Individual Action and Group Activity Recognition in Soccer Videos From a Static Panoramic Camera

Beerend Gerats[1,2], Henri Bouma[2], Wouter Uijens[2], Gwenn Englebienne[1], and Luuk Spreeuwers[1]
[1]*Faculty of EEMCS, University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands*
[2]*Intelligent Imaging, TNO, Oude Waalsdorperweg 63, 2597 AK The Hague, The Netherlands*

Keywords: Action Recognition, Group Activity Recognition, Soccer Match Events, Player Snippets

Abstract: Data and statistics are key to soccer analytics and have important roles in player evaluation and fan engagement. Automatic recognition of soccer events - such as passes and corners - would ease the data gathering process, potentially opening up the market for soccer analytics at non professional clubs. Existing approaches extract events on group level only and rely on television broadcasts or recordings from multiple camera viewpoints. We propose a novel method for the recognition of individual actions and group activities in panoramic videos from a single viewpoint. Three key contributions in the proposed method are (1) player snippets as model input, (2) independent extraction of spatio-temporal features per player, and (3) feature contextualisation using zero-padding and feature suppression in graph attention networks. Our method classifies video samples in eight action and eleven activity types, and reaches accuracies above 75% for ten of these classes.

1 INTRODUCTION

Match event data gained an important role in soccer, from team and player evaluation (Pappalardo, 2019a) to increasing fan engagement (Aalbers and Van Haaren, 2018). The data often describes when, where, what and by whom events are triggered during a professional game. Competition Information Providers (CIPs) manually annotate match event data by a team of three or four persons (Pappalardo, 2019b). The procedure is expensive and time consuming considering the hundreds of annotated games in hundreds of yearly competitions. Automating parts of the annotation process would mitigate the disadvantages of manual annotation.

Over the previous decades, several methods have been proposed for the detection of soccer highlights in television broadcast videos (Giancola, 2018). However, these methods report general events rather than individual player actions. Importantly, match event data annotated by CIPs describe individual ball interactions (e.g. high pass, heading) and may be labelled with general event tags (e.g. corner, goal attempt). Automating annotation of match event data thus requires a shift towards event detection on an individual level, accompanied with detections of general activities.

Methods that simultaneously recognise individual actions and group activities have been evaluated on videos in the Volleyball Dataset (Ibrahim, 2016). However, such methods are not trivially applicable to soccer videos. In this work, we show that a state-of-the-art method in the volleyball domain, the Actor Relation Graph (ARG) (Wu, 2019), has a poor performance in the soccer domain.

Our main contribution is the proposal of a novel method for the automatic recognition of soccer events. The method works with videos that are captured by a static panoramic camera, positioned at the long side of the soccer field. We show that it is possible to recognise actions and activities that occur all over the field from this perspective only. To the best of our knowledge, it is the first method in the soccer domain that infers both individual actions and group activities simultaneously. Three key contributions in the proposed method are:

(1) the use of player snippets as model input;

(2) per-player extraction of spatio-temporal features;

(3) and feature contextualisation using zero-padding and feature suppression in graph attention networks.

This paper is structured as follows. In Section 2, we present a brief overview of methods for event detection and recognition in sport videos. In Section 3, we explain the design and implementation of the proposed method. In Section 4, a new soccer video dataset and evaluation metrics are presented. In Section 5, we present experiments and results. Conclusions are given in Section 6.

B. Gerats, H. Bouma, W. Uijens, G. Englebienne, L. Spreeuwers, "Individual action and group activity recognition in soccer videos from a static panoramic camera," Proc. Int. Conf. Pattern Recognition Applications and Methods ICPRAM, 594 - 601 (2021). https://doi.org/10.5220/0010303505940601

2 RELATED WORK

In this section, we describe methods that detect general events ("group activities") in soccer videos, and we discuss approaches for the simultaneous recognition of actions and group activities, outside the soccer domain.

