An Optical-Flow Integrated Two-by-Two Stream Model for Robust Video-Based Action Recognition

Nartay Aikyn
Nazarbayev University
Astana, Kazakhstan
nartay.aikyn@nu.edu.kz

Muhammad Ahsan
Nazarbayev University
Astana, Kazakhstan
muhammad.ahsan@nu.edu.kz

Muslim Adedamola Alaran
Nazarbayev University
Astana, Kazakhstan
adedamola.alaran@nu.edu.kz

Abstract

Action recognition is a challenging problem in computer vision and has various applications such as video surveillance, human-computer interaction, and sports analysis. In this research proposal, we aim to address the problem of video-based action recognition using deep learning techniques. The proposed methodology involves extracting features from the upper and lower halves of the human body using human body localization and optical flow information. Two by Two streams of deep learning models are employed for classification. Two meta-streams are used to process different human body parts, while for each meta-stream, two sub-streams are used to handle different data modalities. The methodology addresses the issue of recognizing human behavior in various realistic video environments, and enhances the model's ability to handle erroneous visual information resulting from partial occlusion of the human body. The proposed method, which employs Two by Two streams of networks, and transfer learning of video transformer models, achieves a notable accuracy of 92.86% on the UCF sports dataset.

1. Introduction

Recognizing the temporal actions of human bodies has been an essential task in the field of computer vision in which regard many authors had proposed different approaches to capture human action properl [16]. Common techniques for representing human body movements are based on skeleton representation [15] and RGB-D representation [20] [25].Human action recognition is a technique for finding videos that came through content-based video retrieval (CBVR). It is currently a developing area in computer vision research. The simplest input, RGB-D video, simply needs RGB-D cameras to collect; nevertheless, it demands a lot of memory space. Skeletal representation is the most promising in terms of efficiency since it captures important human body parts and can match the accuracy of RGBD-based human action detection [4].

Human actions can be represented as sequences of motions and postures, making video-based action recognition is a crucial task in computer vision. With the recent advancements in deep learning, video-based action recognition has made significant progress. However, the task is still challenging due to the variability of actions, the presence of camera motion and cluttered backgrounds [6].

At its most fundamental level, this issue deals with identifying human behavior and deciphering intention and motive just from observations. Even for humans, this is a challenging task to complete, and misunderstandings are frequent. Automated methods to monitor pedestrian traffic zones and spot risky activity are essential in the field of surveillance. Several of these locations currently have surveillance cameras installed, but only human security professionals are responsible for visual comprehension and risk identification.

Nowadays, precise skeletons can be immediately generated by real-time depth sensors. The skeletal sequences are substantially smaller in size than RGB videos. Skeleton sequences can also handle busy backgrounds and light changes easier. Skeleton sequences are more desirable for action analysis because of these benefits [21].

For action recognition in video sequences, long-term global awareness of the complete action is essential [11] [26]. In order to examine the global information from whole video sequences, action recognition research has allegedly concentrated on using Conditional Random Field (CRF) [23] or Hidden Markov Model (HMM) models. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with Long-Short Term Memory (LSTM) neurons have also been used in deep networks to learn the global representations from the complete sequences for action recognition as desctibed in [7] and [18].

2. Motivation and Data Used

The motivation of this paper is from Avola et al.[1] where authors used a technique based on 2D skeleton and two-branch stacked Long Short-Term Memory (LSTM) with Recurrent Neural Networks (RNNs) cells . The 2D skeletons used in this article are formed from RGB video streams, unlike 3D skeletons, which are often produced by RGB-D cameras. This enables the use of the suggested approach in both indoor and outdoor settings.

The objective is to recognize actions from videos using machine learning classifier(s) and suitable features. It requires sustained careful attention, humans are not well suited to the action of observation. So, it is obvious that clever automated vision-based monitoring systems are needed in order to help a human user in identifying and analyzing risks.

2.1. Data Used and Datasets

The system should take RGB stream as input, and output corresponding Action labels. Some candidate datasets are as follow:

- KTH (6 action, 600 videos),
- Weizmann (10 action, 90 videos),
- UCF sports (10 action, 150 videos),
- JHMDB (21 action, 960 videos)

In this paper, we primarily focused on the UCF sports dataset.

