Two-Stream Figure–Ground Group Activity Recognition System in Soccer Videos

Xu Dong 200708160

School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
x.dong@se20.qmul.ac.uk
Project Supervisor: Professor Ebroul Izquierdo

***Abstract*—In this work, we proposed an effective two stream Figure-Ground system for group activity recognition in a soccer video sequence as shown in Figure 1. Our system is inspired by Figure-Ground perception from principles of Gestalt psychology; Figure–Ground perception states that human visual system tends to perceive a scene into *Ground* (background around the main object) and *Figure* (main objects). The entire system involves *(i)* a *Figure* stream, which models the individuals' action features by using extracted skeleton data to capture spatial features, and *(ii)* a *Ground* stream, which mainly models the background features to capture background and temporal features. Specifically, we utilize 3D ConvNets methods such as I3D and C3D for *Ground* stream features extraction methods and 2D ConvNets for skeleton data extraction. Finally, the two streams are concatenated and fed into a rather shallow network for group activity recognition. The proposed model can be trained in an end-to-end manner with back propagation. In addition, we trained and tested our model on the SoccerNet-v2 dataset with 6 complex soccer game events, and our experiments showed promising results. We also compared 5 different variants for ablation studies and demonstrated that our I3D two-stream network significantly improved the group activity recognition results and achieved the highest (85.30%) recognition. In the end of this paper, we also discuss some weaknesses of our proposed method and further work that can be done in the future.** [1]

***Index Terms*—Group activity recognition, video understanding, skeleton action recognition, sports video understanding, two-stream network.**

Fig. 1. Our Figure–Ground network has a **Figure stream** (red bounding box) and a **Ground stream** (blue background texture) as two streams for our group activity recognition system.

I. Introduction

Group activity recognition is an important problem of video understanding tasks. It has attracted interest from many researchers owing to its practical applications, such as autonomous driving, sports video understanding and surveillance systems. As one of the most typical group activities, soccer games involve multiple players interacting directly and simultaneously with other players to achieve an objective.

Understanding a soccer game event is a challenging task because it requires computer systems to understand more integrated high-level features rather than individual behaviours. For instance, in the laws of soccer games, International Football Association Board (IFAB) states that a player is in an offside position if: "any part of the body is in the opponents' half and nearer to the opponents' goal line than both the ball and the second-last opponent." That requires computer system to not only to predict the action and position of individual players but also to analyze the relative position of other players and background information. In recent years, substantial research efforts have been conducted to design a high accuracy and efficient system for soccer game understanding. Recent automated soccer game analytic methods have achieved promising results on low-level video understanding tasks, such as player detection and tracking (Cioppa et al., 2019; Hurault et al., 2020; Ren et al., 2016; Yang and Li, 2017), soccer ball detection (Deep-diver, 2019; Theagarajan et al., 2018), pitch localization (Homayounfar et al., 2016), and team detection (Istasse et al., 2019). However, only few works have focused on automatically understanding high-level soccer game features, such as camera shot selection (Giancola et al., 2018) and high-level action recognition (offside, penalty and foul) (Deliège et al., 2021; Sanford et al., 2020). To understand such high-level features, machines need to enact multiple levels of inference. Previous approaches (Carreira and Zisserman, 2018) feed the entire video clip into a convolutional neural network, which requires a significant amount of data because of the sparsity of events. Recent novel approaches merged different models in a bottom-up multi-model fashion (Giancola and Ghanem, 2021; Wu et al., 2019a; Zappardino et al., 2021).

[1]Our work can be found at: https://github.com/dx199771/Figure-Ground-Group-Activity-Recognition-System

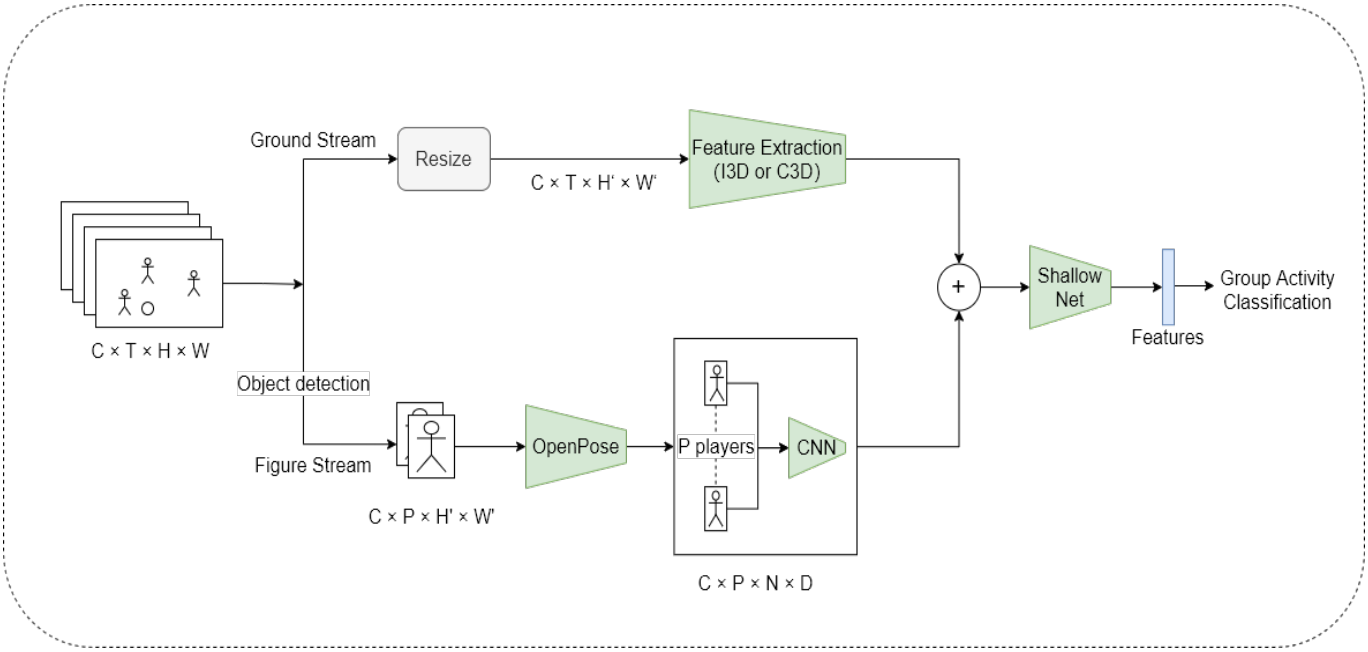

Fig. 2. Activity frame of "Throw-in": Visualization of Detectron2 object detection output (**Bounding boxes**) and OpenPose skeleton output (**Red skeletons**).

These methods extract single object's feature and finally model the relation between objects either by skeleton motion, pivot-actor joint difference or an actor relation graph.

