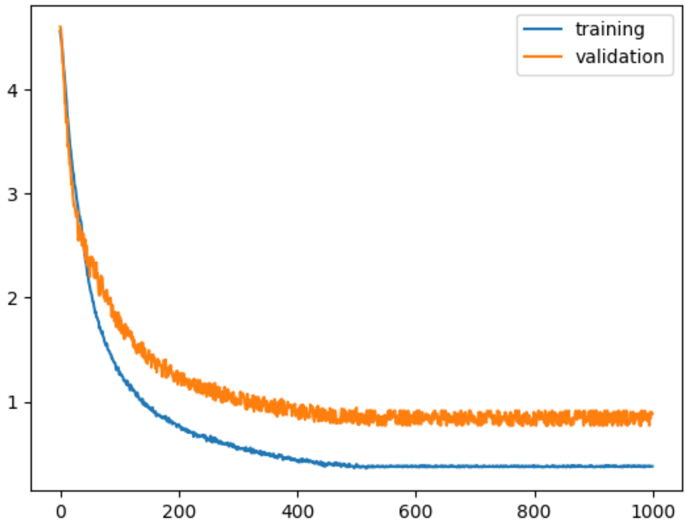

Figure 6. Temporal stream loss curve.

5.3. Spatial Stream

To assess the spatial stream, three scenarios are considered. First, we deployed the original architecture shown in Figure 3 and trained it from scratch on UCF101 with the same configuration as that of the temporal stream. This took a lot of time to train and showed poor results, with an accuracy of just 41.6%. Secondly, we adopted the enhanced spatial stream, for which we first evaluated the pre-trained models listed in Table 3 by training them on our dataset. Finally, we fine-tuned the best-performing model from Table 3 on the enhanced dataset.

Figure 7 shows the performance of enhanced spatial stream by using different pre-trained models and then fine-tuning them on UCF101. We can see that MobileNet performs best, with an accuracy of 75.2%. Moreover, Figure 7 also gives us the idea that almost every pre-trained model we utilized performed better than the original model. Fine-tuning the enhanced spatial stream on UCF101 leads to improvements because the ImageNet and UCF101 datasets are slightly different, and the feature extraction part still needs to learn the dataset through fine-tuning.

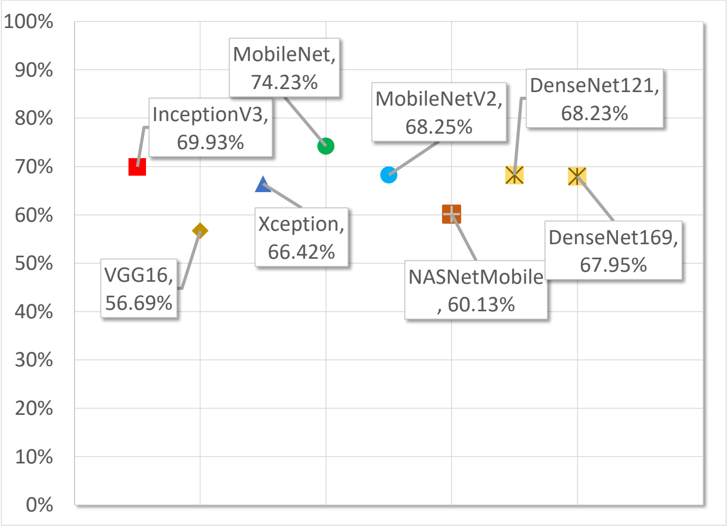

Figure 7. Pre-trained models' performance.

Based on the data given in Figure 7, we picked the MobileNet model since it has the best performance. After that, we trained classification layers and fine-tuned the entire model with the augmented dataset this time. It can be seen in Table 5 that it outperformed the standard MobileNet model, with a 1.53% increase in accuracy.

Table 5. Spatial stream performance on UCF101 (split 1).

Input Configuration	Accuracy (%)
Training from scratch	41.60
MobileNet (fine-tuning on UCF101)	75.23
MobileNet (fine-tuning on augmented UCF101)	76.70

5.3.1. Fusion

In this section, we assess the overall two-stream model. Multiple strategies have been adopted in recent years by researchers to combine the temporal and enhanced spatial streams. One possible approach is to make a stack of joint layers on top of classification layers and then train it, but this leads to overfitting, so we fused the two streams by averaging their softmax scores. The results in Table 6 show the significance of combining both streams, as the overall accuracy is 9.43% greater than the temporal stream results and 12.93% greater than the spatial stream.