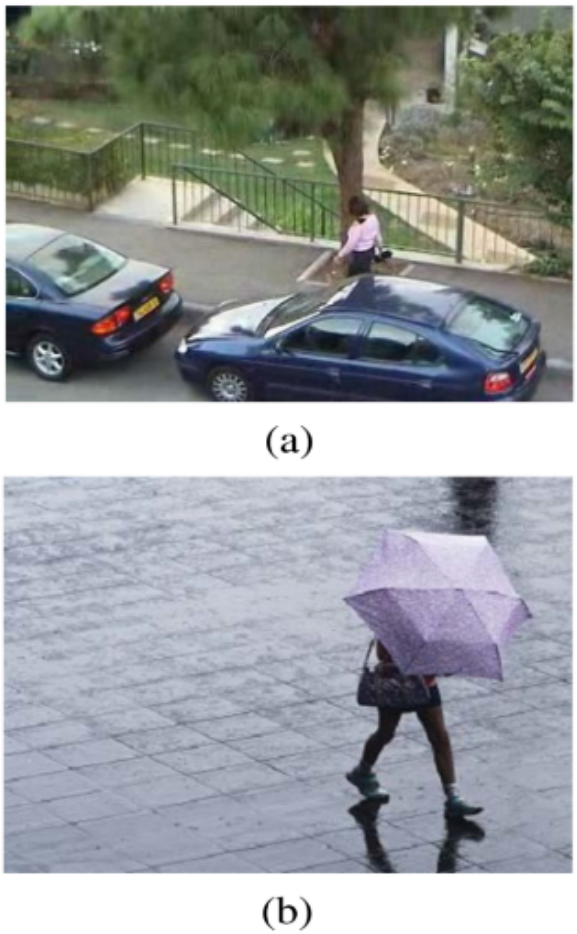

Figure 1. Partial body (subject to change, we will replace it with other images later)

3. Literature Review

In recent years, several deep learning-based approaches have been proposed for video-based action recognition. These include approaches based on 2D Convolutional Neural Networks (CNNs), 3D Convolutional Neural Networks (3D-CNNs), and Recurrent Neural Networks (RNNs). The performance of these methods has been evaluated on various benchmark datasets such as UCF101, HMDB51, and Kinetics.

Robert et al. [2] have looked into the tracking and identification of human behavior using vision-based activity recognition. A sophisticated video surveillance system has been developed to identify and keep an eye on these pedestrians. The region inside the tracked blob surrounding the pedestrian is captured on camera. These images are arranged on a timeline to make a video. Calculating the speed and form of the pedestrian motion involves the Kalman filter. Walking, running, lingering, and falling are the four different pedestrian motions. A warning signal is fired if someone enters a prohibited area, lingers there for a long time, trips over themself, or walks faster than the allowed pace. The system's drawback is that performance changes when the illumination changes. Moreover, it does not distinguish between objects moving at the same pace but using different methods, like a runner and a bike.

Ke et al. [14] recommended employing the discriminator to learn latent long-term global information and local action information for action recognition. McNally et al. [19] proposed a spatio-temporal activation model for action recognition using RGB videos. Zhao et al. [29] introduced the probabilistic model, which is hierarchical. The Bayesian framework improved the ability to identify intra-class variations in the spatial and temporal scope of occurrences. Weiyao et al. [27] proposed the bi-linear pooling and attention network-based multi-modal action recognition model. This model made use of skeletons and RGB movies.

Simonyan et al.[22] suggest a two-stream convolution neural network for RGB video, which is based on RGB images and stacked optical flows and achieves high recognition rates. According to Feichtenhofer et al.[10] the two stream model's fusion method may help to increase the