According to the principles of Gestalt psychology (Koffka, 2013), humans perceive the world by Figure–Ground perception. Figure–Ground perception refers to the human visual system tending to perceive a scene into the main object that is the focus of the visual system (figure) and the background (or ground). The Figure–Ground relationship is complementary. Figure and ground can be united and enhance each other, thereby improving the ability of human perception. Taking inspiration from Figure–Ground perception, our proposed method contains two streams: the *Figure* stream and the *Ground* stream, as shown in **Figure 2**. The *Figure* stream takes individuals' skeleton data and feeds them into neural network to extract individual's action features, whereas the *Ground* stream takes background information features. Finally the two streams' results are concatenated and fed into a rather shallow ConvNets for recognizing the final group activity.

More specifically, the main contributions of our work are:

- We proposed a two-stream (*Figure* and *Ground*) group activity recognition system in soccer videos; our model is inspired by Figure–Ground perception and can be trained in an end-to-end manner with back-propagation.
- We conducted massive experiments on the use of group skeleton data in the *Figure* stream and constructed a recognition model by using a 2D convolutional neural network that can learn spatial features between soccer players.
- We constructed a *Ground* information extraction network based on a 3D ConvNets network (Tran et al., 2015)(Carreira and Zisserman, 2018) that can learn temporal background features between video frames.
- We evaluated our methods on the SoccerNet-v2 dataset, and the experiments showed promising results on 6-label testing dataset (Deliège et al., 2021).
- Our comprehensive ablation experiments also demonstrate that our proposed two-stream network significantly improves the group activity recognition results compared to other baselines.

II. Related Work

Related work can be classified to three parts. In the **Video Action Recognition** part, we compare different novel video action recognition approaches related to our *ground* stream; however, these models are used for single-label or individual actions rather than complex group activity. In the **Skeleton Data Representation** part, we introduce skeleton data representations and skeleton feature extraction methods related to our *figure* stream. In the **Group Activity Understanding** part, we introduce and compare different group activity recognition methods.

A. Video Action Recognition

Video action recognition is one of the most important tasks in video understanding. Over the last decade, there has been increasing research interest in video action recognition, and many high-quality action recognition datasets have emerged. The earlier approaches mostly utilized hand-crafted features for video action recognition (Peng et al., 2014; Sargana et al., 2017; Wang and Schmid, 2013). Wang et al. (2011) focuses on feature trajectories as a way to represent video features and proposed a dense trajectories method (DT). DT samples

dense points from each video frame and tracks them based on displacement information in a dense optical flow field. The results showed robust performance on fast irregular motions and shot boundaries and outperformed the current methods. Improved dense trajectories (IDT) (Wang and Schmid, 2013) improved the initial Dense Trajectories method by taking camera motion into account. Also, IDT employs a human detector to determine the position frame of the person for better detection performance. However, hand-crafted features show heavy computational cost problem and suffer from poor universality.

With the rise of deep learning and neural network, 2D ConvNets were firstly used for action recognition. DeepVideo (Karpathy et al., 2014) proposed a multi-resolution 2D ConvNets model on video frames and finds the connectivity in a time domain. Although this work's idea inspired many later work, it shows poor results on the UCF101 dataset (Soomro et al., 2012) (22.5% less than hand-crafted results). In contrast to 2D image data, video essentially has temporal information. Hence, it is important to describe the temporal relationship between frames. Researchers began exploring Recurrent Neural Networks (RNNs) in action recognition tasks, especially the Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997). By combining ConvNets and LSTM, there are several works for video action recognition that used the two-stream networks settings. Two-stream LSTM (Gammulle et al., 2017) utilizes a deep fusion network to exploit spatio-temporal features from CNNs and LSTM models. The results showed higher accuracy than state-of-art results on three widely used datasets. However, specific actions such as 'pick', 'run' and 'stand' still showed lower accuracy values. Inspired by soft-Attention LSTM (Xu et al., 2016), VideoLSTM (Li et al., 2016) proposes a correlation-based spatial attention method with a motion-based attention method that futher improved the video recognition performance and outperformed three widely used dataset.

Due to the high computational cost and low adaptation on other large datasets of previous methods, a novel 3D ConvNets (C3D) were proposed by Tran et al. (2015). C3D adopts an easy and natural way to understand a video as a 3D tensor. The spatio-temporal characteristic enables C3D to outperform the 2D ConvNets on various datasets. However, C3D has more parameters than 2D ConvNets, and C3D does not use ImageNet pre-training model, which makes it difficult to train in practice. Based on C3D, a novel Two-Stream Inflated 3D ConvNets (I3D) was introduced by Carreira and Zisserman (2018). I3D takes video frame sequences as input and feeds them into stacked 3D convolutional layers, which makes it able to extract seamless spatio-temporal features. Also, I3D employs 2D pretrained-ImageNet-Inception-v1 as a weight initialization method, expands it into three dimension and achieves the state-of-art results on UCF-101 and HMDB-51 action classification datasets.

In our work, we adopt a I3D-like method as our *Ground* stream feature extraction model because I3D can extract spatio-temporal information from videos and perform well on short-range temporal video sequences. Furthermore, miscellaneous action recognition approaches such as multi-stream networks, flow-mimic approaches, frame/Clip sampling and rank pooling methods also showed promising results on action recognition benchmarks. These methods are summarized and compared in the work of Zhu et al. (2020). We also list and compare some widely used action recognition methods' performance on three datasets in **Table IV**.

B. Skeleton-Based Action Recognition

In addition to motion, appearance and trajectories such image/sequence of frame information, body pose (skeleton) can also provide important information about human actions without a scene context. In recent years, skeleton-based action recognition methods have became the most active research in the action understanding field due to its recognition naturalness and compactness. From a neural networks point of view, skeleton-based action recognition methods can be classified into three categories: RNN-based, CNN-based and GCN-based (Ren et al., 2020).

RNN-based methods take output from the last frame as input and feed it into the RNN model, which is a reasonable way for processing sequential data. However, because most RNN-based methods cannot extract spatial data, the results demonstrated poor performance on action recognition tasks. To solve this problems, several RNN-based method variants focus on modelling both spatial and temporal skeleton data. Wang and Wang (2017) composed a novel two-stream method that adds spatial configurations. Xie et al. (2018) employed a temporal Attention Recalibration Module (TARM) in an RNN+CNN network and significantly improved the results on four benchmarks. Although there are substantial research has been devoted to solving the problems of a single frame's skeleton relationship and the large viewpoint problems of RNN-based methods, the RNN-based methods still showed poor performance.