2.1 Event Detection in Soccer

The aim of event detection is to detect temporal boundaries of a match event or camera shot, and to classify the isolated samples accordingly (Tavassolipour et al., 2013). Goal (attempts), corners, cards, shoots, penalties, fouls and offsides have been detected in television broadcast videos. Recent methods use a 3D-Convolutional Neural Network (CNN) (Khan, 2018) or combine a CNN and a Recurrent Neural Network (RNN) (Jiang et al., 2016) consecutively. Often, these methods rely on the detection of cinematic features, based on general ways for television production teams to record soccer events on camera (Ekin et al., 2003). For example, a goal attempt is often followed by a slow-motion shot of the event. We consider these dependencies undesirable as it limits the applicability of a model to broadcast videos only. Performances range between 82.0% (Tavassolipour et al., 2013) and 95.5% (Vanderplaetse and Dupont, 2020) multi-class accuracy (MCA), for the recognition of seven and four activity classes respectively.

Others combine recordings of twelve (Zhang, 2019) or fourteen (Tsunoda, 2017) static cameras, positioned around the field. The latter approach reaches 70.2% MCA for the recognition of three classes. We argue that multiple-camera setups are expensive in purchase and require large computational resources. Our method is designed for event recognition in videos from one static panoramic camera, which are more accessible for non professional clubs.

Soccer videos contain a majority of background pixels due to the size of the field. Nevertheless, most of the methods mentioned above classify events directly from video frames. Zhang et al. (2019) propose to detect events from latent player embeddings, created by a U-encoder on pixels in player bounding boxes. Our method creates latent player embeddings also, from normalised player snippets instead of bounding boxes.

2.2 Action and Group Activity Recognition

Three types of deep learning networks can be discovered in state-of-the-art action and group activity recognition methods: spatio-temporal, multiple-stream and hybrid networks. Spatio-temporal networks, such as a 3D-CNN (Ji, 2012), search for volumetric patterns at different scales of the input videos. The I3D CNN (Carreira and Zisserman, 2017) is a multiple-stream network that recognises actions from RGB and optical flow videos. The network appears to give better results in group activity recognition than a standard CNN (Azar, 2019). In a hybrid network, two networks are combined consecutively (Kong and Fu, 2018). The approach is popular for group activity recognition. First, a CNN extracts individual features and creates a latent embedding per group member. We will refer to this phase as *feature extraction*. Second, a different network explores inter-human relations to update the embeddings accordingly. We will refer to this phase as *feature contextualisation*. RNNs (Tsunoda, 2017) and Graph Convolutional Networks (Ibrahim and Mori, 2018) are often used for the latter phase.

Our method uses a hybrid network with I3D for feature extraction and graph attention networks (GATs) (Veličković, 2017) for feature contextualisation. We have not yet seen these networks being applied to event recognition in the soccer domain.

2.3 Actor Relation Graph as Baseline

It is difficult to compare our method with state-of-the-art in soccer event detection, because each method is evaluated with another dataset. The sets vary in event types, number of classes and input videos, while none of the methods recognise individual actions.

The Actor Relation Graph (ARG) is a hybrid network that uses an Inception-V3 CNN (Szegedy, 2016) for feature extraction and uses GATs with self-attention (Vaswani, 2017) for feature contextualisation. The method reaches state-of-the-art performance in action and group activity recognition on Volleyball Dataset videos. Because the domain is related to soccer, and an open-source implementation is available, the ARG is selected as baseline.

3 PROPOSED METHOD

We start this section with an overview of the proposed method architecture, and note where it differs from the baseline approach. Thereafter, the architecture is explained along four phases in the data pipeline: data pre-processing, feature extraction, feature contextualisation and the generation of predictions. Last, we provide implementation details.

