Action Recognition in Australian Rules Football through Deep Learning

Stephen Kong Luan, Hongwei Yin, and Richard Sinnott

School of Computing and Information Systems, University of Melbourne, Australia
rsinnott@unimelb.edu.au

Abstract. Understanding player's actions and activities in sports is crucial to analyze player and team performance. Within Australian Rules football, such data is typically captured manually by multiple (paid) spectators working for sports data analytics companies. This data is augmented with data from GPS tracking devices in player clothing. This paper focuses on exploring the feasibility of action recognition in Australian rules football through deep learning and use of 3-dimensional Convolutional Neural Networks (3D CNNs). We identify several key actions that players perform: kick, pass, mark and contested mark, as well as non-action events such as images of the crowd or players running with the ball. We explore various state-of-the-art deep learning architectures and developed a custom data set containing over 500 video clips targeted specifically to Australian rules football. We fine-tune a variety of models and achieve a top-1 accuracy of 77.45% using R2 + 1D ResNet-152. We also consider team and player identification and tracking using You Only Look Once (YOLO) and Simple Online and Realtime Tracking with a deep association metric (DeepSORT) algorithms. To the best of our knowledge, this is the first paper to address the topic of action recognition in Australian rules football.

Keywords: Action recognition · 3D CNN · Australian rules football.

1 Introduction

Action recognition has been explored by many researchers over the past decade. The typical objective is to detect and recognize human actions in a range of environments and scenarios. Action recognition, unlike object detection, needs to consider both spatial and temporal information in order to make classifications. In this paper we focus on using 3-dimensional Convolutional Neural Networks (3D CNNs) to achieve action recognition for players in Australian rules football.

Australian rules football, commonly referred to as "footy" in Australia, is a popular contact sport played between two 18-player teams on a large oval. The premier league is the Australian Football League (AFL). The ultimate aim is to kick the ball between 4 goal posts for a score (6 points if the ball goes through the middle two posts) or a minor score (1 point if the ball goes through the one of the inner/outer posts). This is achieved by players doing a range of actions

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

to move the ball across the pitch. These include kicking, passing (punching the ball), catching, running (up to 15 metres whilst carrying the ball) and tackling.

The understanding of player actions and player movements in sports are crucial to analyse player and team performances. Counting the number of effective actions that take place during a match is key to this. This paper focuses on development of a machine learning application that is able to detect and recognize player actions through the use of deep artificial neural networks.

2 Literature review

Prior to deep learning, approaches based on hand-engineered features for computer vision tasks were the primary method used for action recognition. Improved Dense Trajectories (IDT) [26] is representative of such approaches. This achieved good accuracy and robustness, however hand engineering features is limited. Deep learning architectures based on CNNs have achieved unparalleled performance in the field of computer vision. Deep Video developed by Karpathy et al. [17] was one of the first approaches to apply 2D CNNs for action recognition tasks. This used pre-trained 2D CNNs applied to every frame of the video and fusion techniques to learn spatio-temporal relationships. However, its performance on the UCF-101 data set [20] was worse than IDT, indicating that 2D CNNs alone are sub-optimal for action recognition tasks since they do not adequately capture spatio-temporal information.

Two-stream networks such as [19] add a stream of optical flow information [11] as a representation of motion besides the conventional RGB stream. The approach used two parallel streams that were combined with fusion based techniques. This approach was based on 2D CNNs and achieved similar results to IDT. This approach sparked a series of research efforts focused on improving two-stream networks. This included works focused on improvement in fusion [6], and use of recurrent neural networks including Long Short-Term Memory (LSTM) [4,15]. Other methods include Temporal Segment Networks (TSN) [27] capable of understanding long range video content by splitting a video into consecutive temporal segments, and multi-stream networks that consider other contextual information such as human poses, objects and audio in video. The framework of two-stream networks was widely adopted by many researchers, however, a major limitation of two-stream networks was that optical flows require pre-processing and hence require considerable hand-engineering of features. Generating optical flows for videos can be both computationally and storage demanding. This also affected the scale of training data sets required.