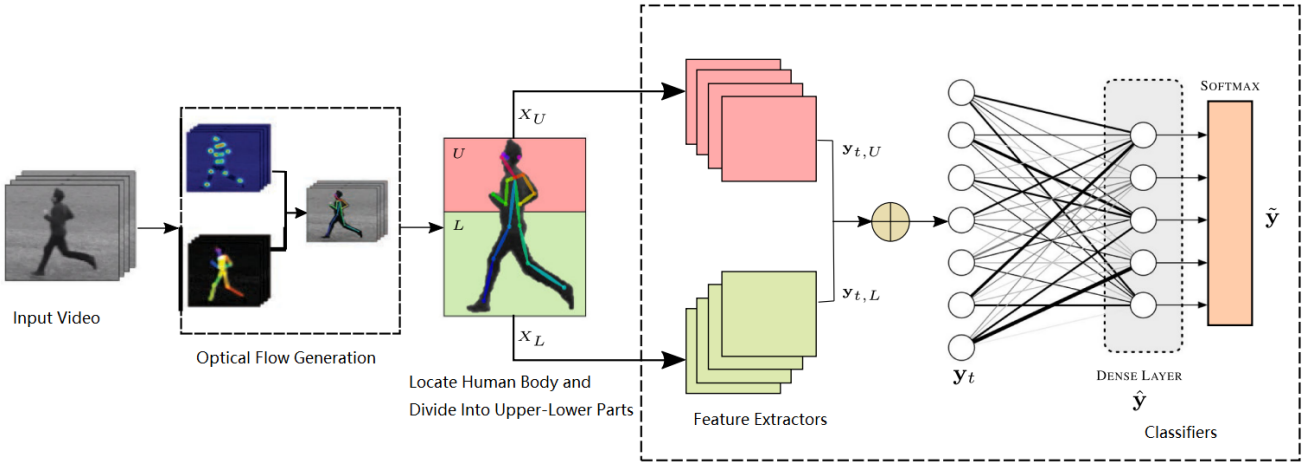

Figure 2. Architecture of Proposed Model, in courtesy of Avola et al. (subject to change, we will replace it with original graph later)[1]

recognition accuracy even further. With continuous video frames as input, Tran et al. [24] suggest the C3d model to more effectively extract video's temporal properties. Wang et al. produced TSN [26] that uses uniform sampling from the video to depict the video segments. By combining various feature levels suthors from [28] made TPN which obtained pyramid features and rich semantics.

Two-stream networks, in particular from [22], [9] employ two distinct CNNs on per-frame visual appearances and optical flows, respectively, and an average pooling operation for temporal aggregation. One of its variations, TSN [26], suggests that video clips be represented by sampling from evenly spaced segments. And in order to better capture information in the temporal dimension, TSM [17] and TRN [30] substitute the average pooling procedure with an interpretable relational module and a shift module, respectively. In order to concurrently model temporal and spatial semantics, methods [24], [13] in the second category alternatively employ 3D CNNs that stack 3D convolutions.

Avola et al.[1] suggested to use 2D skeletal data and two-branch stacked RNNs with LSTM cells to recognize human actions. The methodology uses a 3D-DenseNet-based supplementary network to handle the missing skeletal data and a C-RNN-GAN-based strategy to support those datasets with sparse data. RGB data comparison tests reveal that the suggested technique outperforms the existing literature on three benchmarks. Instead, the technique exhibits strong performance in the remaining datasets, even in the presence of noisy, or limited, data and perspective changes. Their strategy was broadly applicable in uncontrolled indoor and outdoor contexts and was comparable to methods that use 3D skeletons.

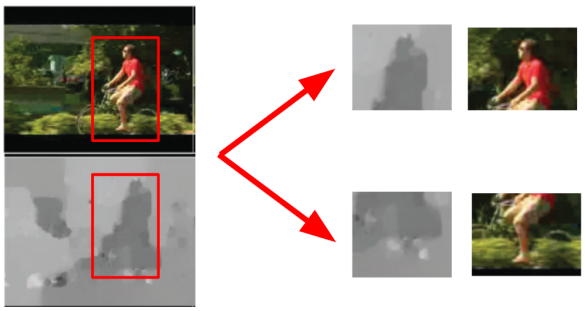

Figure 3. Example of Detecting human body, and splitting the RGB frame and motion flows

4. Proposed Methodology

Our proposed methodology entails locating the human body from the RGB video input and extracting the upper and lower halves of the human body as the major input features for each video frame. The proposed architecture is illustrated in Figure 2. We also utilize the optical flow information extracted from the original movie as additional features. To classify these two sets of data, we employ two-by-two streams of deep learning models. The classification outcomes of these streams will be combined to produce the final result. This approach enhances the model's ability to handle erroneous visual information resulting from the partial occlusion of the human body, as depicted in Figure 1.