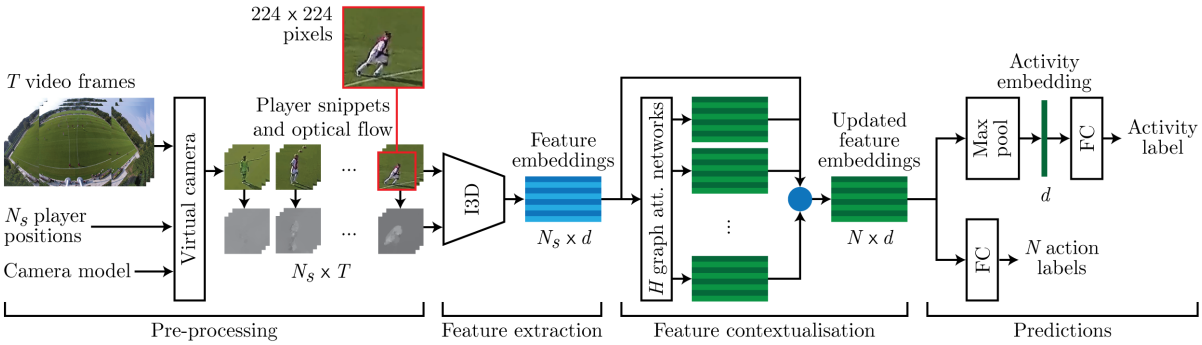


Figure 1: Architecture of the proposed method. T is the number of consecutive frames considered in one activity. N is a pre-defined number of persons to be detected on the soccer field. $N_s = \min(N_d, N)$, with N_d the number of automatically detected players by an ACF person detector. d is the dimensionality of a player embedding.

3.1 Overview of Architecture

The data pipeline of the proposed method is displayed in Figure 1. The method pre-processes the data by generating player snippets using a virtual camera algorithm. Such an algorithm synthesises frames from a raw video stream where the camera virtually zooms and rotates, while normalising for lens distortion (Matsui, 1998). The virtual camera algorithm takes raw video frames, player positions and a camera model as input. Spatio-temporal features are extracted from RGB and optical flow videos by the I3D network. The resulting player embeddings are updated with multi-head self-attention in GATs. The model outputs an action label per player and one shared activity label per sample.

The Actor Relation Graph (ARG), extracts features from the whole frame at once. The video frames are sub-sampled to 1280×720 pixels such that a feature map can be generated from the full scene. Thereafter, a standard sized feature map is cut-out for every player with RoIAlign. Only spatial features are captured by a standard CNN. Similar to the proposed method, the ARG uses GATs with self-attention to contextualise player embeddings.

3.2 Player snippets as Model Input

A soccer field is about 40 times larger than a volleyball field, meaning that the distance between a player and the camera can become much larger, players become smaller and more pixels in the video capture irrelevant background. Therefore, the proposed method uses high resolution player representations as model input in the form of *player snippets*.

To create player snippets, the positions of all players must be known in field coordinates such that a virtual camera algorithm can zoom in on these positions. We denote a field coordinate with (X, Y), where $(0, 0)$ is the centre spot of the field. An Aggregated Channel Features (ACF) person detector (Dollár, 2014) returns N_d bounding boxes for persons located within the soccer field lines. The detector is applied to all T consecutive frames considered in one activity. Player trajectories are created from bounding boxes that relate to the same player in consecutive frames, using tracking software (Bouma, 2013). When trajectories are shorter than T frames, they are linearly interpolated and extrapolated. For each activity, we select the N_s trajectories with the largest mean confidence of the person detector over T frames. Here $N_s = \min(N_d, N)$, with N(=23) the number of persons that we strive to detect (22 players and one referee).

With the camera model, pixel coordinates (x,y) can be transformed into real-world coordinates (X,Y,Z) by the projection on a virtual plane at height Z. The bottom-centre pixel in each bounding box is projected onto the ground ($Z = 0.0$ meters) to be transformed into a field coordinate. Finally, a virtual camera zooms in on position $(X_i^t, Y_i^t, 0.8\text{m})$ for player i in frame t. The zoomed image is cut-out from the original frame and resized to 224×224 pixels, the standard resolution for I3D input.

The use of player snippets gives two benefits for action and activity recognition. First, the snippets provide high resolution representations for all players, including those located far from the camera. Second, the virtual camera normalises for the rotated horizon present at most field positions in our dataset.

3.3 I3D for Feature Extraction

The proposed method uses a two-stream I3D network (Carreira and Zisserman, 2017) to create player feature embeddings. The CNN extracts spatio-temporal features, which we expect to be more informative than spatial features only, as they could describe movement over time. For example, it could distinguish

between ball movements towards and away from a player. We experiment with two temporal window sizes for I3D, being 0.32 and 0.48 seconds, corresponding to $T = 9$ (baseline) and $T = 13$ frames.