3D CNNs can be thought of as a natural way to understand video content. Since video is a series of consecutive frames of images, a 3-dimensional convolutional filter can be applied to both the spatial and temporal domain. Initial research was explored by [13] in 2012, then in 2015 by Tran et al. [22] who proposed a 3D neural network architecture called C3D using $3 \times 3 \times 3$ convolutional kernels. They demonstrated that 3D CNNs were better at learning spatio-temporal features than 2D CNNs. The introduction of C3D marked the

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

start of a new chapter in action recognition. 3D CNNs were shown to be suited to extracting and learning spatio-temporal features from video - a core demand for real-time action recognition. However C3D were difficult to train with the training process usually taking weeks on large data sets, due to the cost incurred in training with an overwhelming number of parameters in the full 3D architecture.

In 2017, Carreira et al. [2] proposed Inflated 3D ConvNets (I3D), which utilized transfer learning and outperformed all other models using the UCF-101 data set. I3D avoided the necessity for training from scratch by using some well-developed 2D CNN architectures that were pre-trained on large scale data sets such as the ImageNet [3]. I3D added an additional temporal dimension, where the model weights were used. The proposed I3D model was implemented for both the two-stream and single stream approach. Weights from an Inception-V1 model [12] pre-trained on ImageNet were used and trained on the Kinetics-400 data set [18]. This was subsequently fine-tuned on the UCF-101 data set to achieve a top-1 accuracy of 95.1% with RGB stream only. I3D demonstrated that 3D CNNs could benefit from the weights of 2D CNNs pre-trained on large scale data. This has since become a popular strategy adopted by many that has sparked a model benchmark standard based on the Kinetics-400 data.

Tran et al. [24] proposed the R2+1D architecture in 2018. This focused on factorizing spatio-temporal 3D convolutions into 2D spatial convolutional blocks and 1D temporal convolutional blocks. This decomposition provided simplicity for model optimization and improved the efficiency of training, while also enhancing the model's ability to represent complex functions by increasing the number of non-linearities through adding Rectified Linear Unit activation functions (ReLU) between the 2D and 1D blocks. The R2+1D model used the Deep Residual Network (ResNet) [10] architecture as the backbone and achieved similar performances to I3D on data sets such as Kinetics-400 and UCF-101.

Non-local blocks proposed by Wang et al. in 2018 [28] introduced a new form of operational building block that was able to capture long range temporal features similar to the self attention mechanism [25]. This was compatible to most architectures with minimal effort. The authors implemented their model by adding non-local blocks into the I3D architecture and achieved consistent improvement of performance over the original model using several data sets. In 2019 Tran et al. [23] proposed Channel Separated Networks (CSN) which focused on factorizing 3D CNNs by separating channel-wide interactions and spatio-temporal interactions by introducing regularization measures into the architecture to improve the overall accuracy. CSN are regarded as an efficient and lightweight architecture, where the model interaction-reduced channel-separated network (ir-CSN) using a ResNet-152 backbone reported a top-1 accuracy of 79.2% on the Kinetics-400 data set.

Feichtenhofer et al. [5] proposed the SlowFast networks framework. This consisted of a fast and a slow stream. The fast stream was used for extracting temporal motion features at a high frame rate, whilst the slow stream was used for extracting spatial features at a low frame rate. These two streams were later fused by lateral connections, commonly seen in two-stream network models. However,

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

the architecture of SlowFast networks was fundamentally different to two-stream networks since it was based on streams of different temporal frame rates and not two separate streams of spatial and temporal features. The SlowFast network provided a generic and efficient framework that could be applied to various spatio-temporal architectures. Furthermore, the fast stream was lightweight as the channel capacity was greatly reduced by only focusing on temporal features. The proposed network used ResNet architecture as the backbone and achieved a better performance than I3D and R2+1D on the Kinetics-400 data set.

Another similar framework was the Temporal Pyramid Network (TPN) proposed by Yang et al. [30]. This used a pyramid structure for processing frames at multiple feature levels to capture the variation in speed for different actions - so called visual tempos. TPN had the ability to use various 3D or 2D architectures as the backbone, where the set of hierarchical features extracted by the backbone undergoes down-sampling with a spatial module and a temporal rate module for processing features rich in both visual tempos and spatial information. These could then be aggregated by an information flow process. TPN used a ResNet-101 backbone and achieved better performance in Kinetics-400 over the SlowFast network.

3 Australian Rules Football Data Set

A well-defined and high-quality data set is crucial for action recognition tasks. This should contain enough samples for deep neural networks to extract motion patterns, and offer enough variance for different scenarios and camera positions for performance analysis. No such data set exists for AFL, hence we construct our own action recognition data set for AFL games. In this process, we referred to some well-known data sets for video content understanding including Youtube-8M [1], UCF 101 [20], Kinetics-400 [18], SoccerNet [8] and others. All the training and testing videos used here were retrieved from YouTube.