Different to RNN-based methods, CNN-based approaches can easily extract high-level semantics owing to their natural characteristics. Some works focused on modelling skeleton data to two dimensions pseudo-images and feeding them into a ConvNets to learn features (Caetano et al., 2019; Ke et al., 2017; Wang et al., 2018) with some manually designed transformations. These methods either compare the magnitude values and direction of the skeleton joints or divide the skeleton data into three different clips and learn features separately. Although these methods carefully designed the transformations and network settings, they still show inferior results because 2D pseudo-image methods cannot take advantages of the locality of ConvNets. More recently, Duan et al. (2021) exploits a 3D-CNN based approach named PoseC3D, which is based on a 3D heatmap stack to represent skeleton data. PoseC3D is more robust against noise data, more effective on learning spatio-temporal representations and shows better results on various datasets.

Recently, many skeleton-based action recognition methods have employed graph convolutional networks (GCNs) to ex-

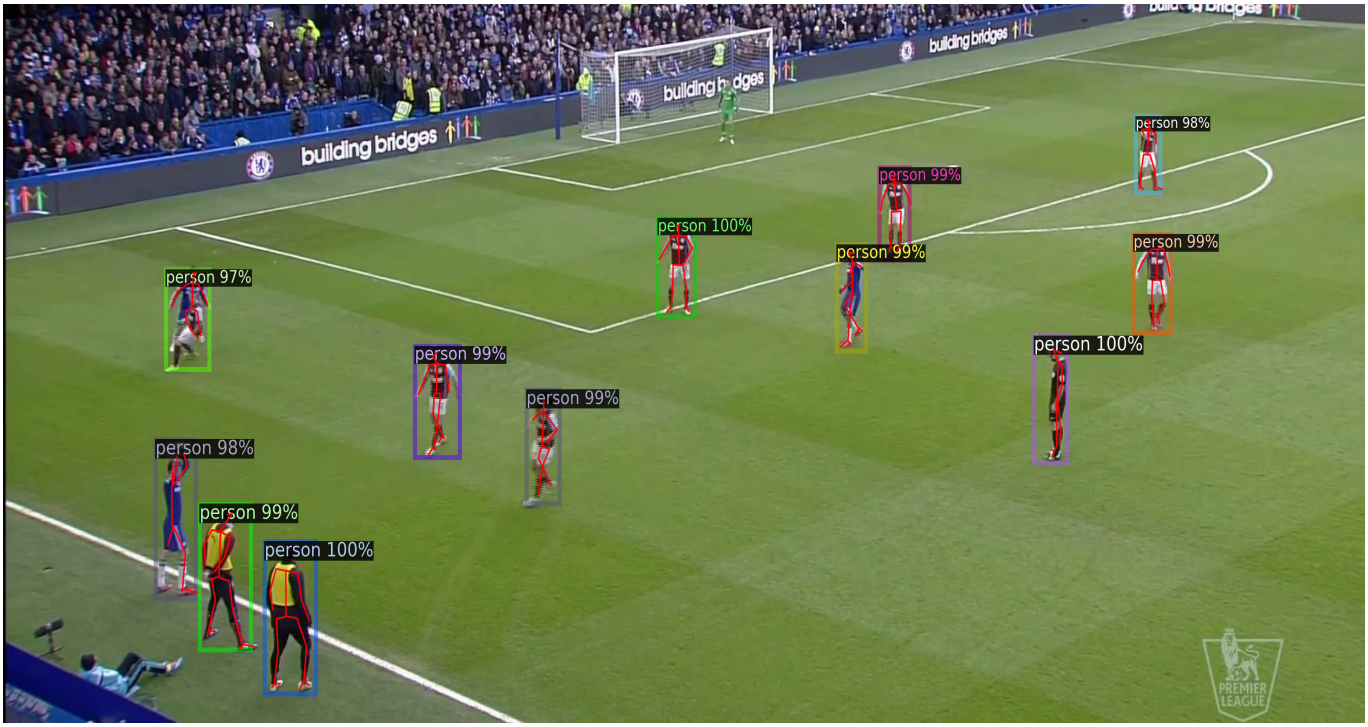

Fig. 3. Activity frame of "Throw-in": Visualization of Detectron2 object detection output (**Bounding boxes**) and OpenPose skeleton output (**Red skeletons**).

tract skeleton features. GCN-based methods model human skeleton data as spatial temporal graphs. The first and well-known GCN method is ST-GCN (Yan et al., 2018), which models both spatial and temporal features from extracted graphs ST-GCN shows strong expressive power and excellent generalization ability and achieves state-of-the-art results on two datasets. Other GCN-based methods improve ST-GCN (Papadopoulos et al., 2019; Song et al., 2021) by either reducing the training time and memory or enhancing the recognition accuracy. Despite GCN-based methods having achieved substantial success in skeleton-based action recognition, it still shows poor robustness and scalability (Zhu et al., 2019).

In this work, we create a simple and effective CNN-based method for extracting skeleton features as our *Figure* stream network settings; details are shown in the **Methodology** section. Furthermore, we also summarize some widely used skeleton-based action recognition methods and compare the recognition results in **Table VI** and **Table V**.

C. Group Activity Understanding

Different from video action recognition and skeleton-based action recognition tasks, group activity recognition focuses on more complex activity in scenes with multiple persons. Earlier group activity recognition approaches mostly focus on extracting hand-crafted features, followed by probabilistic graphical models for group activity recognition (Amer et al., 2012). Amer et al. (2012) proposed a three-layered AND-OR graph to localize and detect group activities. Although AND-OR approaches show good results on the UCLA campus dataset, it demonstrated poor universality on other datasets.

More recently, deep learning approaches have demonstrated significant performance on group activity recognition. Some approaches included Deng et al. (2016); Donahue et al. (2016); Ibrahim et al. (2016a) focus on using an RNN-based method and temporal information to represent group activity features. Other works considered CNN-based methods or GCN-based methods to represent group activity. Azar et al. (2019) proposed a Convolutional Relational Machine (CRM) method that utilizes the spatial relations between actors to recognize group activity. Wu et al. (2019a) flexibly and efficiently learned actors' relations as a graph (ARG) for modelling relation between actors. The ARG can be automatically learned in an end-to-end manner, and the recognition accuracy achieved state-of-art on two datasets by using this method. Additionally, Zappardino et al. (2021) employed skeleton data to train an end-to-end system for group activity recognition. The (Zappardino et al., 2021) method represents the skeleton data in three different branches: group sequence of skeletons, Skeleton Motion and Pivot-actor joint differences. The method shows highly competitive results on the Volleyball dataset (Ibrahim et al., 2016b).

Our work differs from these approaches. In this paper, we attempt to merge individuals' skeleton information and background information for modelling group activity features. Similar to the work of Zappardino et al. (2021), we employed individual skeleton data as our *figure* stream data representation to model *Figure* stream features. Furthermore we employed the 3DCNN-based method for our *Ground* stream feature extraction.