Table 6. Fusion results of two-stream network on UCF101 (split 1).

Input Configuration	Accuracy (%)
Temporal Stream	80.20
Enhanced Spatial Stream	76.70
Fusion by Averaging	89.63

5.3.2. Comparison with State of the Art

Finally, we compare the overall results of our approach with the state-of-the-art methodologies by comparing the mean accuracies over three splits of UCF101. For that, the temporal stream was trained on dense optical flow images which were extracted beforehand, with a stack of L = 5 frames. The spatial stream on the other hand used pretrained MobileNet. Further improvements were achieved by fine tuning the spatial stream on the augmented data. Both streams were fused in the end by averaging their softmax scores to produce the results. We first compared the results of both streams with other state-of-the-art methods. Table 7 shows the comparisons as well as the models used by other methodologies in the motion and appearance stream. We can see that our spatial stream performed much better than the original spatial stream in [13], i.e., an increase of 6% in accuracy. For the temporal stream, we used the same model as [13].

Table 7. Comparison of appearance and motion path with other SOA models.

Method	Appearance		Motion	
	Model	Acc (%)	Model	Acc (%)
K. Simonyan [13]	AlexNet	73.00	CNN	83.70
C. Feichtenhofer [23]	VGG-16	82.60	VGG-16	86.25
C. Feichtenhofer [22]	ResNet	82.29	ResNets	87.0
L. Shi [27]	ResNeXt	85.20	ResNeXt	87.00
Proposed	MobileNet	79.00	CNN	82.60

The results in Table 8 show the overall comparison with state-of-the-art methods, and we observe that our results performed well compared to almost all of them, with an accuracy of 91.20%.

Table 8. Comparison with other SOA models (mean accuracy).

Model	Accuracy (%)
Two-Stream Network [13]	86.90
Two-Stream Network Fusion [23]	91.40
Residual Two-Stream Network [22]	91.70
Residual Frames Two-Stream Network [29]	91.80
Temporal Stream Network	82.60
Enhanced Spatial Stream Network	79.00
Proposed Two-Stream Network	**91.20**

In this research, our goal was to recreate the existing two-stream network for HAR by enhancing its spatial stream and, therefore, we only compared our method with some corresponding methods, shown in Tables 7 and 8. Our spatial stream result outperformed the original two-stream model [13]. In [22,23,29] researchers used very deep networks which require a lot of computation power and time to train them. Keeping in view the edge they have over us in terms of computational power, our model still outperformed most of them when comparing the overall accuracy over three splits, as shown in Table 8.

6. Conclusions

The key goal of our research was to develop a reliable HAR network. Various techniques were discussed in this paper to cope with the problems faced by two-stream networks including overfitting, and their effects were measured. Different configurations of inputs were also considered, and the results were compared with published research. Here, we proposed an enhanced form of the original spatial stream. Several strategies were deployed to reduce the overfitting issue posed by the insufficient datasets. Using data augmentation and transfer learning, a critical improvement in HAR has been attained. The empirical experiments have shown that the proposed architecture's results are better than the top-ranked model in terms of accuracy, with 91.2% accuracy on the UCF101 dataset.

While the currently applied methodology provides good results for the used dataset, future research into new alternatives for the proposed system may enhance precision. Using transfer learning in the temporal stream can provide good results. Furthermore, by adding more activities to the existing datasets, overfitting can be further minimized. Furthermore, the current study offers foundational principles for future researchers to investigate more configurations in the stated architecture, which will aid in the HAR system's high achievement level.

Author Contributions: S.K. Investigation, Writing—original draft; A.H. Supervision, Methodology, Writing—original draft; F.H. Methodology, Conceptualization; A.P. Resources, Validation; F.R. Resources, Validation; M.A. Methodology, Writing—review & editing; W.A. Conceptualization, Writing—review & editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Researchers Supporting Project number (RSPD2023R636), King Saud University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The data used in this research work is a publicly available dataset available at https://doi.org/10.48550/arXiv.1212.0402 (accessed on 10 February 2023).