4.1. Feature Extraction

To extract robust features from video-based action recognition, we follow a multi-step process:

1. **Image Modality Generation**: We generate optical flow images from RGB videos to capture motion information.

2. **Human Body Detection**: We use Haarcascade Full-body detection from Opencv [12] to detect the human

body within the video frames.

3. **Body Splitting**: We crop and split the RGB, optical flow, and other modalities into upper and lower body parts, as illustrated in Figure 3. By doing so, we aim to remove the background and focus on the actions performed by the human body.

4. **Data Resizing**: We resize the cropped and split data to ensure consistent dimensions across all modalities.

5. **Feature Extraction**: We employ deep learning models such as Transformers or CNNs to extract high-dimensional features from each body part. The extracted features are expected to capture action-related information and provide robust representations for classification.

It is worth noting that after each feature extraction step, we decrease the feature dimensionality (e.g., cropping removes the background and body splitting removes either upper or lower body). By doing so, we expect the model to learn more advanced representations, which are more robust and action-related. The ultimate goal is for the model to learn action-related features rather than simply spotting objects in the background.

4.2. Optical Flow

Optical flow refers to the visible movement of objects, surfaces, and boundaries in a visual setting, which arises due to the relative motion between the observer and the scene [5]. In this work, we implement Gunnar Farneback [8] algorithm for dense optical flow via the OpenCV [3] platform. Once the video frames have been read and image preprocessing has been performed, we obtain a 2-channel array that contains optical flow vectors (u,v). These vectors represent the magnitude and direction of the apparent motion. The magnitude of the vectors is used to determine the intensity of the resultant image, while the direction of the vectors is mapped to the hue value of the image. For optimal visibility, we set the strength of the hue to a maximum of 255.

4.3. Deep Learning Model

The deep learning models in our system are responsible for both deep feature extraction and classification tasks. To achieve robust recognition, we propose a two-by-two streams model. This model consists of two meta streams, each handling one of the splits of the upper and lower body. Each meta stream contains two sub-streams, one for RGB and the other for optical flow data. During processing, the features from the sub-streams are merged, and the output of the two meta streams is concatenated and fed to the classifier model. If data is missing from one stream, the counterpart meta stream can still generate a result. Otherwise, the results from all streams are combined to generate the final result, as shown in Figure 5.