Optical flow images are generated from the player snippets using the TV-L1 algorithm (Zach et al., 2007). I3D processes RGB and optical flow videos in two separate streams. For each stream, the videos are given to the network in batches of N_s video samples, where N_s relates to the number of detected players. Then, I3D provides a d-dimensional player embedding through d logits. The result of both streams is added element-wise, in a late fusion fashion. For each activity sample, I3D returns one $N_s \times d$ feature matrix. Previously, we experimented with d=1024 (baseline) and d=256, and found the latter to be more optimal for our model. All presented results are with d=256.

3.4 Graph Attention Networks for Feature Contextualisation

State-of-the-art methods for group activity recognition have shown that attention is a useful mechanism for contextualisation (Gavrilyuk, 2020). We will follow this approach and use multi-head self-attention (Vaswani, 2017). Before feature contextualisation, we apply layer normalisation over each embedding independently and a ReLU activation thereafter.

Graph attention networks. We construct H(=64) GATs, where each graph contains N(=23) vertices representing the players and one referee. We adopt the ARG approach, where the magnitudes of attention between players depend on inter-player relationships and relative player distance (see Equation 1). A distance-mask $D \in [0,1]^{N \times N}$ prunes player-pairs from the graph when they are physically too far from each other. $D_{i,j} = 0$ if the distance between player i and j is larger than μ, and $D_{i,j} = 1$ otherwise. We use $\mu = 20.8$ meters, which is 0.2 times the width of a soccer court. It is comparable to the original ARG implementation, using 0.2 times the image width.

$$G^{(h)} = \sigma\left(\frac{D\left(W_Q^{(h)}E + b_Q^{(h)}\right)\left(W_K^{(h)}E + b_K^{(h)}\right)^T}{\sqrt{d}}\right) \quad (1)$$

with σ the softmax function, $E \in \mathbb{R}^{N \times d}$ the original player embeddings and $G^{(h)} \in \mathbb{R}^{N \times N}$ the graph attention matrix from graph h. Weight matrices $W_Q^{(h)}, W_K^{(h)} \in \mathbb{R}^{d \times d}$ and biases $b_Q^{(h)}, b_K^{(h)} \in \mathbb{R}^d$ linearly transform the player embeddings to *query* and *key* embeddings.

Graph convolution. The original player embeddings are non-linearly transformed via a graph convolution layer, as in Equation 2, also adopted from the ARG approach. Layer normalisation is applied over all embeddings in one graph.

$$\tilde{E}^{(h)} = \mathrm{ReLU}\left(\mathrm{LayerNorm}\left(G^{(h)}EW_V^{(h)}\right)\right) \quad (2)$$

with $\tilde{E}^{(h)} \in \mathbb{R}^{N \times d}$ the updated context features from graph h and weight matrix $W_V^{(h)} \in \mathbb{R}^{d \times d}$ that transforms the collection of contextualised features to *value* embeddings.

Missing player detections. Where it was possible to process N_s players per sample in the previous phases, feature contextualisation requires precisely N feature embeddings. This is required, as the model processes feature matrices with dimensionality BS $\times N \times d$, with BS=batch size. Code implementation[1] of the ARG reveals that the method duplicates $N - N_s$ player embeddings in each graph, to fill in for the missing players. The proposed method fills the gaps with d-dimensional vectors containing zeros only ("zero-padding") instead. These embeddings are thus ignored by the self-attention mechanism.

Multi-head attention. The context features from multiple graphs are combined using a fusion function. Previously, we found $H = 64$ to be optimal for our model. The authors of the ARG propose to add the contextualised embeddings element-wise. The proposed method uses a concatenation operation (see Equation 3) instead of an addition operation.

$$E' = E + \mathrm{Concat}\left(\tilde{E}^{(1)}, \tilde{E}^{(2)}, \ldots, \tilde{E}^{(H)}\right)W_O \quad (3)$$

with E' the final feature embeddings before label prediction, a residual connection to the original embeddings E and weight matrix $W_O \in \mathbb{R}^{H \times d \times d}$. The embeddings from H graphs are concatenated and linearly transformed through W_O.