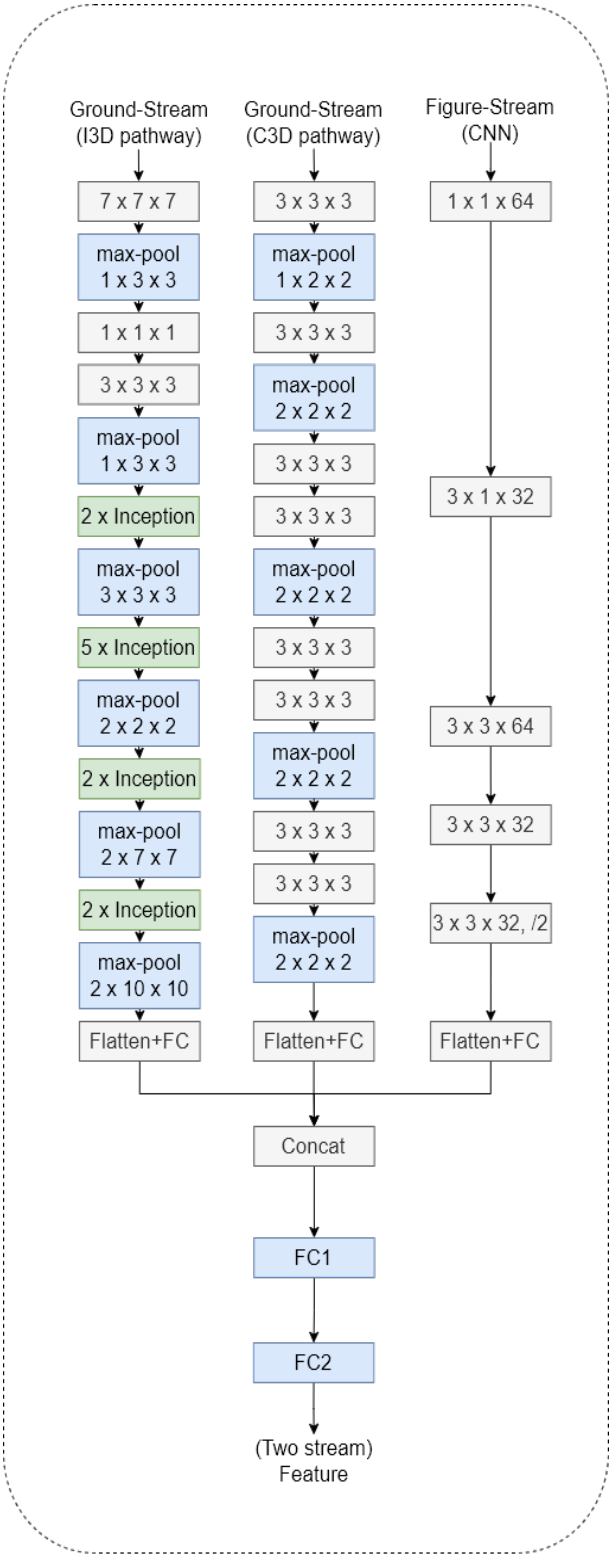

Fig. 4. Architecture of our proposed two-stream model. Blue box denote pooling and fully connection layer, Grey box denotes convolutional layers and green box denotes inception layers of I3D (Carreira and Zisserman, 2018) inception module.

III. Methodology

In this section, we present our two-stream (*Figure* stream and *Ground* stream) approach for soccer game group activity recognition using both background information and skeleton data as shown in **Figure 2**. We outline our two-stream network by introducing different streams separately. Besides, we also introduce the two-stream network settings and three different network variants (two-stream, single I3D variant, single C3D variant, single skeleton variant) for ablation studies as shown in **Figure 4**.

A. Single Frame Skeleton Representation (Figure stream)

For the training stage, a short soccer game clip with n frames are firstly fed into the *Figure* stream after normalization; the middle frame's players bounding boxes are then obtained through Faster R-CNN (Ren et al., 2016) using the Facebook Detectron2 library (Wu et al., 2019b). Then, each bounding box is fed into OpenPose (Cao et al., 2019), and a group of skeleton data is obtained as **FS**. **FS** can be represented as a $C \times P \times N \times D$ tensor. Here, C denotes the number of channels, P is the number of player in a frame, N is the number of keypoints of one individual's skeleton and D is the keypoint coordinate dimension. Therefore, the group skeletons data can be represented as $\mathbf{FS} = \{\mathbf{S}^1, \mathbf{S}^2, ..., \mathbf{S}^P\}$, and the single skeleton of a person p as

$$\mathbf{S}^p = [J_1^p, J_2^p, ..., J_N^p] \tag{1}$$

where $J = [x, y, p]$ is the skeleton coordinate and precision generated by the OpenPose library.

Because each frame may contains different numbers of skeleton data, the network requires consistent number of skeleton input. To normalize the data, here, we fill frames that has inadequate data by duplicating existing data. As for frames with exceeding skeleton data, we extract specific numbers of skeleton data with the highest confidence score.

After obtaining the skeleton data, as for two-stream variant configurations, the skeleton data are firstly fed into the *Figure* stream and a 3 dimension features are output, after flatting the features and going through the fully connected layer. The final output comprises a one dimension 1024 features. As for the single CNN skeleton variant, the skeleton data go through two fully connected layers before go through the *figure* stream and outputting a one dimension 512 features for the classifier.

B. Background feature representation (Ground stream)

Background features are extracted from a video feature extraction network, and here we adopt Inflated 3D CNN (I3D) (Carreira and Zisserman, 2018) and 3D Convolutional neural network (C3D) (Tran et al., 2015). For training stage, a short soccer game clip with n frames is firstly resized and normalized and input to the network with size $C \times T \times H \times W$, where C is the number of channels, T is the number of frames, and H and W denote the height and width of the video, respectively. The video sequences **GF** can thus be represented as

$$\mathbf{GF} = [F^1, F^2, ...F^t] \tag{2}$$

where $F = [h, w]$ is the height and width of a single frame.

In the two-stream variant configuration, after obtaining the input video sequences representations, the sequences data are then fed into the *Ground* stream (I3D or C3D) directly, and a 4 dimension features are output. The features are then flattened and go through a fully connected layer and finally generate a one dimension 1024 features. For the single I3D and C3D, variants, the frames data are fed into either I3D or C3D and a 4 dimension features are output. The features then go through a $1 \times 1 \times 1$ convolutional layer, and a 2 dimension features are output which will be used in the classifier.

