Efficient Action Recognition for Drones: A Comparative Study of Lightweight and Traditional Models

Mohammed El Amine Mokhtari, Matei Mancas, Virginie Vandenbulcke,
Elias Ennadifi, Sohaib Laraba, Mohamed Tazir, Bernard Gosselin

University of UMONS - Numerdiart, Boulevard Dolez, 31 – 7000 Mons - Belgium

Abstract: Action recognition is a critical task for unmanned aerial vehicles (UAVs), or drones, in various applications such as surveillance or search and rescue. However, deploying accurate action recognition models on drones is challenging due to the limited processing power and memory capacity of the on-board computing devices. In this study, we propose a lightweight model for action recognition that can be deployed on drones. We compare our proposed model with traditional models such as LSTMs and Temporal segment networks in terms of accuracy, speed, and model size. Our proposed model achieves competitive accuracy while being 20 times smaller than traditional models. These results demonstrate that the proposed model is a promising solution for real-time action recognition on drones with limited computational resources.

Keywords: Action recognition, computer vision, machine learning, MobileNet

1. Introduction

Recent advances in unmanned aerial vehicles (UAVs), also known as drones, have led to their increased use in various applications such as surveillance, search and rescue, and environmental monitoring. One important application of drones is in the field of action recognition, which involves automatically detecting and classifying human actions in video sequences. This technology can be used to improve situational awareness and decision making in many critical applications.

Despite the many advantages of using drones for action recognition, there are several challenges that need to be addressed. One of the main challenges is the limited processing power and memory capacity of the on-board computing devices. The models used for action recognition should be small, fast, and accurate to be suitable for deployment on drones. Even in the case of very small drones where images are only sent to a computer with no processing onboard, the action recognition obviously needs to be very fast.

In this study, we aim to address the research problem by proposing a lightweight model for action recognition that can be deployed on drones. Specifically, we investigate the performance of traditional models such as Long Short-Term Memory networks (LSTMs) and Temporal Segment Networks (TSNs) and compare them with lightweight image classification models such as 2D or 3D MobileNet.

In the following sections, we will present a comprehensive review of the state-of-the-art models for action recognition and the datasets used for evaluation. We will then compare the performance of these models on a custom dataset based on Kinetics600 in the Results and Discussion section. Finally, we will conclude our study with an analysis of the findings and provide suggestions for future research.

2. Literature Review

In this section, we provide an overview of the current state-of-the-art in action recognition. We discuss traditional models such as LSTMs [1,2,3,4] and TSNs [5], as well as more recent architectures using 3D convolutions [6,7] and lightweight models such as MobileNet. We also highlight the challenges and limitations of deploying action recognition models on drones.

2.1 Overview of action recognition

Action recognition [8,9] is the process of detecting and classifying human actions in video sequences. It has become an increasingly important technology in recent years due to its wide range of applications, such as surveillance [10], human-computer interaction [11], and video indexing and retrieval. In particular, action recognition is becoming critical for Unmanned Aerial Vehicles (UAVs), or drones, as they are increasingly used in various applications, such as search and rescue, environmental monitoring, and military operations.

In drone-based applications [12], action recognition can provide valuable information for situational awareness and decision-making. For example, in a search and rescue scenario, drones can be used to detect human actions such as waving for help or signalling for attention. In a surveillance application, drones can be used to detect suspicious activities such as trespassing or breaking and entering.

However, deploying accurate action recognition models on drones is challenging due to the limited processing power and memory capacity of the on-board computing devices. The models used for action recognition should be small, fast, and accurate to be suitable for deployment on drones. Therefore, developing lightweight models for action recognition [13,14] is crucial for successful implementation of drones in various applications.

2.2 Related Work

In recent years, significant progress has been made in the field of action recognition using deep learning techniques. Many methods have been proposed, including two-stream CNNs, 3D CNNs, and attention-based models. These methods have achieved state-of-the-art performance on various action recognition benchmarks. In this section, we will review some of the most notable works in action recognition using deep learning.