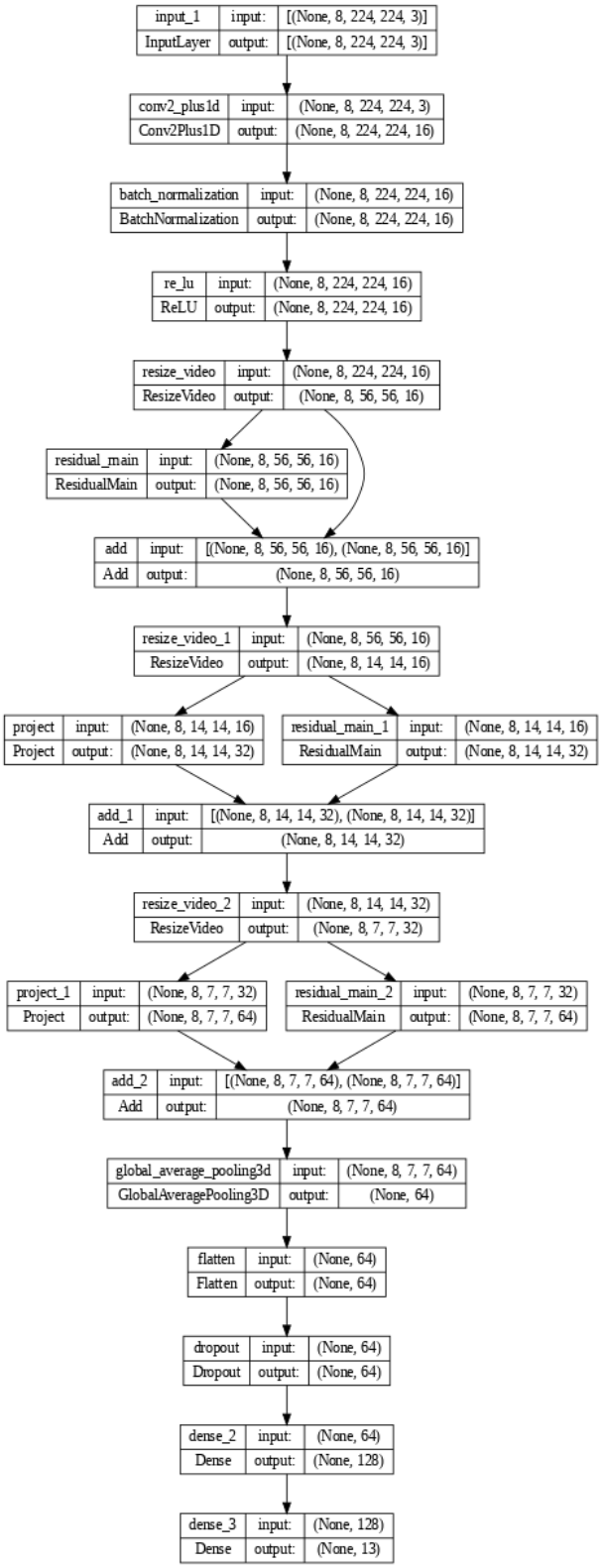

Figure 4. Baseline 3D CNN architecture

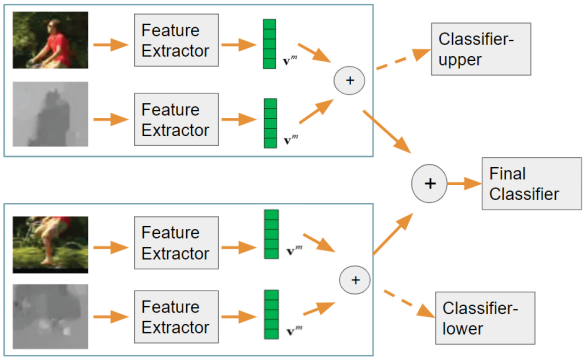

Figure 5. 2x2 stream architecture. The blue box shows the two meta streams, while the orange arrow represents sub-streams.

For preliminary results, we used a ResNet-like 3D-CNN model as the baseline, as illustrated in Figure 4. This simplified ResNet CNN model consists of 3 residual blocks, serving as the feature extractor for both RGB and optical flow data.

Furthermore, we experimented with our architecture by employing pre-trained video transformer models as the backbone. In this transfer-learning setting, the feature extractors are fixed, and only the newly attached fully connected (FC) layers are allowed to change their weights during training. The addition of the transformer model, which can capture long-term dependencies, resulted in notable performance improvements, as discussed in the next section.

5. Experimental Result

5.1. Experimental Setup

To evaluate the effectiveness of our proposed model, we conducted experiments on the UCF sports dataset, which consists of 150 videos in 13 different actions. We used the ground truth action localization labels provided by the dataset to locate the human body.

For our experiments, we randomly selected 25% of the dataset as the test set, and the remaining 75% as the training set. Despite the limited number of samples in the dataset, we did not use any data augmentation techniques in our current work. Instead, we utilized dropout layers to prevent overfitting during training.

All reported results are based on the test set, unless specified otherwise. In the next section, we discuss the results of our experiments and compare the performance of our proposed model with the baseline ResNet-like 3D-CNN model.

5.2. Preliminary Results

We present the preliminary test results of our proposed 2x2 stream design with different backbones in Table 5.1. The experiments were conducted on both full-size RGB/optical flow videos as well as upper and lower body splits of the video.

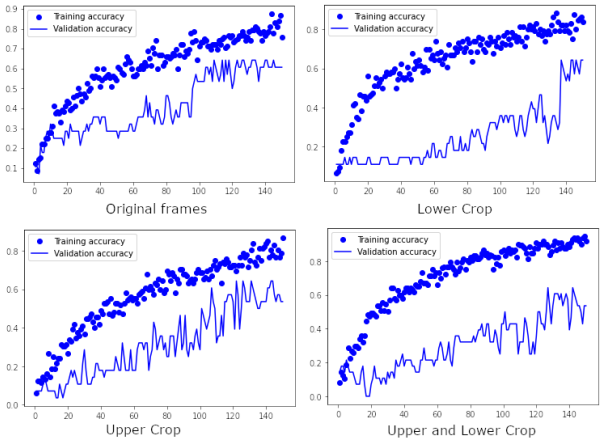

Figure 6. Training and Validation accuracy when using the original frames, lower crop only, upper crop only, and both crops

For the transfer learning backbone, we used TimeSformer, Resnet3D 18, and Swin3D Tini, all of which are pre-trained on the Kinetics 400 dataset, except for the baseline model, which was trained from scratch.