C. Two-streams Classification

The *Ground* and *Figure* stream features are simultaneously fed into the network with different network configurations. However, their weights are not shared and trained separately. Each stream outputs a 1024 tensor. In terms of loss function, We utilizes binary cross entropy with logits as our loss function because it provides more numerically stable loss value than using a plain combined Sigmoid and BCELoss function. For group activity classification methods, we proposed two strategies:

1) Concatenating descriptors: The final two streams' feature maps are concatenated, obtaining a 2×1024 tensor to represent the action's features. After going through two fully connection layers and calculating the binary cross entropy with logits loss. The final actions are then classified by the classifier. The final loss function is shown as below:

$$Loss = L_T(y^H, \hat{y}^H) \tag{3}$$

Here, L_T denotes the cross-entropy loss of our two-stream concatenated network, y^H is the ground-truth labels of group activity. $\hat{y}^H$ is the output prediction value.

2) Using multiple classifiers: Different to concatenating descriptors method, the final two streams' features are fed into different classifiers and obtained two classification results. The final loss function is shown as below:

$$Loss = L_F(y^G, \hat{y}^G) + \alpha L_G(y^I, \hat{y}^I) \tag{4}$$

Here, L_F denotes the cross-entropy loss of the *Figure* stream, where L_G denotes the cross-entropy loss of *Ground* stream. y^G and y^I are the ground-truth labels of group activity. $\hat{y}^G$ and $\hat{y}^I$ are the output prediction value. Here, we introduce variable α for balancing these two losses. The higher α stream means there will be more weight assigned to this stream.

IV. Experiment and results

In this section, we firstly introduce our dataset as well as the implementation details of our network. Then, we empirically demonstrate and compare the efficiency and effectiveness on the SoccerNet-v2 dataset of 5 different variants for ablation studies.

TABLE I
SOCCER-NET V2 TRAINING AND TESTING ACTION LABELS PERCENTAGE AMONG 6 LABELS.

Action Labels	Training	Testing	Total Action
Substitution	3.97%	4.40%	4.18%
Corner	7.53%	8.70%	8.11%
Yellow card	2.89%	2.90%	2.90%
Clearance	6.26%	8.30%	7.28%
Throw-in	24.39%	21.3%	22.85%
Ball out of play	**54.96%**	**54.40%**	**54.68%**

TABLE II
MODEL PARAMETERS AND FLOPS COMPARISON WITH 320×320 INPUT DATA SEQUENCES

Method	Parameters	Flops
Inflated 3D ConvNets (I3D)	24.88M	179.04GMAC
3D ConvNets (C3D)	**65.83M**	**974.26GMAC**
Figure ConvNets	1.16M	0.01GMAC
Two-stream	18.88M	0.08GMAC

A. Dataset

In this work, we conduct our experiments on the SoccerNet dataset (Deliège et al., 2021; Giancola et al., 2018). SoccerNet is a novel large-scale soccer game dataset with three 3 different annotation categories: actions, camera shots, and replays. In this work, we only used actions annotations. In particular, SoccerNet actions annotations contains 500 full soccer games' actions, with 17 different labels; see **Appendix A**. The total number of actions reached 110,458 which is equal to, on average, 221 actions per game. To simplify our training, we only train our network with 6 labels: *Corner, Throw-in, Clearance, Goal, Yellow card* and *Substitution* with 10184 labels and test our network with 3330 labels. The percentage of training and testing labels number are shown in **Table I**.

B. Implementation Details

We adopt stochastic gradient descent with momentum to learn the network parameters with an initial learning rate $lr = 0.01$, momentum $\mu = 0.9$. The learning rate decreases by a factor of 10 every 10 epochs. The input of the *Ground* stream is a 50-frame image sequence of size 320×320, where the input of the *Figure* stream has 8 skeleton data with the highest detection confidence value. We set our loss function balanced factor as $\alpha = 0.4$. Furthermore, we train our network for 50 epochs with a mini-batch size of 4 (due to GPU memory limitations). As for data augmentation, we utilized Random Crop, horizontally flipping as our data augmentation methods. During the experiments, we found that our network suffered from an overfitting problem. To minimize the overfitting risk, we tried to either add a dropout layer with 0.5 probability or batch normalization function during the training process in *Ground* stream, and we add a 0.3 probability dropout layer in the *Figure* stream. In terms of the device and framework (includes baseline models and different variants), we trained our models on a machine with i7-8700k cpu, NVIDIA GTX1080

TABLE III
RESULTS OF 5 NETWORK CONFIGURATIONS ACTION RECOGNITION RESULTS AMONG 6 LABELS.

Method	Substitution	Corner	Yellow card	Clearance	Throw-in	Ball out of play	Overall
Counts(testing)	132	261	87	249	639	1632	3000
Ground-stream with I3D	51.98%	30.02%	10.20%	15.67%	73.18%	91.71%	45.46%
Ground-stream with C3D	26.81%	13.79%	34.48%	1.20%	30.79%	22.13%	21.53%
Figure-stream (2D-Skeleton)	50.93%	89.27%	46.79%	**96.86%**	71.36%	96.23%	75.24%
Two-stream with I3D ground	**70.45%**	**95.40%**	**62.07%**	89.16%	**96.24%**	**98.5%**	**85.30%**
Two-stream with C3D ground	46.97%	17.24%	47.13%	16.30%	54.77%	76.59%	43.17%

8GB GPU and 32GB RAM, we utilized PyTorch as our deep learning framework.

C. Model parameters comparision

Model parameter is an important factor of model efficiency. As shown in Table II, Single 3D ConvNets(C3D) and Inflated 3D ConvNets(I3D) have the most parameters and Flops. That is mainly because these two networks are used to extract video information, which contains a large amount of data (320×320×50); also, I3D has more width network (Inception Net), and C3D has more depth network (up to 15 layers). Figure ConvNets has the fewest parameters and Flops due to shallow network layers and a small input data size. In practice, processing a video sequences on C3D two-stream variant spent around five hours and spent three and half hours on two-stream I3D variant on our machine.

D. Baseline and Different Variant Results for Ablation Studies

Although our network can be trained in an end-to-end manner, it can also be implemented in different variants. In this section, we firstly illustrate our three single-stream and two two-stream variants of our proposed method on the Soccernet-v2 dataset, and then we compare the baseline results with four different variants for ablation studies purposes. We test the following variants:

1) ***Ground stream with I3D****:* This variant only has an Inflated 3D ConvNets (I3D) model for extracting background information only without considering *Figure* stream information (skeleton data).

2) ***Ground stream with C3D****:* This variant only has a 3D ConvNets (C3D) model for extracting background information only without considering *Figure* stream information (skeleton data).

3) ***Figure stream (2D-Skeleton)****:* This variant is for extracting *Figure* information (skeleton data) only wihtout considering *Ground* stream features.

4) ***Two stream with I3D ground****:* This variant contains two streams for extracting both *Ground* and *Figure* stream information. It adopts I3D for extracting *Ground* stream features. It can be trained in an end-to-end manner.