Implicit bias We consider players that are interacting with the ball or that are involved in a duel as *active*. Players that are not *active* are referred to as *passive*. Soccer games contain an implicit bias that when no duel occurs, only one player is interacting with the ball ("*active*"). A model with feature contextualisation should explore this inter-player relation, and predict fewer "false positives" (*passive* players that are wrongly recognised as *active*); especially in activities with no duel where multiple players are initially recognised as *active*. The duplication padding

[1] github.com/wjchaoGit/Group-Activity-Recognition

strategy could accidentally duplicate *active* players in activities where usually only one is interacting with the ball. Zero-padding avoids this issue and is expected to strengthen the implicit bias. Besides, the network must find a way to diminish large activations that relate to *active* classes in *passive* players. We argue that adding non-negative embedding values as fusion operation does not diminish any activations. We call this *feature accumulation*. With the concatenation operation, context features can be added as well as subtracted. We call the latter *feature suppression*.

3.5 Predictions

The refined player representations E' are grouped in a $N \times d$ feature matrix. Thereafter, two output streams predict the action labels and activity label, separately. Both classifications are performed through a fully-connected (FC) layer and softmax. In the activity stream, max pooling is applied to the feature matrix beforehand, to obtain one d-dimensional vector.

3.6 Implementation Details

The I3D model is trained without feature contextualisation first, and is initiated with model parameters pre-trained on ImageNet (Deng, 2009) and Kinetics (Kay, 2017). Hyperparameters that resulted in the best performing ARG on soccer videos were selected to train all models with. The model is trained in 20 epochs, with learning rate 1×10^{-5} (5×10^{-6} starting from epoch 15), dropout probability of 0.3, no weight decay and using an Adam optimiser (Kingma and Ba, 2014). Cross entropy with class weights is used to calculate the action and activity prediction losses.

The GATs are trained afterwards, with all I3D layers frozen. Hyperparameters are unchanged, except for the number of epochs (40 instead of 20).

4 EXPERIMENTAL SETUP

4.1 The Soccer Dataset

We constructed a new dataset including soccer videos from one panoramic camera, which contains four sensors that are positioned side-by-side and together capture the whole field from one static perspective (see Figure 2). A camera model is constructed for each sensor by calibration with the field dimensions. Videos from the four sensors are combined in one video stream that has a resolution of 3840 $\times$ 2160 pixels at 25 frames per second. During 280 minutes in four soccer games, we annotated 3717 activities in eleven categories. The exact frame that an event occurred was registered, i.e. the moment of ball contact, when the ball leaves the players' foot/hands, or when the referee blows its whistle. This frame is the middle frame in temporal window T. The training set contains 2801 events from game one, game two and the first half of game three. The validation set consists of 403 events in the second half of game three. All 513 events in game four are kept in the test set.

Using an ACF person detector, we obtained bounding boxes for persons located inside the field lines, for each activity. The detected persons were annotated with an action label, in eight categories, or as an incorrect detection. The number of action and activity instances can be seen in Table 7 for the train, validation and test set. In total, 83818 samples were annotated with an individual action label. Note that 94.6% of the detections have a *passive* action label.

Figure 2: Example frame of the raw video stream.

4.2 Evaluation Metrics

In related work, performance in action and group activity recognition is often reported in multi-class accuracy (MCA). However, as the Soccer Dataset is highly unbalanced, this metric gives a too optimistic view. Therefore, we calculate a Matthew Correlation Coefficient (MCC) (Matthews, 1975) per class label. The metric is independent of class imbalance. To avoid reporting nineteen MCCs at every evaluation, we average the scores over all actions and all activities. The result is two mean MCC (MMCC) scores.

5 EXPERIMENTS AND RESULTS

Experiments are discussed along three phases: data pre-processing, feature extraction and feature contextualisation. Experiments in Sections 5.1-5.3 use the validation set for evaluation. In Section 5.4, we evaluate the ability of the ARG as baseline and the proposed method to generalise to samples in the test set.

5.1 Sub-sampling and Player Snippets

In soccer, players can be positioned far from the camera due to the large playing field. The ARG as used by Wu et al. (2019) sub-samples the video frames to a standard resolution. This causes soccer players that are farthest from the camera to be represented with very few pixels. In Figure 3 (a) it can be seen that for a soccer player located at the goal line, the body pose is difficult to recognise. The effect of using player snippets instead, with and without horizon normalisation (see Figure 3 (b) and (c)), is evaluated by training an Inception-V3 with the three different inputs.