As AFL games are popular in Australia, there are more than enough videos on YouTube, including real match recordings, training session recordings, tutorial guides etc. However, manually creating and labelling data from video content (individual frames) is a challenging and time-consuming task. In order to feed enough frames and information for temporal feature extraction into deep learning models, we set the standard that each video clip should be at least 16 frames in length and it should be not a long-distance shot with low resolution of action tasks.

Players in an AFL match are highly mobile hence actions only exist for a very limited amount of time and are often interfered with by other players through tackles. As a result, actions sometime may end up in failure. This brings significant challenges to the construction process of the data set, e.g. judging the actual completeness of actions. This work focuses on recognizing the patterns and features of attempted actions, and pays less attention to whether the action has been completed or not. All action clips within the data set have a high level

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

of observable features, where the actual completeness of those actions was less of a concern.

In AFL games, some actions like marks (catching the ball kicked by a player on the same team) have a specific condition that needs to be met. According to AFL rules, a mark is only valid when a player takes control of the ball for a sufficient amount of time, in which the ball has been kicked from at least 15 meters away and does not touch the ground and has not been touched by another player. We aim to identify specific action patterns based only on the camera images and as such we do not consider the precision of whether the kicker was 15 metres away. Marks can be separated into marks and contested marks, where the latter is when multiple players attempt to catch (or knock the ball away) at the same time.

The videos from YouTube comprise many meaningless frames. We clip videos from longer videos and label them into five different classes:

(1) Kick: This class refers to the action whereby a player kicks the ball. The ball could come from various sources: the player himself holding the ball in front and dropping/kicking it, or kicking it directly off the ground.

(2) Mark: A player catches a kicked ball for sufficient time to be judged to be in control of the ball and without the ball being touched/interfered with by another player.

(3) Contested mark: Contested mark, is a special form of mark. This refers to the action that one player is trying to catch the ball and one or more opponents are either also trying to catch the ball at the same time or they are trying to punch the ball away.

(4) Pass: A player passes (punches) the ball to another player in the same team.

(5) Non-Action: This class includes players running, crowds cheering etc. This class is used to control the model performance as during the match there are many non-action frames. Without this class, the model would always try to classify video content into the previous four classes.

The details of each class in the data set are shown in Table 1, and example of each action class is shown in Figure 1. Compared to other classes, the non-action class has a relatively low number of instances in the data set. The reason is that this class spans many different scenes, and too many instances in this class would drive the attention of the model away from key features of the four key action classes.

There are several challenges when using a data set for action recognition. Some actions share the same proportion of representations. One example is marking and passing the ball. In a video clip of relatively long distance passing, if the camera does not capture the whole passing process, e.g. it starts from somewhere in the middle, the representing features of this action might be similar to a mark action, i.e. someone catches the ball. The data set could also be modified by combing two classes of mark and contested mark, as sometimes it is hard to identify a mark compared to a contested mark. If a player is trying to catch the ball, and in the background an opponent is also trying to catch the ball, but

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

Class	# of instances		
	Training	Testing	**Total**
kick	158	20	178
contested mark	94	20	114
mark	61	20	81
pass	83	21	104
non-action	66	21	87
Total	**462**	**102**	**564**

Table 1: Number of Instances of Each Class

Fig. 1: Kick, Contested Mark, Mark, Completed Pass

they do have not any physical contact at any time from one angle it may be considered as a mark. From a different camera angle, where there appears to be some degree of physical contact, it might seem more like a contested mark.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

4 Implementation and Discussion of Results

Given the complexity and diversity of the architectures mentioned above, we use the Gluon CV toolkit [9]. This provides a Pytorch model implementation, and importantly, the ability to train custom data sets. In order to fully utilize the benefit of transfer learning and to compensate for the limited amount of data, we used models pre-trained on largely scaled action recognition data sets such as the Kinetics-400, and then fine-tune those models using the custom AFL data set. The final implementation involves a slightly modified version of Gluon CV which includes a few algorithmic alterations and some minor bug fixes. The architectures and pre-trained models we used along with their specifications and top-1 accuracy on Kinetics-400 are listed below in Table 2 [31]. Here R2+1D ResNet-50 model was calculated using a $112 \times 112 \times 3 \times 16$ input data size, R2+1D ResNet-152 model was calculated using a $112 \times 112 \times 3 \times 32$ input data size, and all other models were calculated based on a $224 \times 224 \times 3 \times 32$ input data size.