5) ***Two stream with C3D ground****:* This variant contains two streams for extracting both *Ground* stream information and *Figure* stream information. It adopts C3D for extracting *Ground* stream features. It is the end-to-end version.

The comparison group activity recognition results of the above variants are shown in **Table III**. All the obtained variants' results include data augmentation and methods that can prevent overfitting. In terms of a single **ground stream with I3D** variant, the overall results reached 45.46%. The ball out of play had the highest recognition results (91.71%) whereas the lowest results were found for the Yellow card, which had 10.20% recognition accuracy. As for single **ground stream with C3D** variant, the overall results only had 21.53% accuracy, which is lower than single I3D results. The highest recognition label is Yellow card and the lowest label is clearance (only 1.20%).

In terms of **Figure-stream (2D-Skeleton)** variants. The overall results reached 75.24% recognition accuracy, which is higher than single Ground-stream with I3D and C3D variants. The highest result label is Clearance (96.85%) and the lowest label is Yellow Card (46.79%)

Regarding the **two-stream with I3D ground** and **two-stream with C3D ground** variants, the overall results of Two-stream with I3D ground variant has the highest overall results (85.30%) among five variants, which is significantly higher than the single stream variants. However the Clearance label has slightly lower results than single figure-stream (96.86%). The **Two-stream with C3D ground** has lower overall results than Two-stream I3D variant, and some of the labels such as Substitution, Corner, Clearance in single Figure-stream variant even higher than **Two-stream with C3D ground** results. In conclusion, It is clear that whether in single stream I3D, C3D, or 2D-skeleton variants, two-stream with I3D ground has higher performance results than these single stream variants.

V. DISCUSSION

As shown in **Table III**, we observed that the two-stream with I3D ground model provided the best results among the 5 variants. The reason that two stream with I3D ground had the highest recognition results is mainly because this variant integrates both *figure* features obtained by players' skeleton data and *ground* information from the action frames sequence therefore giving the model the ability to extract spatio-temporal features. Conversely, only using ground-stream variants led to the worst results. Primarily because only using

Ground stream cannot model complex group activity. It is clear that our proposed two-stream model has significantly improved recognition performance compared to only using a one-stream model.

Furthermore, we investigated that single label's recognition accuracy. The confusion matrix obtained for our SoccerNet-v2 dataset using a two-stream approach with I3D ground variants is shown in **Figure 5**. We observed that the model performs particularly well on labels such as Corner, Throw-in, Ball out of play events. However, labels such as Yellow card and Substitution had lower recognition results. Primarily because such events have more diverse individual actions (i.e., Substitution can be displayed in different ways in various angles), and background information and fewer training samples which is difficult for neural network to learn a generalized model to represents group activity features.

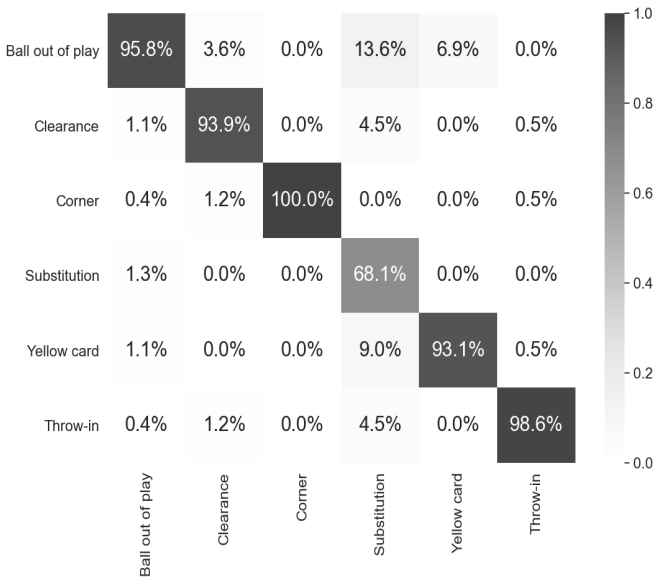

Fig. 5. Confusion matrix for the SoccerNet-v2 6 labels testing dataset obtained using our two-stream with I3D ground model.

VI. Limitations and Further work

The present two-stream figure–ground group activity recognition system showed promising results on the SoccerNet-v2 dataset. However, it still has some weaknesses due to the time and computational resource limitations. Therefore, further work should be preformed to improve the system efficiency and performance in the following aspects:

1) Class imbalanced data and enlarge dataset: Effectively classifying imbalanced data is an important issue in the deep learning area. In our work, there was a serious problem of imbalanced data. As shown in I, class labels such as Yellow card only had 294 training samples, whereas class label Ball out of play had 5594 samples. This difference will lead to poor recognition performance, especially for the minority class. To solve the class imbalanced data problem, we could adopt methods such as Online Hard Example Mining (Shrivastava et al., 2016), SMOTE: Synthetic Minority Over-sampling Technique (Chawla et al., 2002) and few shot learning methods to improve the system performance. Additionally, we only trained our system on 6 labels with 10184 labels. In further work, we will train the system on all 17 labels with more data from the SoccerNet-v2 dataset, which may improve the recognition accuracy.

2) Model improvements: A crucial extension of this work is to use skeleton sequences instead of single-frame skeleton data for extracting *Figure* stream features. Single-frame skeleton data only consider the spatial information that may lose temporal features and, thus, lead to poor recognition performance. Furthermore, batch size is an important hyper-parameter during deep learning processes. The more samples used during training process, the more accurate the model will be. Due to the computational limitations, in our experiment, each batch size only contains two training samples and can be increased further.

3) Model efficiency: Model efficiency is also an important factor that can be improved. As shown in Table II, *Ground* streams such as I3D and C3D have the most parameters and Flops. This could be improved by using pre-trained models such as transfer learning methods (Zhuang et al., 2020) or utilizing high-efficiency optimization methods and loss function.

VII. Conclusion

In this work, we have shown a novel two-stream group activity recognition system. The system is inspired by Gestalt's figure-ground perception from psychology and can be trained in an end-to-end manner. The main idea of figure-ground perception is that the human visual system tends to divide a scene into figures and background, which both provide important information for recognizing events and objects. In our system, the *Figure* stream utilizes players' skeleton information to extracts main objects' action information, and the *Ground* stream extracts background features behind the main objects by inputting the image sequences into the network. We also evaluated our system on the SoccerNet-v2 dataset with 6 labels. We compared the efficiency and recognition accuracy among 5 different variants for ablation studies. The final experiment showed promising results on the SoccerNet-v2 dataset. Finally, we also discussed some weaknesses of our system that can be further developed in the future.