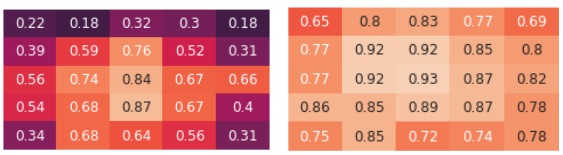
(a) Sub-sampled (b) Player snippet (c) Pl.sn. + norm.

Figure 3: Sample of a player located at the left goal line.

In Table 1 it can be seen that using player snippets increases the MMCC in action recognition from 0.163 to 0.264 and in activity recognition from 0.364 to 0.437. Normalisation of the rotated horizon further increases these scores to 0.358 (actions) and 0.616 (activities). In Figure 4, the accuracies for action recognition are provided per field region, where the top regions are farthest away from the camera. It can be seen that the use of player snippets not only improves action recognition, but also results in more uniform performance scores over all field regions. Sub-sampling gives the poorest results for players that are rotated or that are positioned far away from the camera. In the next experiments, player snippets with horizon normalisation are used as model input.

Table 1: MMCCs for sub-sampling (baseline) and player snippets, with and without horizon normalisation.

Model input	Norm.	Actions	Activities
Sub-sampling	✗	0.163	0.364
Player snippets	✗	0.264	0.437
Player snippets	✓	**0.358**	**0.616**



(a) Sub-sampling (b) Player snippets + norm.

Figure 4: Accuracies of action predictions at different parts of the soccer field. Top regions are farthest from the camera.

5.2 Inception-V3 and I3D

We compared Inception-V3 and I3D as backbone for feature extraction. The former extracts spatial features and uses temporal fusion via element-wise addition. I3D explores spatio-temporal features.

The previous experiment was carried out using Inception-V3 with a batch size of four. In Table 2, it can be seen that decreasing the batch size to one increases action recognition performance to an MMCC of 0.500. It can also be seen that spatio-temporal features are particularly important for the recognition of individual actions, increasing performance to an MMCC of 0.646. Both networks give comparable results for activity recognition. We reason that group activities develop over multiple seconds, rather than the short time period that we selected for one sample (0.32 seconds). Increasing the temporal window to 0.48 seconds results in MMCCs of 0.658 for action and 0.641 for activity recognition. I3D, with a 0.48s temporal window, is used in the next experiment.

Table 2: MMCCs when using Inception-V3 (baseline) or I3D as backbone for feature extraction.

Back-bone	Temp. window	Batch size	Actions	Activities
Inc-V3	0.32 sec	4	0.358	0.616
Inc-V3	0.32 sec	1	0.500	0.615
I3D	0.32 sec	1	0.646	0.619
I3D	0.48 sec	1	**0.658**	**0.641**

5.3 Padding and Fusion Function

The action classifications made in the previous experiments are without feature contextualisation. Putting player embeddings into the context of other players is expected to improve model predictions. However, in Table 3 can be seen that using GATs with an additive fusion function and duplication for missing player detections does not improve upon the model without contextualisation. Using zero-padding and concatenation does improve the results towards MMCCs of 0.687 for action and 0.676 for activity recognition.

Table 3: MMCCs with or without (first row) feature contextualisation, using different fusion functions and padding strategies for missing player detections.

Fusion	Padding	Actions	Activities
-	-	0.658	0.641
Addition	Duplication	0.651	0.628
Addition	Zero	0.669	0.607
Concat.	Zero	**0.687**	**0.676**

In Table 4 can be seen that the number of false positives (FPs) can be reduced with feature contextualisation, in activities where only one player can interact with the ball. When a network recognises multiple players as *active* (X$\geq$2) in one activity, it should discover that only one of these is correct. However, the duplication padding strategy weakens this implicit bias, resulting in more FPs than a model without contextualisation. Zero-padding and the suppression of features via the concatenation operation reduces the number of initial FPs with 83%.

Table 4: Number of misclassified *passive* players that are recognised as *active* (false positives), in all activities where only one player can interact with the ball.