We observe that the 2x2 stream model with TimeSformer as the backbone achieved the highest accuracy of 92.86%. This indicates that our proposed 2x2 stream model with transfer learning can capture more advanced representations and achieve better performance in action recognition tasks. The results suggest that the integration of optical flow data and the splitting of the video into upper and lower body parts help to improve the performance of the action recognition task.

The training and validation accuracy over epochs for different types of input are illustrated in Figure 6. The noticeable gap between the training and validation accuracy indicates that the primary issue with the baseline model is overfitting.

5.3. Separate stream vs. stacked data

Table 5.3 presents the experimental results for using separate streams for each data modality versus stacking the data via the image channel. The results show that using separate streams for different modalities performs better than stacking the data. Specifically, the two-stream model outperforms the one-stream model for all input types. Therefore, this paper adopts the separate stream method to handle the modality.

5.4. Discussion

In the Discussion section, we analyzed the preliminary results and identified some possible reasons for the lack of significant differences among the methods with the same back-boned methods. Firstly, the UCF sports dataset may

Backbone model	Input Feature	1 stream	2 stream	2x2 stream
Baseline CNN	RGB - full	60.71%	64.29%	-
	Op.flow - full	-		
	RGB - upper	57.14%	57.14%	64.29%
	Op.flow - upper	-		
	RGB - lower	64.29%	60.71%	
	Op.flow - lower	-		
TimeSformer	RGB - full	89.29%	89.29%	-
	Op.flow - full	71.43%		
	RGB - upper	85.71%	89.29%	**92.86%**
	Op.flow - upper	64.29%		
	RGB - lower	89.29%	85.71%	
	Op.flow - lower	67.86%		
Resnet3D 18	RGB - full	67.86%	71.43%	-
	Op.flow - full	42.86%		
	RGB - upper	60.71%	71.43%	71.43%
	Op.flow - upper	46.43%		
	RGB - lower	67.86%	67.86%	
	Op.flow - lower	46.43%		
Swin3D Tiniy	RGB - full	82.14%	85.71%	-
	Op.flow - full	67.86%		
	RGB - upper	67.86%	75.00%	85.71%
	Op.flow - upper	39.29%		
	RGB - lower	89.29%	85.71%	
	Op.flow - lower	60.71%		

Table 1. Experimental results for different backbones with different inputs

Input	Two stream	One stream
RGB + Op.flow - full	**64.29%**	60.71%
RGB + Op.flow - upper	**57.14%**	53.57%
RGB + Op.flow - lower	**60.71%**	57.14%

Table 2. Experimental results for Separate stream (two streams for two data modalities) vs. stacked data (one stream for two data modalities, data are stacked via image channel)

not be the best fit for our proposed method since recognizing complex sport actions requires an understanding of the relationship between the upper and lower body, and there is no partial occlusion of the human body. Secondly, recognizing a sport action may necessitate the inclusion of object information from the background, but our design intentionally cropped out the background. Thirdly, overfitting to the small dataset size may have occurred, which could limit the model's generalization ability.

To address these issues, several steps can be taken. Firstly, we can test our method on different datasets to determine its effectiveness across various scenarios. Secondly, we can apply regularization techniques such as data augmentation to reduce overfitting and improve the model's generalization performance. We can also try a smaller model architecture to avoid overfitting.

In our experiments, we found that RGB performed better than optical flow when used alone. However, we observed that combining optical flow with RGB improved the accuracy of the model. Furthermore, we noticed that the lower part of the body usually yielded better results than the upper part. In terms of deep learning models, we found that transformer models performed better than CNNs. In addition, our experimental results show that using separate streams for different modalities performs better than stacking the data.

In summary, our preliminary results suggest that our proposed methodology can be improved by testing on different datasets, applying regularization techniques, and experimenting with different model designs.