ShuffleNet: ShuffleNet [15] is a method proposed for efficient convolutional neural networks for image classification. The method uses group convolutions and a channel shuffle operation (*Figure 1*) to reduce the number of parameters and computation cost while maintaining high accuracy. ShuffleNet units are designed based on the bottleneck design principle and consist of a depthwise convolution, followed by a pointwise group convolution and a channel shuffle operation. This structure allows for efficient computation, especially on mobile devices with limited computing power. ShuffleNet has been shown to outperform other efficient CNN architectures on benchmark datasets for image classification, including MobileNet and SqueezeNet.

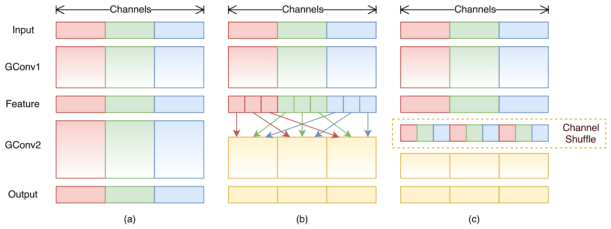


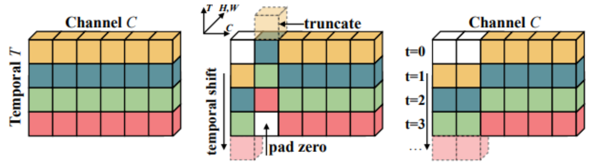
Figure 1: Channel shuffle with two stacked group convolutions. Adapted from [15].

Transformers: recently, Transformers [16] have emerged as a popular approach for modelling temporal dependencies in videos. In the context of action recognition [17], the use of Transformers has shown promising results, particularly in capturing long-term temporal dependencies and modelling complex interactions between actions. One approach is to use a spatial-temporal Transformer to model the spatial and temporal features of video frames jointly. Another approach is to use a hybrid architecture that combines a Convolutional Neural Network (CNN) with a Transformer to leverage the strengths of both models. These methods have shown to achieve state-of-the-art performance on several benchmarks, including Kinetics [25] and Something-Something [26] datasets. However, they typically require large amounts of training data and significant computational resources, which can be a limitation for practical applications.

LSTM: Long Short-Term Memory (LSTM) networks have been widely used in action recognition due to their ability to model long-term temporal dependencies. One major advantage of LSTMs [21] is their ability to capture complex temporal dynamics, making them well-suited for recognizing actions that involve multiple phases or exhibit complex temporal patterns. Additionally, LSTMs can be easily trained end-to-end, allowing for efficient learning of spatiotemporal representations from raw video data.

However, LSTMs also have some limitations. One major disadvantage is their computational complexity, which can make them challenging to use for real-time action recognition applications. Additionally, LSTMs may struggle to model the high-dimensional input data present in many video datasets, leading to overfitting or poor performance. Finally, while LSTMs can model long-term dependencies, they may struggle to capture context at larger temporal scales, making them less effective for recognizing actions that occur over very long periods of time.

TSM: The Temporal Shift Module (TSM) is a recently proposed method for efficient video understanding [20] that achieves state-of-the-art results in action recognition. The key idea of TSM is to shift the feature maps in the temporal dimension (*Figure 2*) by a learned amount, instead of computing additional convolutional layers or optical flow. This allows for significant computational savings, while still maintaining high accuracy. TSM has been shown to outperform other efficient methods like MobileNetV2 and ShuffleNet, and even achieves similar accuracy to much larger and computationally expensive models like ID3 and SlowFast (described in the next sections). However, one limitation of TSM is that it requires the input video to be uniformly sampled, which may not always be possible in practice. Overall, TSM is a promising approach for efficient video understanding, and its success highlights the potential of using simple but effective modifications to improve the efficiency of video analysis tasks.



(a) The original tensor without shift. **(b)** Offline temporal shift (bi-direction). **(c)** Online temporal shift (uni-direction).