Fusion function	Padding	# FPs, when X are recognised as active: X=1	X$\geq$2	**Total #FPs**
-	-	5	36	**41**
Addition	Dupl.	3	43	**46**
Addition	Zero	3	18	**21**
Concat.	Zero	1	6	**7**

5.4 Test Samples From an Unseen Game

We evaluated the ARG (Wu, 2019) and the proposed method on test samples from an unseen game. The former is not able to generalise to these samples and gains MMCCs equal to random predictions (see Table 5). The domain gap between volleyball and soccer is too large for the method to be applied right away. Nevertheless, with the proposed adaptations it is possible to gain MMCCs of 0.623 for action and 0.632 for activity recognition. Compared to the scores on the validation set, the performance drop is only 0.064 and 0.044 respectively. Our method reaches 98.7% MCA for action and 75.2% MCA for activity recognition. However, recall that the dataset is unbalanced.

Table 5: Performance of the baseline and proposed method, both trained with soccer videos, on test samples from an unseen game.

Method	Actions		Activities	
	MMCC	MCA	MMCC	MCA
ARG	0.005	21.6%	0.000	7.2%
Proposed	**0.623**	**98.7%**	**0.632**	**75.2%**

For group activity recognition, we can compare the results with related work in the soccer domain (see Table 6). Our method recognises a large number of activity classes, while it cannot rely on cinematic features in television broadcasts and does not use video input from multiple camera positions. Nevertheless, the maximum reduction in MCA is 20.3%.

Table 6: MCA for activity recognition in soccer videos, with #C the number of classes.

Method	Video input	#C	MCA
Tavassolipour, 2013	TV broadc.	7	82.0%
Jiang, 2016	TV broadc.	4	89.1%
Khan, 2018	TV broadc.	4	94.5%
Vanderplaetse, 2020	TV broadc.	4	95.5%
Tsunoda, 2017	Multi cam.	3	70.2%
Ours	Panorama	11	75.2%

In Table 7, the correlation coefficients and accuracies can be seen per class. The proposed method recognises four actions and six activities with an MCC above 0.7 and with an accuracy above 75% (see classes in bold).

6 CONCLUSION

We proposed a novel method for the recognition of individual actions and group activities in soccer videos from a static panoramic camera. We showed that it is possible to recognise four actions and six activities with accuracies above 75%, in videos captured from a single perspective. We introduced three novel aspects: (1) player snippets and horizon normalisation, (2) spatio-temporal feature extraction, and (3) the use of context information by graph attention networks that use zero-padding and feature suppression.

ACKNOWLEDGEMENTS

The soccer video recordings, player detections and tracks, camera models and virtual camera software were provided by the Netherlands Organisation for Applied Scientific Research (TNO).

REFERENCES

Aalbers, B. and Van Haaren, J. (2018). Distinguishing between roles of football players in play-by-play match event data. In *Int. Workshop on Machine Learn. and Data Mining for Sports Analytics*, pages 31–41. Springer.

Azar, S. M. e. a. (2019). Convolutional relational machine for group activity recognition. In *Proc. of the IEEE Conf. on CVPR*, pages 7892–7901.

Bouma, H. e. a. (2013). Real-time tracking and fast retrieval of persons in multiple surveillance cameras of a shopping mall. In *Multisensor, Multisource Inf. Fusion: Architect., Algorithms, and Appl. 2013*, volume 8756, page 87560A. Int. Soc. for Opt. and Photon.

Table 7: Performance of the proposed method on test samples from an unseen game. Three activity classes describe different kinds of duels: air duel (A), duel with one player in ball possession (P), duel with a loose ball (L). "Ball OOB" is an abbreviation for "ball out-of-bounds". The numbers of training/validation/test samples are provided in the last three rows.

	Passive	Heading	Interception	Dribble	**Play ball**	**In duel**	**Throw in**	Keeper	Duel (A)	Duel (P)	Duel (L)	**Play freely**	Free kick	**Kick off**	**Goal kick**	**Corner**	**Throw in**	Whistle	**Ball OOB**
MCC	**.96**	.47	.00	.56	**.84**	**.85**	**.85**	.47	.58	.59	.33	**.74**	.43	**.76**	**.86**	**.81**	**.84**	.28	**.73**
Acc. (%)	**100**	40	00	50	**88**	**79**	**100**	35	36	64	39	**82**	43	**100**	**78**	**90**	**99**	32	**87**
# Instances	Actions								Activities										
Training	59459	70	67	303	1262	1575	121	34	59	221	468	1541	59	13	41	22	121	74	182
Validation	8710	6	11	40	202	189	15	4	10	26	53	231	7	2	9	9	15	9	32
Test	11140	15	17	44	221	27	35	8	11	32	85	255	13	2	9	10	35	12	49

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. of the IEEE Conf. on CVPR*, pages 6299–6308.