Model	**Pre-trained**	**#Mil parameters**	**GFLOPS**	**Accuracy (%)**
I3D ResNet-50	ImageNet	28.863	33.275	74.87
I3D ResNet-101 Non-Local	ImageNet	61.780	66.326	75.81
I3D SlowFast ResNet-101	ImageNet	60.359	342.696	78.57
R2+1D ResNet-50	-	53.950	65.543	74.92
SlowFast-8x8 ResNet-101	-	62.827	96.794	76.95
TPN ResNet-101	-	99.705	374.048	79.70
R2+1D ResNet-152* [7]	IG65M	118.227	252.900	81.34
irCSN ResNet-152* [7]	IG65M	29.704	74.758	**83.18**

Table 2: Model Specifications

All model architectures are in 3D. I3D and I3D SlowFast models were based on inflated 2D ResNet pre-trained on ImageNet. irCSN and R2+1D ResNet-152 were pre-trained on IG-65M, and all other models were trained from scratch. All models used the Kinetics-400 data set for training [9].

The final training dataset was randomly split into training and validation data sets in the ratio of 70% and 30% respectively. A sub-clip of 16 frames was evenly sampled from each video clip at a regular interval depending on the clip's length. The number of input frames was selected as most actions happen in a short time period. If the sampled frames were less than 16, replacements would be randomly selected from the rest of the frames. The sampled frames would then be processed by standard data augmentation techniques, where it would

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

be first resized to a resolution of 340×256, while R2+1D resized the frames to 171×128,. The frames were then subject to a random resize with bi-linear interpolation and a random crop size 224×224. The crop size for R2+1D was 112×112. Following this, the frames were randomly flipped along the horizontal axis with a probability of 0.5, and normalized with means of $(0.485, 0.456, 0406)$ and standard deviations of $(0.229, 0.224, 0.225)$ with respect to each channel.

The training process used stochastic gradient descent (SGD) as the optimizer, with custom values of learning rate, momentum and weight decay, which were specific to each model. The value of learning rate plays a very important role in the model training process, where the correct learning rates will allow the algorithms to converge, whereas the wrong learning rates will result in the model not generalizing at all. Since we fine-tune pre-trained models, the initial learning rate was set much lower than the original model. The common values of the learning rate were 0.01 and 0.001, with a momentum of 0.9, a weight decay of $1e^{-5}$, and learning rate policy set to either step or cosine, depending on each model's architecture and level of complexity. Cross entropy loss was used for the model criterion with class weights taken into consideration since the training data set was imbalanced between the different classes. The number of epochs was set at 30 with an early stopping technique used to prevent over-fitting. The epoch with the lowest validation loss was saved as the best weight.

The top-1 accuracy on the testing data set for the fine-tuned models is shown in Table 3.

Model	**Accuracy (%)**
I3D ResNet-50	56.86
I3D ResNet-101 Non-Local	61.77
SlowFast-8x8 ResNet-101	69.61
TPN ResNet-101	70.59
I3D SlowFast ResNet-101	71.57
R2+1D ResNet-50	72.55
irCSN ResNet-152	74.51
R2+1D ResNet-152	**77.45**

Table 3: Top-1 Accuracy on the AFL test data set

As seen, the best performing model was the R2+1D ResNet-152 model pre-trained on the (very large) IG65M dataset. This achieved a top-1 accuracy of 77.45%. The final classification of action recognition results are shown in Table 4. As seen, the classification for marks had the lowest recall of 0.55, while contested marks had a recall of 0.85. This is possibly due to marks and contested marks being difficult to distinguish in some circumstances due to the presence of other players in the background. The classification for non-action has the lowest precision of 0.57 and the lowest f1 score of 0.65. The reason for this is that the non-action class is very broad and contains many sub-classes, such as scenes of audiences and players running and cheering. Splitting the class into multiple

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

distinct classes in the future may improve the non-action accuracy. Among all classes, the classification of kicks has the highest f1-score at 0.89, since a kick has arguably the most distinct and recognizable features.