6. Conclusion

In conclusion, our proposed methodology for action recognition using two-stream deep learning models based on the upper and lower parts of the human body, and their corresponding optical flow features, has shown promising results. Our preliminary experiments have shown an accuracy of 92.86 % on the UCF sports dataset using a 2x2 streamed method with pretrained transformers as backbone. However, our results have highlighted some limitations of

our method, such as overfitting and the need for a larger dataset with more complex action sequences.

In future research, we plan to apply regularization techniques, test our method on different datasets, and try smaller model designs. We also plan to explore other deep learning models, tune the hyperparameters, and train for longer periods to achieve better results and compare with state-of-the-art methods. Overall, our proposed methodology has the potential to improve the accuracy of action recognition, especially in scenarios with partial occlusion of the human body. we also intend to solve the overfitting problem that was discovered in our preliminary studies by using regularization approaches, such as dropout and weight decay. Additionally, we intend to investigate the use of smaller model designs that would lower the risk of overfitting and possibly enhance the generalizability of our methodology. In order to further enhance the performance of our technique, we also intend to research the utilization of attention mechanisms and the possibility of integrating it with other modalities, such as audio and depth, to provide even better outcomes in difficult situations.

References

[1] D. Avola, M. Cascio, L. Cinque, G. L. Foresti, C. Massaroni, and E. Rodola. 2-d skeleton-based action recognition via two-branch stacked lstm-rnns. *IEEE Transactions on Multimedia*, 22(10):2481–2496, 2019.

[2] R. Bodor, B. Jackson, and N. Papanikolopoulos. Vision-based human tracking and activity recognition. In *Proc. of the 11th Mediterranean Conf. on Control and Automation*, volume 1, pages 1–6. Citeseer, 2003.

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[4] J. Cha, M. Saqlain, D. Kim, S. Lee, S. Lee, and S. Baek. Learning 3d skeletal representation from transformer for action recognition. *IEEE Access*, 10:67541–67550, 2022.

[5] W. Chen, Z. Wang, Q. Wu, J. Liang, and Z. Chai. Implementing dense optical flow computation on a heterogeneous fpga soc in c. 13(3), 2016.

[6] C. J. Dhamsania and T. V. Ratanpara. A survey on human action recognition from videos. In *2016 online international conference on green engineering and technologies (IC-GET)*, pages 1–5. IEEE, 2016.

[7] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[8] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003.

[9] C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4768–4777, 2017.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1933–1941, 2016.

[11] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars. Modeling video evolution for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5378–5387, 2015.

[12] M. C.-S. Hannes Kruppa and B. Schiele. Fast and robust face finding via local context.

[13] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[14] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid. Learning latent global network for skeleton-based action prediction. *IEEE Transactions on Image Processing*, 29:959–970, 2019.

[15] J. Li, X. Xie, Q. Pan, Y. Cao, Z. Zhao, and G. Shi. Sgm-net: Skeleton-guided multimodal network for action recognition. *Pattern Recognition*, 104:107356, 2020.

[16] X. Li, Y. Hou, P. Wang, Z. Gao, M. Xu, and W. Li. Trear: Transformer-based rgb-d egocentric action recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(1):246–252, 2021.

[17] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019.

[18] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1647–1656, 2017.

[19] W. McNally, K. Vats, A. Wong, and J. McPhee. Evopose2d: Pushing the boundaries of 2d human pose estimation using accelerated neuroevolution with weight transfer. *IEEE Access*, 9:139403–139414, 2021.

[20] J. Munro and D. Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 122–132, 2020.

[21] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[22] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[23] Y. Song, L.-P. Morency, and R. Davis. Action recognition by hierarchical sequence summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3562–3569, 2013.

[24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[25] H. Wang, Z. Song, W. Li, and P. Wang. A hybrid network for large-scale action recognition from rgb and depth modalities. *Sensors*, 20(11):3305, 2020.

[26] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[27] X. Weiyao, W. Muqing, Z. Min, and X. Ting. Fusion of skeleton and rgb features for rgb-d human action recognition. *IEEE Sensors Journal*, 21(17):19157–19164, 2021.

[28] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020.

[29] R. Zhao, W. Xu, H. Su, and Q. Ji. Bayesian hierarchical dynamic model for human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7733–7742, 2019.

[30] B. Zhou, A. Andonian, A. Oliva, and A. Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018.