Deng, J. e. a. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE Conf. on CVPR*, pages 248–255. Ieee.

Dollár, P. e. a. (2014). Fast feature pyramids for object detection. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 36(8):1532–1545.

Ekin, A., Tekalp, A. M., and Mehrotra, R. (2003). Automatic soccer video analysis and summarization. *IEEE Trans. on Image Process.*, 12(7):796–807.

Gavrilyuk, K. e. a. (2020). Actor-transformers for group activity recognition. In *Proc. of the IEEE/CVF Conf. on CVPR*, pages 839–848.

Giancola, S. e. a. (2018). Soccernet: A scalable dataset for action spotting in soccer videos. In *Proc. of the IEEE Conf. on CVPR Workshops*, pages 1711–1721.

Ibrahim, M. S. and Mori, G. (2018). Hierarchical relational networks for group activity recognition and retrieval. In *Proc. of the ECCV*, pages 721–736.

Ibrahim, M. S. e. a. (2016). A hierarchical deep temporal model for group activity recognition. In *Proc. of the IEEE Conf. on CVPR*, pages 1971–1980.

Ji, S. e. a. (2012). 3d convolutional neural networks for human action recognition. *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 35(1):221–231.

Jiang, H., Lu, Y., and Xue, J. (2016). Automatic soccer video event detection based on a deep neural network combined cnn and rnn. In *IEEE 28th Int. Conf. ICTAI*, pages 490–494. IEEE.

Kay, W. e. a. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Khan, M. Z. e. a. (2018). Learning deep c3d features for soccer video event detection. In *14th Int. Conf. ICET*, pages 1–6. IEEE.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kong, Y. and Fu, Y. (2018). Human action recognition and prediction: A survey. *arXiv preprint arXiv:1806.11230*.

Matsui, K. e. a. (1998). Soccer image sequence computed by a virtual camera. In *Proc. 1998 IEEE Comput. Soc. Conf. on Comput. Vision and Pattern Recognit. (Cat. No. 98CB36231)*, pages 860–865. IEEE.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Struct.*, 405(2):442–451.

Pappalardo, L. e. a. (2019a). Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM TIST*, 10(5):59.

Pappalardo, L. e. a. (2019b). A public data set of spatio-temporal match events in soccer competitions. *Scientific data*, 6(1):1–15.

Szegedy, C. e. a. (2016). Rethinking the inception architecture for computer vision. In *Proc. of the IEEE Conf. on CVPR*, pages 2818–2826.

Tavassolipour, M., Karimian, M., and Kasaei, S. (2013). Event detection and summarization in soccer videos using bayesian network and copula. *IEEE Trans. on Circuits and Syst. for Video Techn.*, 24(2):291–304.

Tsunoda, T. e. a. (2017). Football action recognition using hierarchical lstm. In *Proc. of the IEEE Conf. on CVPR Workshops*, pages 99–107.

Vanderplaetse, B. and Dupont, S. (2020). Improved soccer action spotting using both audio and video streams. In *Proc. of the IEEE/CVF Conf. on CVPR Workshops*, pages 896–897.

Vaswani, A. e. a. (2017). Attention is all you need. In *Advances in Neural Inf. Process. Syst.*, pages 5998–6008.

Veličković, P. e. a. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Wu, J. e. a. (2019). Learning actor relation graphs for group activity recognition. In *Proc. of the IEEE Conf. CVPR*, pages 9964–9974.

Zach, C., Pock, T., and Bischof, H. (2007). A duality based approach for realtime tv-l 1 optical flow. In *Joint Pattern Recognit. Symp.*, pages 214–223. Springer.

Zhang, K. e. a. (2019). An automatic multi-camera-based event extraction system for real soccer videos. *Pattern Anal. and Appl.*, pages 1–13.