Figure 2: TSM performs efficient temporal modelling by moving the feature map along the temporal dimension. Adapted from [20].

TSN: Temporal Segment Networks (TSN) is a widely used method for deep action recognition [5] that aims to capture the temporal dynamics of actions in videos. The method divides the video frames into several equally spaced segments and extracts features from each segment using a pre-trained 2D CNN. The features are then aggregated across segments using a consensus function, such as averaging or max pooling, to obtain a final video-level feature representation (*Figure 3*). This allows TSN to capture both spatial and temporal information in videos, making it highly effective for action recognition.TSN has been shown to outperform many other state-of-the-art methods on various benchmark datasets for action recognition. One of the key advantages of TSN is its ability to handle long-term temporal information by dividing the video frames into segments, enabling it to capture both local and global temporal dynamics. Additionally, TSN is highly scalable and can be easily adapted to different architectures and input modalities. However, one limitation of TSN is its reliance on pre-training a 2D CNN on large-scale image classification datasets, which can limit its performance on action recognition tasks with limited labelled data. Another limitation is the computational complexity of processing multiple video segments, which can make training and inference time-consuming. Despite these limitations, TSN remains a highly effective method for action recognition and has inspired many other approaches in the field.

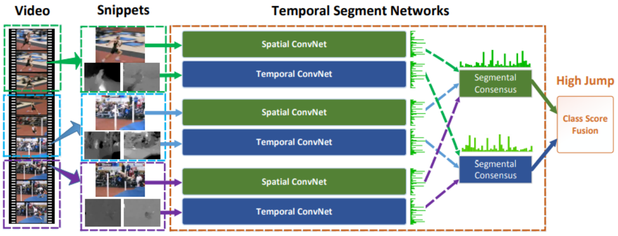


Figure 3: Architecture of TSN Model. Adapted from [5].

SlowFast: SlowFast is a widely used method for action recognition that consists of two pathways: a slow pathway and a fast pathway. The slow pathway captures the spatial information of the video by processing a subset of frames at a lower frame rate, while the fast pathway processes the remaining frames at a higher frame rate to capture the temporal information of the video. The features from the two pathways are then combined using a fusion layer.The slow pathway is typically implemented using a 2D CNN, while the fast pathway can be implemented using either a 2D CNN or optical flow. This design allows SlowFast to effectively capture both spatial and temporal information in videos, making it highly effective for action recognition.

2.3 Limitations

While the models discussed in the previous section have shown great promise in achieving high accuracy in action recognition tasks, they may not be suitable for deployment on drones due to their large size and high computational demands. In order to be used on drones, action recognition models need to be lightweight and computationally efficient, while still maintaining a high level of accuracy. Therefore, it is important to investigate and develop models that are specifically designed for deployment on drones with limited processing power and memory capacity.

2.4 Alternative Solutions

Given the limitations of traditional models for action recognition in drone applications, we turned to lightweight models as an alternative solution. Specifically, we investigated the performance of two popular models in this category: MobileNetV2 and 3D MobileNetV2. These models are based on the efficient architecture of MobileNet, which is designed to reduce the number of parameters and computations while maintaining high accuracy. MobileNetV2 and 3D MobileNetV2 have shown to be effective in various computer vision tasks, including image classification and action recognition. In this study, we evaluate the performance of these models for action recognition in drone applications and compare them with traditional models such as LSTMs and Temporal segment networks.

2D MobileNetV2: MobileNetV2 [22] is a lightweight convolutional neural network (CNN) architecture that was proposed as an efficient alternative to traditional CNNs for image classification tasks. The architecture of MobileNetV2 is based on depthwise separable convolutions, which decompose the standard convolution operation into two separate operations: a depthwise convolution that applies a single filter to each input channel, followed by a pointwise convolution that applies a 1x1 filter to combine the output channels of the depthwise convolution. This reduces the number of computations and parameters required by the network while still achieving high accuracy. In addition to depthwise separable convolutions, MobileNetV2 also employs linear bottleneck modules, which use a non-linear activation function in the middle of each convolutional layer to increase the representational power of the network. These design choices make MobileNetV2 highly efficient and suitable for deployment on devices with limited computational resources, such as mobile phones and drones. MobileNetV2 has achieved state-of-the-art results on several benchmark datasets for image classification and has been successfully applied to other computer vision tasks, including object detection and semantic segmentation.