Action	Precision	Recall	F1-score
Kick	1.00	0.80	0.89
Contested mark	0.74	0.85	0.79
Mark	0.85	0.55	0.67
Pass	0.86	0.90	0.88
Non-action	0.57	0.76	0.65

Table 4: Final Classification Results

The results for the top-1 accuracy of the AFL testing data set are generally consistent with the model performance using the Kinetics-400 dataset, however the R2+1D ResNet-50 model achieved some noteworthy improvements. The model I3D ResNet-50 performed poorly with a top-1 accuracy of 56.86%, whilst the model I3D ResNet-101 Non-Local only achieved an accuracy of 61.77%. It might be inferred that the inflated 2D ResNets (I3D) are limited in their ability to capture spatio-temporal features, while R2+1D is more capable in this regard as it utilizes the factorization of the 3D ResNet architecture. It was also found that non-local blocks may not be suitable for Australian rules football, as they are designed to capture long range temporal features. Actions in AFL are relatively fast and diverse which results in the model under-performing.

It was found that the performance of models generally depends on their backbone architecture. The complexity of the ResNet architecture is closely related to the prediction accuracy, hence it could be argued that the more complex the architecture is, the more likely the model will generalize and make the right predictions. Comparing ResNet-50 with ResNet-152, there is a significant difference in complexity and number of parameters, which could be one reason for the relatively large performance difference. Another major factor to consider is that both R2+1D ResNet-152 and irCSN used IG65M for model pre-training and hence benefit from the very largely scaled data set. It is also interesting to note that R2+1D uses a 112×112 resolution input after data augmentation, whilst the rest of the models use a 224×224 input. Despite this R2+1D is still able to produce some of the best results overall.

SlowFast and TPN networks both model visual tempos in video clips. When incorporating I3D into SlowFast network, the model I3D SlowFast ResNet-101 performed evidently better than the other I3D models, indicating that the SlowFast networks are capable at better extracting spatio-temporal features and that modelling visual tempos improves the overall model performance. However, SlowFast is a more strict framework that limits the number of frames of different streams, whilst TPN is more flexible due to its pyramid structure. As a result, TPN ResNet-101 performed slightly better than SlowFast ResNet-101.

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

There are several important limitations to the presented models. Firstly, incomplete actions will likely be classified as actions. As shown in Figure 2(a), an incomplete contested mark has been classified as a contested mark. This is due to the incomplete action sharing a lot of similar features to a completed action. The model does not always possess the ability to recognise whether the ball has been cleanly caught (or not). Secondly, the model tends to perform poorly in complex scenes and environments. From Figure 2(b), it can be seen that there are many players present in the background and a player is tackling another player who has the ball. In this case, the model mis-classifies the scenario into a pass as it is similar to the scenarios of pass in the training data set.

(a) Incomplete contested mark **(b)** Realistic complex environment

Fig. 2: Mis-classified actions

5 Team Identification and Associated Limitations

Many action events depend on distinguishing teams, e.g. a completed pass requires the ball to be passed by a player within the same team. Team identification is thus important to any Australian football model. In this work, we utilize the You-Only-Look-Once (YOLO) v5 [14] framework and the DeepSORT algorithm [29] to identify and track multiple objects at the same time. The implementation of this module inputs raw frames to be classified, filters and then keeps player location information in each frame. The DeepSORT algorithm is capable of tracking object movement across different frames, and assign unique IDs to team players.

As with many team sports, AFL players wear team jerseys with colors representing their team. In this way, audiences are able to identify (distinguish) players from the two teams. We apply color distribution extractors to images to extract the differences in player jersey colors. The distribution can then be used as an input to construct a high-dimensional features such as KMeans clustering [16] to cluster players into groups. A screenshot of the results of the application is shown in Figure 3.

The performance of this module is limited by the resolution of the video. With a higher resolution, the jersey color of players in the foreground is clear but