VIII. Acknowledgment

I would like to express my particular gratitude to my Supervisor Professor Ebroul Izquierdo, who gave me a lot of guidance and support on project ideas and throughout the project. I also thanks Mr Bilal Hassan for providing me a lot of advice on programming and many valuable and useful academic resources. Thanks to Andrew Boyd for proofreading and Deliège et al. (2021) for sharing the datasets.

References

M. R. Amer, D. Xie, M. Zhao, S. Todorovic, and S.-C. Zhu. Cost-sensitive top-down/bottom-up inference for multiscale activity recognition. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part IV*, ECCV'12, page 187–200, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642337642. doi: 10.1007/978-3-642-33765-9_14. URL https://doi.org/10.1007/978-3-642-33765-9_14.

S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi. Convolutional relational machine for group activity recognition, 2019.

C. Caetano, J. Sena, F. Brémond, J. A. dos Santos, and W. R. Schwartz. Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition, 2019.

Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.

J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16: 321–357, Jun 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953.

A. Cioppa, A. Deliege, M. Istasse, C. De Vleeschouwer, and M. Van Droogenbroeck. Arthus: Adaptive real-time human segmentation in sports through online distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Deep-diver. Soccer ball detection using yolov2 (darkflow). 2019. URL https://github.com/deep-diver/Soccer-Ball-Detection-YOLOv2.

A. Deliège, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. V. Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos, 2021.

Z. Deng, A. Vahdat, H. Hu, and G. Mori. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition, 2016.

J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description, 2016.

H. Duan, Y. Zhao, K. Chen, D. Shao, D. Lin, and B. Dai. Revisiting skeleton-based action recognition, 2021.

H. Gammulle, S. Denman, S. Sridharan, and C. Fookes. Two stream lstm: A deep fusion framework for human action recognition, 2017.

S. Giancola and B. Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts, 2021.

S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun 2018. doi: 10.1109/cvprw.2018.00223. URL http://dx.doi.org/10.1109/CVPRW.2018.00223.

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

N. Homayounfar, S. Fidler, and R. Urtasun. Soccer field localization from a single image, 2016.

S. Hurault, C. Ballester, and G. Haro. Self-supervised small soccer player detection and tracking. *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, Oct 2020. doi: 10.1145/3422844.3423054. URL http://dx.doi.org/10.1145/3422844.3423054.

M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition, 2016a.

M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. Hierarchical deep temporal models for group activity recognition. 2016b.

M. Istasse, J. Moreau, and C. D. Vleeschouwer. Associative embedding for game-agnostic team discrimination, 2019.

A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. doi: 10.1109/CVPR.2014.223.

Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. doi: 10.1109/cvpr.2017.486. URL http://dx.doi.org/10.1109/CVPR.2017.486.

K. Koffka. *Principles of Gestalt psychology*. Routledge, 2013.

Z. Li, E. Gavves, M. Jain, and C. G. M. Snoek. Videolstm convolves, attends and flows for action recognition, 2016.

K. Papadopoulos, E. Ghorbel, D. Aouada, and B. Ottersten. Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition, 2019.

X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, 2014.

B. Ren, M. Liu, R. Ding, and H. Liu. A survey on 3d skeleton-based action recognition using learning method, 2020.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

R. Sanford, S. Gorji, L. G. Hafemann, B. Pourbabaee, and M. Javan. Group activity detection from trajectory and video data in soccer, 2020.

A. B. Sargana, P. Angelov, and Z. Habib. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *Applied Sciences*, 7:110, 01 2017. doi: 10.3390/app7010110.

A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining, 2016.

Y.-F. Song, Z. Zhang, C. Shan, and L. Wang. Constructing stronger and faster baselines for skeleton-based action recognition, 2021.

K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild, 2012.

R. Theagarajan, F. Pala, X. Zhang, and B. Bhanu. Soccer: Who has the ball? generating visual analytics and player statistics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1830–18308, 2018. doi: 10.1109/CVPRW.2018.00227.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks, 2015.

H. Wang and C. Schmid. Action recognition with improved trajectories. In *2013 IEEE International Conference on Computer Vision*, pages 3551–3558, 2013. doi: 10.1109/ICCV.2013.441.

H. Wang and L. Wang. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks, 2017.

H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, page 3169–3176, USA, 2011. IEEE Computer Society. ISBN 9781457703942. doi: 10.1109/CVPR.2011.5995407. URL https://doi.org/10.1109/CVPR.2011.5995407.

P. Wang, W. Li, C. Li, and Y. Hou. Action recognition based on joint trajectory maps with convolutional neural networks. *Knowledge-Based Systems*, 158:43–53, 2018. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2018.05.029. URL https://www.sciencedirect.com/science/article/pii/S0950705118302582.

J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu. Learning actor relation graphs for group activity recognition, 2019a.

Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019b.

C. Xie, C. Li, B. Zhang, C. Chen, J. Han, C. Zou, and J. Liu. Memory attention networks for skeleton-based action recognition, 2018.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2016.

S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, 2018.

Y. Yang and D. Li. Robust player detection and tracking in broadcast soccer video based on enhanced particle filter. *Journal of Visual Communication and Image Representation*, 46:81–94, 2017. ISSN 1047-3203. doi: https://doi.org/10.1016/j.jvcir.2017.03.008. URL https://www.sciencedirect.com/science/article/pii/S1047320317300640.

F. Zappardino, T. Uricchio, L. Seidenari, and A. del Bimbo. Learning group activities from skeletons without individual action labels. *2020 25th International Conference on Pattern Recognition (ICPR)*, Jan 2021. doi: 10.1109/icpr48806.2021.9413195. URL http://dx.doi.org/10.1109/ICPR48806.2021.9413195.

D. Zhu, Z. Zhang, P. Cui, and W. Zhu. Robust graph convolutional networks against adversarial attacks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019.

Y. Zhu, X. Li, C. Liu, M. Zolfaghari, Y. Xiong, C. Wu, Z. Zhang, J. Tighe, R. Manmatha, and M. Li. A comprehensive study of deep video action recognition, 2020.

F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning, 2020.

APPENDIX

WIDELY USED MODELS COMPARISON (ACTION RECOGNITION MODELS)

TABLE IV
RESULTS OF WIDELY USED MODELS IN ACTION RECOGNITION ON THREE DATASETS.

Model	**Backbone**	**UCF101**	**HMDB51**	**Kinetics400**
DeepVideo	AlexNet	65.4%	-	-
Two-stream	CNN-M	88.0 %	59.4 %	-
Fusion	CNN-M	92.5 %	65.4 %	-
TSN	CNN-M	94.0%	68.5	73.9%
C3D	VGG-16	82.3 %	56.8%	59.5 %
I3D	BN-Inception	**95.6%**	**74.8%**	71.1 %
P3D	ResNet50	88.6 %	-	71.%
ResNet3D	ResNeXt101	94.5%	70.2 %	65.1 %
SlowFast	ResNet101-NL	-	-	**79.8%**

WIDELY USED MODELS COMPARISON (SKELETON-BASED ACTION RECOGNITION MODELS)

TABLE V
RESULTS OF WIDELY USED MODELS IN SKELETON-BASED ACTION RECOGNITION ON NTU-RGB+D.