3D MobileNetV2: 3D MobileNetV2 is a variant of MobileNetV2 that is designed for 3D convolutional neural networks (CNNs) in video analysis tasks such as action recognition. Similar to MobileNetV2 for image classification, 3D MobileNetV2 employs depthwise separable convolutions in the spatial and temporal dimensions of video data to reduce the computational cost while maintaining high accuracy. Additionally, 3D MobileNetV2 introduces a squeeze-and-excitation module, which adaptively weights the feature maps based on their importance, and a feature aggregation module, which combines the features across the spatial and temporal dimensions to capture both spatial and temporal information. These design choices make 3D MobileNetV2 highly efficient and effective for action recognition tasks in drones with limited computational resources. 3D MobileNetV2 has achieved competitive results on several benchmark datasets for action recognition and has been shown to outperform other efficient models such as MobileNetV1 and TSM.

2.5 Datasets

Action recognition models are typically trained on large-scale datasets that provide a diverse range of action classes and variations in lighting, camera angles, and actors. One of the most widely used datasets for action recognition is Kinetics [23], which is a large-scale benchmark dataset that contains over 400, 600, or 700 human action classes depending on the version of the dataset. Kinetics videos are collected from YouTube and cover a wide range of activities, such as sports, cooking, dancing, and more. The dataset is split into training, validation, and test sets, and is commonly used for pre-training and fine-tuning action recognition models. Another popular dataset for action recognition is the UCF101 [24] dataset, which contains 101 action classes and a total of 13,320 video clips. The dataset covers a variety of human activities, such as sports, cooking, and instrument playing. Like Kinetics, UCF101 is also commonly used for pre-training and fine-tuning action recognition models. Other datasets that have been used for action recognition include HMDB51, which contains 51 action classes, and Something-Something, which contains 174 action classes and focuses on fine-grained actions such as "open and close" or "take and put". Overall, these datasets provide valuable resources for training and evaluating action recognition models and have been used in many state-of-the-art approaches to the task.

Used Dataset: In this study, we created a custom dataset for efficient action recognition on drones. We selected five action classes from the Kinetics600 dataset. For each class, we randomly sampled 30 videos from the Kinetics600 dataset to ensure a diverse set of videos for each action. The dataset includes a total of 150 videos.Our approach of selecting a subset of the Kinetics600 dataset allows us to reduce the computational requirements for training and testing the action recognition model, while still maintaining diversity in the action classes. This is important for real-time action recognition on drones with limited computational resources.

3. Results and Discussion

In our study, we used a custom dataset based on Kinetics600. We randomly selected 5 action classes from the dataset and chose 30 videos for each class, resulting in a total of 150 videos. In addition, we created a sixth class called "other" that contained 30 videos of actions not included in the original 5 classes, to help the model better generalize to unseen actions.

We used pre-trained checkpoints for TSN, TSM, and SlowFast, while MobileNetV2 2D and 3D were trained from scratch using the custom dataset. We fine-tuned the pre-trained checkpoints on our custom dataset for TSN, TSM, and SlowFast.

Regarding the small number of training images, we found that MobileNetV2 3D was able to achieve high accuracy despite the limited training data. This is likely due to the efficient architecture of the model, which enables it to learn effective spatiotemporal representations with relatively few parameters. Additionally, the use of data augmentation techniques such as random cropping and flipping helped to further improve the generalization performance of the model.