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

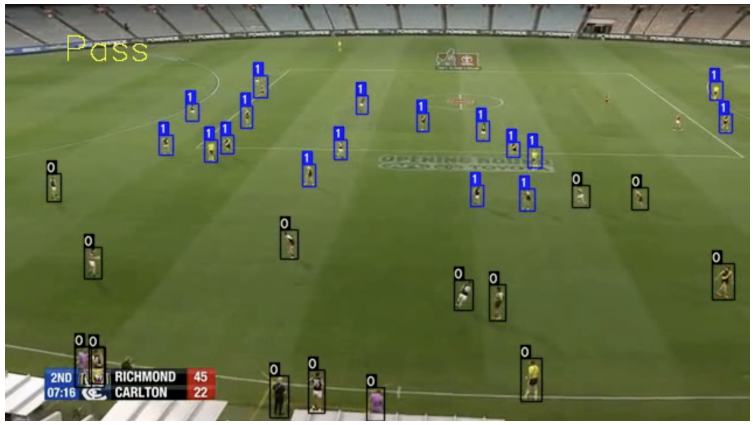

Fig. 3: Team Identification Classification Example

those in the far background is less clear. Another challenge faced is rapid camera movement and viewpoint changes. In real matches, sudden viewpoint changes from long-distance views to close-up views (and vice versa) happens continually. Ideally (from the model perspective) there would be a single camera angle - akin to what a spectator sees in a game, but this never happens in reality when games are shown on television. These continuous viewpoint changes make it challenging to track a specific player's movement. Nevertheless, the team identification is able to distinguish the teams within a few milliseconds. The performance of the system also greatly depends on teams wearing clearly identifiable jerseys. This is always the case however so does not limit the model. If players get especially muddy for example this might be an issue, but this is a rarity in Australia.

6 Conclusions and Future Work

This paper explored the feasibility of action recognition for Australian rules football using 3D CNN architectures. Various action recognition models including state-of-the-art models pre-trained on large-scale data sets were utilised. We fine-tune those models on a newly developed AFL data set, and reported a 77.45% top-1 accuracy for the best performing model R2+1D ResNet-152. A smoothing strategy allowed the algorithm to localize the frame range for actions in long video segments. We also developed a team identification solution and an action recognition application that showed both the potential and viability of applying real time end-to-end action recognition to AFL matches.

There are many future extensions to the work. The team identification framework opens up further improvements on action recognition in AFL matches for specific teams. Actions such as pass and contested mark require additional team information in order to be classified correctly. Moreover, the use of attention mechanisms in machine learning and use of transformers such as Bidirectional Encoder Representations from Transformers (BERT) [21] has the ability

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

to model contextual information with mechanisms for self attention. This could be useful in scenes that contain multiple players and allow to achieve a higher prediction accuracy.

Examples of the application of the models and the source code are available at: https://youtu.be/I7490fyuiK8 and https://github.com/stephenkl/Research-project respectively. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: YouTube-8M: A Large-Scale Video Classification Benchmark. CoRR **abs/1609.08675** (2016), `http://arxiv.org/abs/1609.08675`, arXiv: 1609.08675
2. Carreira, J., Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733. IEEE, Honolulu, HI (Jul 2017). https://doi.org/10.1109/CVPR.2017.502, `http://ieeexplore.ieee.org/document/8099985/`
3. Deng, J., Dong, W., Socher, R., Li, L.J., Kai Li, Li Fei-Fei: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE, Miami, FL (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848, `https://ieeexplore.ieee.org/document/5206848/`
4. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2625–2634. IEEE, Boston, MA, USA (Jun 2015). https://doi.org/10.1109/CVPR.2015.7298878, `http://ieeexplore.ieee.org/document/7298878/`
5. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast Networks for Video Recognition. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6201–6210. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00630, `https://ieeexplore.ieee.org/document/9008780/`
6. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional Two-Stream Network Fusion for Video Action Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1933–1941. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.213, `http://ieeexplore.ieee.org/document/7780582/`
7. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12038–12047. IEEE, Long Beach, CA, USA (Jun 2019). https://doi.org/10.1109/CVPR.2019.01232, `https://ieeexplore.ieee.org/document/8953267/`
8. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. CoRR **abs/1804.04527** (2018), `http://arxiv.org/abs/1804.04527`, arXiv: 1804.04527