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

References

1. Aggarwal, J.K.; Ryoo, M.S. Human Activity Analysis: A Review. *ACM Comput. Surv.* **2011**, *43*, 1–43. [CrossRef]
2. Lavee, G.; Rivlin, E.; Rudzsky, M. Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video. *IEEE Trans. Syst. Man Cybern. Part C* **2009**, *39*, 489–504. [CrossRef]
3. Batool, S.; Hassan, A.; Khattak, M.A.K.; Shahzad, A.; Farooq, M.U. IoTAuth: IoT Sensor Data Analytics for User Authentication Using Discriminative Feature Analysis. *IEEE Access* **2022**, *10*, 59115–59124. [CrossRef]
4. Keyvanpour, M.R.; Vahidian, S.; Ramezani, M. HMR-vid: A comparative analytical survey on human motion recognition in video data. *Multimed. Tools Appl.* **2020**, *79*, 31819–31863. [CrossRef]
5. Qiu, Z.; Yao, T.; Mei, T. Learning Spatio-Temporal Representation with Pseudo-3D Residual Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
6. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 221–231. [CrossRef] [PubMed]
7. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7794–7803.
8. Zhao, Y.; Xiong, Y.; Lin, D. Trajectory convolution for action recognition. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Volume 31.
9. Chaquet, J.M.; Carmona, E.J.; Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. *Comput. Vis. Image Underst.* **2013**, *117*, 633–659. [CrossRef]
10. Caruccio, L.; Polese, G.; Tortora, G.; Iannone, D. EDCAR: A knowledge representation framework to enhance automatic video surveillance. *Expert Syst. Appl.* **2019**, *131*, 190–207. [CrossRef]
11. Xiao, Z.; Jiang, J.; Ming, Z. High-Level Video Event Modeling, Recognition, and Reasoning via Petri Net. *IEEE Access* **2019**, *7*, 129376–129386. [CrossRef]
12. Zhang, J.; Shum, H.P.H.; Han, J.; Shao, L. Action Recognition from Arbitrary Views Using Transferable Dictionary Learning. *IEEE Trans. Image Process.* **2018**, *27*, 4709–4723. [CrossRef] [PubMed]
13. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 1, pp. 568–576.
14. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
15. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3D convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 11–18 December 2015.
16. Taylor, G.W.; Fergus, R.; LeCun, Y.; Bregler, C. Convolutional learning of spatio-temporal features. In *Computer Vision–ECCV 2010, Proceedings of the 11th European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010*; Springer: Berlin/Heidelberg, Germany, 2010.
17. Weimer, D.; Scholz-Reiter, B.; Shpitalni, M. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. *CIRP Ann.—Manuf. Technol.* **2016**, *65*, 417–420. [CrossRef]
18. Donahue, J.; Hendricks, L.A.; Rohrbach, M.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2625–2634. [CrossRef] [PubMed]
19. Li, H.; Chen, J.; Hu, R.; Yu, M.; Chen, H.; Xu, Z. Action recognition using visual attention with reinforcement learning. In *MultiMedia Modeling, Proceedings of the 25th International Conference, MMM 2019, Thessaloniki, Greece, 8–11 January 2019*; Springer: Berlin/Heidelberg, Germany, 2019.
20. 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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
21. 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.
22. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Spatiotemporal residual networks for video action recognition. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.
23. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
24. Han, Y.; Zhang, P.; Zhuo, T.; Huang, W.; Zhang, Y. Going deeper with two-stream ConvNets for action recognition in video surveillance. *Pattern Recognit. Lett.* **2018**, *107*, 83–90. [CrossRef]
25. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *arXiv* **2012**, arXiv:1212.0402. [CrossRef]

26. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004; Volume 3, pp. 32–36.
27. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 12018–12027.
28. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. MATNet: Motion-Attentive Transition Network for Zero-Shot Video Object Segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 8326–8338. [CrossRef] [PubMed]
29. Tao, L.; Wang, X.; Yamasak, T. Rethinking motion representation: Residual frames with 3D convnets. *IEEE Trans. Image Process.* **2021**, *30*, 9231–9244. [CrossRef] [PubMed]
30. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
31. Pan, S.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
33. Shi, X.; Lv, F.; Seng, D.; Zhang, J.; Chen, J.; Xing, B. Visualizing and understanding graph convolutional network. *Multimed. Tools Appl.* **2021**, *80*, 8355–8375. [CrossRef]

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.