	TSN	TSM	SlowF	Mb2D	Mb3D
Acc	83%	82%	85%	71%	**94%**
Ckp	181MB	180MB	200MB	8.8MB	**8.9MB**
Fps	25	25	32	20	**30**

Table 1: Results in terms of accuracy and size

The accuracy achieved by TSN and TSM were 83% and 82%, respectively, while SlowFast achieved 85%. These models were designed specifically for action recognition and are widely used in state-of-the-art research. However, their model sizes were relatively large, ranging from 180 MB to 200 MB, which may be a challenge for deployment on drones with limited computational resources.In contrast, MobileNetV2 2D achieved 71% accuracy with a much smaller model size of 8.8 MB. This model is a lightweight image classification model that is commonly used for mobile and embedded devices. However, its accuracy is lower than the other models tested in this study, which suggests that it may not be suitable for more complex action recognition tasks.Finally, MobileNetV2 3D achieved the highest accuracy of 94% with a model size of only 8.9 MB. This model is a variant of MobileNetV2 that is specifically designed for action recognition using 3D convolutions. The high accuracy achieved by MobileNetV2 3D is particularly impressive given its small model size, which makes it well-suited for deployment on drones with limited computational resources.While we did not test traditional models like LSTMs and Transformers in our study due to their large size and high computational requirements, it's important to acknowledge their potential in action recognition. However, the practical challenges of implementing them on drones make it necessary to focus on lightweight models that are specifically designed for drone-based action recognition, like MobileNetV2 3D.Our results demonstrate that lightweight models such as MobileNetV2 3D are a promising solution for efficient action recognition on drones. These models can achieve high accuracy while also being small enough to be easily implemented on the on-board computing devices of drones. Future work could focus on extending our custom dataset to include additional action classes and testing the performance of other lightweight models for action recognition on drones.

4. Acknowledgement

The authors acknowledge Region Wallonne (Belgium) for the research funding (DroneSecurity project), TRAIL/ARIAC Belgian AI project and the company Sky-Hero for their collaboration.

5. Conclusion

In this study, we investigated the problem of efficient action recognition for drones and compared the performance of various state-of-the-art models. Our results show that lightweight models such as MobileNetV2 3D are a promising solution for efficient action recognition on drones. MobileNetV2 3D achieved the highest accuracy of 94% while also having a small model size of only 8.97 MB, making it well-suited for deployment on drones with limited computational resources. We also evaluated the performance of other models, including TSN, TSM, and SlowFast, and found that they achieved high accuracy, but their larger model sizes may be a challenge for deployment on drones. MobileNetV2 2D achieved a small model size, but its accuracy was lower than the other models tested.

Our experiments highlight the importance of developing lightweight models for action recognition on drones, as the limited processing power and memory capacity of on-board computing devices present significant challenges. The use of efficient models can enable real-time action recognition on drones and improve situational awareness in various applications such as surveillance, search and rescue, and environmental monitoring.Future work could focus on extending our custom dataset to include additional action classes and testing the performance of other lightweight models for action recognition on drones. Additionally, we can explore techniques such as transfer learning and data augmentation to improve the accuracy of the models further.In conclusion, our study demonstrates the feasibility of using lightweight models for efficient action recognition on drones, and highlights the potential for further advancements in this area to enable effective drone-based action recognition in real-world scenarios.

6. References

[1] Ullah, Amin, et al.: "*Action Recognition in Video Sequences Using Deep Bi-directional LSTM with CNN Features*", IEEE Access, Volume 6, Issue not available, IEEE, 2017.

[2] Li, Chuankun, et al.: "*Skeleton-based Action Recognition Using LSTM and CNN*", 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, 2017.

[3] Liu, Jun, et al.: "*Spatio-temporal LSTM with Trust Gates for 3D Human Action Recognition*", Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 2016.

[4] Gammulle, Harshala, et al.: "*Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition*", 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, 2017.

[5] Wang, Limin, et al: "*Temporal segment networks: Towards good practices for deep action recognition*", European conference on computer vision. Springer, Cham, Amsterdam, The Netherlands, 2016.

[6] Huo, Yuqi, et al.: "*Mobile Video Action Recognition*", arXiv preprint, arXiv:1908.10155, 2019.