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

9. Guo, J., He, H., He, T., Lausen, L., Li, M., Lin, H., Shi, X., Wang, C., Xie, J., Zha, S., Zhang, A., Zhang, H., Zhang, Z., Zhang, Z., Zheng, S., Zhu, Y.: GluonCV and GluonNLP: Deep Learning in Computer Vision and Natural Language Processing. arXiv:1907.04433 [cs, stat] (Feb 2020), http://arxiv.org/abs/1907.04433, arXiv: 1907.04433
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (Jun 2016). https://doi.org/10.1109/CVPR.2016.90, http://ieeexplore.ieee.org/document/7780459/
11. Horn, B.K., Schunck, B.G.: Determining optical flow. Artificial Intelligence **17**(1-3), 185–203 (Aug 1981). https://doi.org/10.1016/0004-3702(81)90024-2, https://linkinghub.elsevier.com/retrieve/pii/0004370281900242
12. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. pp. 448–456. ICML'15, JMLR.org, Lille, France (Jul 2015)
13. Ji, S., Xu, W., Yang, M., Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(1), 221–231 (Jan 2013). https://doi.org/10.1109/TPAMI.2012.59
14. Jocher, G., Stoken, A., Chaurasia, A., Borovec, J., NanoCode012, Taoxie, Kwon, Y., Michael, K., Changyu, L., Fang, J., V, A., Laughing, tkianai, yxNONG, Skalski, P., Hogan, A., Nadar, J., imyhxy, Mammana, L., AlexWang1900, Fati, C., Montes, D., Hajek, J., Diaconu, L., Minh, M.T., Marc, albinxavi, fatih, oleg, wanghaoyang0106: ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support (Oct 2021), https://doi.org/10.5281/zenodo.5563715
15. Joe Yue-Hei Ng, Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4694–4702. IEEE, Boston, MA, USA (Jun 2015). https://doi.org/10.1109/CVPR.2015.7299101, http://ieeexplore.ieee.org/document/7299101/
16. Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A.: An efficient k-means clustering algorithm: analysis and implementation. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7), 881–892 (2002). https://doi.org/10.1109/TPAMI.2002.1017616
17. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-Scale Video Classification with Convolutional Neural Networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1725–1732. IEEE, Columbus, OH, USA (Jun 2014). https://doi.org/10.1109/CVPR.2014.223, https://ieeexplore.ieee.org/document/6909619
18. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs] (May 2017), http://arxiv.org/abs/1705.06950, arXiv: 1705.06950
19. 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 - Volume 1. pp. 568–576. NIPS'14, MIT Press, Montreal, Canada (Dec 2014)

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47

20. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402 [cs] (Dec 2012), http://arxiv.org/abs/1212.0402, arXiv: 1212.0402
21. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: VL-BERT: Pre-training of Generic Visual-Linguistic Representations. CoRR **abs/1908.08530** (2019), http://arxiv.org/abs/1908.08530, arXiv: 1908.08530
22. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 4489–4497. IEEE, Santiago, Chile (Dec 2015). https://doi.org/10.1109/ICCV.2015.510, http://ieeexplore.ieee.org/document/7410867/
23. Tran, D., Wang, H., Feiszli, M., Torresani, L.: Video Classification With Channel-Separated Convolutional Networks. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5551–5560. IEEE, Seoul, Korea (South) (Oct 2019). https://doi.org/10.1109/ICCV.2019.00565, https://ieeexplore.ieee.org/document/9008828/
24. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A Closer Look at Spatiotemporal Convolutions for Action Recognition. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (Jun 2018). https://doi.org/10.1109/CVPR.2018.00675, iSSN: 2575-7075
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. pp. 6000–6010. NIPS'17, Curran Associates Inc., Long Beach, California, USA (Dec 2017)
26. Wang, H., Schmid, C.: Action Recognition with Improved Trajectories. In: 2013 IEEE International Conference on Computer Vision. pp. 3551–3558 (Dec 2013). https://doi.org/10.1109/ICCV.2013.441, iSSN: 2380-7504
27. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 20–36. Lecture Notes in Computer Science, Springer International Publishing, Cham (2016)
28. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local Neural Networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7794–7803. IEEE, Salt Lake City, UT, USA (Jun 2018). https://doi.org/10.1109/CVPR.2018.00813, https://ieeexplore.ieee.org/document/8578911/
29. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing (ICIP). pp. 3645–3649. IEEE, Beijing (Sep 2017). https://doi.org/10.1109/ICIP.2017.8296962, http://ieeexplore.ieee.org/document/8296962/
30. Yang, C., Xu, Y., Shi, J., Dai, B., Zhou, B.: Temporal Pyramid Network for Action Recognition. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 588–597. IEEE, Seattle, WA, USA (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00067, https://ieeexplore.ieee.org/document/9157586/
31. Zhu, Y., Li, X., Liu, C., Zolfaghari, M., Xiong, Y., Wu, C., Zhang, Z., Tighe, J., Manmatha, R., Li, M.: A Comprehensive Study of Deep Video Action Recognition. arXiv:2012.06567 [cs] (Dec 2020), http://arxiv.org/abs/2012.06567, arXiv: 2012.06567

ICCS Camera Ready Version 2022
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-08757-8_47