[7] Ji, Shuiwang, et al.: "*3D Convolutional Neural Networks for Human Action Recognition*", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 35, Issue 1, IEEE, 2012.

[8] Jhuang, Hueihan, et al.: "*Towards Understanding Action Recognition*", Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 2013.

[9] Herath, Samitha, Mehrtash Harandi, and Fatih Porikli: "*Going Deeper into Action Recognition: A Survey*", Image and Vision Computing, Volume 60, Elsevier, 2017.

[10] Ullah, Amin, et al.: "*Action Recognition Using Optimized Deep Autoencoder and CNN for Surveillance Data Streams of Non-stationary Environments*", Future Generation Computer Systems, Volume 96, Issue not available, Elsevier, 2019.

[11] Lazar, Jonathan, Jinjuan Heidi Feng, and Harry Hochheiser: "*Research Methods in Human-Computer Interaction*", Morgan Kaufmann, 2017.

[12] Sultani, Waqas, and Mubarak Shah: "*Human Action Recognition in Drone Videos Using a Few Aerial Training Examples*", Computer Vision and Image Understanding, Volume 206, Elsevier, 2021.

[13] Cheng, Ke, et al.: "*Extremely Lightweight Skeleton-based Action Recognition with ShiftGCN++*", IEEE Transactions on Image Processing, Volume 30, Issue not available, IEEE, 2021.

[14] Kozlov, Alexander, Vadim Andronov, and Yana Gritsenko: "*Lightweight Network Architecture for Real-time Action Recognition*", Proceedings of the 35th Annual ACM Symposium on Applied Computing, Brno, Czech Republic, 2020

[15] Kopuklu, Okan, et al: "*Resource efficient 3d convolutional neural networks*", Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, South Korea, 2019.

[16] Dosovitskiy, Alexey, et al.: "*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*", arXiv preprint, arXiv:2010.11929, 2020.

[17] Ulhaq, Anwaar, et al: "*Vision Transformers for Action Recognition: A Survey*", arXiv preprint arXiv:2209.05700, 2022.

[18] Koot, Raivo, and Haiping Lu: "*Videolightformer: Lightweight action recognition using transformers*", arXiv preprint arXiv:2107.00451, 2021.

[19] Cha, Junuk, et al.: "*Learning 3D Skeletal Representation from Transformer for Action Recognition*", IEEE Access, Volume 10, Issue not available, IEEE, 2022.

[20] Lin, Ji, Chuang Gan, and Song Han, "*Tsm: Temporal shift module for efficient video understanding*", Proceedings of the IEEE/CVF international conference on computer vision, Long Beach, California 2019.

[21] Greff, Klaus, et al.: "*LSTM: A Search Space Odyssey*", IEEE Transactions on Neural Networks and Learning Systems, Volume 28, Issue 10, IEEE, 2016.

[22] Sandler, Mark, et al: "*Mobilenetv2: Inverted residuals and linear bottlenecks*", Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, Utah, 2018.

[23] Kay, Will, et al.: "*The Kinetics Human Action Video Dataset*", arXiv preprint, arXiv:1705.06950, 2017.

[24] Soomro, Khurram, Amir Roshan Zamir, and Mubarak Shah: "*UCF101: A Dataset of 101 Human Actions*

Classes from Videos in the Wild", arXiv preprint, arXiv:1212.0402, 2012.

[25] DeepMind: "Kinetics Dataset", [Online]. Available: https://www.deepmind.com/open-source/kinetics

[26] Qualcomm Developer Network: "Something-Something Dataset", [Online]. Available: https://developer.qualcomm.com/software/ai-datasets/something-something

7. Glossary

TSN:	Temporal Segment Networks
TSM:	Temporal Shift Module
Acc:	Accuracy
Ckp:	Checkpoint
SlowF:	SlowFast
Mb2D:	MobileNetV2 2D
Mb3D:	MobileNetV2 3D
Fps:	Frame per Second