Methods	Accuracy(CS)	Accuracy(CV)
Lie Group	50.10%	52.80%
ST-LSTM+TS	69.20%	70.30%
ST-GCN	81.50%	88.30%
RNN+CNN+Attention	82.70%	93.20%
Graph Convolutioanl Netowkrs	86.80%	94.20%
AGC-LSTM	89.20%	95.00%
Two stream adaptive GCN	88.50%	95.10%
Directed Graph Neural Networks	**89.90%**	**96.10%**

TABLE VI
RESULTS OF WIDELY USED MODELS IN SKELETON-BASED ACTION RECOGNITION ON NTU-RGB+D-120.

Methods	Accuracy(CS)	Accuracy(CSe)
Spatio-Temporal LSTM	55.70%	57.90%
Global Context-Aware Attention LSTM	58.30%	59.20%
Two-Stream Attention LSTM	61.20%	63.30%
TSRJI	67.90%	62.80%
FGCN	85.40%	87.40%
PoseC3D	86.90%	**90.30%**
EfficientGCN-B4	**88.30%**	89.10%

ANNOTATION GUIDELINES

Here, we summarize all the annotation guidelines' text descriptions as below to help better understand the action labels (Deliège et al., 2021):

- Ball out of play: Moment when the ball crosses one of the outer field lines.
- Throw-in: Moment when the player throws the ball
- Foul: Moment when the foul is committed
- Indirect free-kick: Moment when the player shoots, to resume the game after a foul, with no intention to score
- Clearance (goal-kick): Moment when the goalkeeper shoots
- Shots on target: Moment when the player shoots, with the intention to score, and the ball goes in the direction of the goal frame
- Shots off target: Moment when the player shoots, with the intention to score, but the ball does not go in the direction of the goal frame
- Corner: Moment when the player shoots the corner
- Substitution: Moment when the replaced player crosses one of the outer field lines
- Kick-off: Moment when, at the beginning of a halftime or after a goal, the two players in the central circle make the first pass

- Yellow card: Moment when the referee shows the player the yellow card
- Offside: Moment when the side referee raises his flag
- Direct free-kick: Moment when the player shoots, to resume the game after a foul, with the intention to score or if the other team forms a wall
- Goal: Moment when the ball crosses the line
- Penalty: Moment when the player shoots the penalty
- Yellow then red card: Moment when the referee shows the player the red card
- Red card: Moment when the referee shows the player the red card

MSc Project - Reflective Essay

Project Title:	Two-Stream Figure–Ground Group Activity Recognition System in Soccer Videos
Student Name:	Xu Dong
Student Number:	200708160
Supervisor Name:	Professor Ebroul Izquierdo
Programme of Study:	Media and Arts Technology MSc

Introduction

The aim of this work is to investigate a video understanding system that can recognize complex group activity in soccer games. In order to achieve this aim, there are several objectives that need to be considered: A suitable dataset needs to be collected or found for model training and testing; A video understanding system needs to be constructed that can effectively recognize the group activity in the dataset. An efficient training implementation procedure needs to be investigated to avoid any unexpected issues such as overfitting. Inspired by the principle of Gestalt psychology (Koffka,2013), humans perceive the world by Figure-ground perception. I proposed a figure-ground two stream group activity recognition system and conducted an experiment on Soccernet-v2 (Deli`ege et al., 2021) dataset. In this reflective essay, I will briefly discuss the methodology that I adopted in the system; I will analyze the strengths and weaknesses with other related work; I will discuss the possibilities for further work; I will also critically analyze the relationship between theory and practical work produced. Finally I will discuss the awareness of Legal, Social Ethical Issues and Sustainability of the system.

Methodology

To better understand the system, I briefly introduce the methodologies in this section. The system contains two streams that are inspired by figure-ground perception from psychology terms. The *figure* stream mainly models the main objects in a scene such as each player in a soccer game (here the figure information is the skeleton data of players) while the *ground* stream extracts the background information around the main objects (here the background information is equivalent to the frames sequence). The final two streams system can be trained by concatenate *figure* and *ground* streams and the final trained system is able to recognize complex group activity in soccer games.

Analysis of strengths/weaknesses

The two stream figure-ground group activity recognition system has several strengths and weaknesses. In terms of strengths:

1. The system is different to previous work which only uses either spatio-temporal feature extraction models or skeleton extraction models. In this work, I proposed a novel two stream networks inspired by the psychology term “figure-ground perception” that combine both figure information and ground features to represent group activity features. Furthermore, from my best knowledge this is the first group activity system that uses both *figure* and *ground* information.
2. After reviewing a lot of literature and critically analyzing the related approaches. In terms of figure stream network, The system employs a novel 2D convolutional neural network model to extract players' skeleton features from

game scenes. The figure stream ConvNets model contains five convolutional layers and two fully connected layers to modelling players' skeleton features. In terms of ground stream feature extraction methods. I conduct experiments on two classic video understanding models: Inflated 3D ConvNets (I3D) (J. Carreira, 2018) and 3D convolutional networks (C3D) (Tran et al., 2015). I modified some layers of I3D and C3D such as the pooling method and receptive field size to make the model more robust and trained with high efficiency.

3. During the experiment process, I trained these models separately and finally concatenated the extracted features to represent the group activity. In order to minimize the overfitting problem. I adopt batch normalization, dropout methods. I compared each variant's recognition results and training speed with the baseline for ablation studies.

Although the system shows promising results on SoccerNet-v2 datasets with 6 labels, it still suffers from a lot of weakness. In terms of weakness:

1. The system is only trained with 6 labels and 10184 labels due to the time limitations. This insufficient data may lead to the performance of the system not meeting expectations.
2. During the experiment training process, the *figure* stream can be only trained with two batch sizes per epoch due to the GPU memory limitations which may affect the final system performance.
3. To simplify our training process, in *figure* stream,I only use single frame's skeleton data to train *figure* stream models.
4. I only train and test the model on one dataset, which may lead to low generalization ability.
5. I observed that the imbalance of data is an important issue of our system, which leads to serious imbalance recognition accuracy.

Possibilities for further work

In the last section, I listed several strengths and weaknesses of our system. In order to solve the possible weaknesses, further work can be done and further improve the system efficiency and performance in the following sections:

1. In order to solve the first problem of insufficient data, I plan to train the system with more data and more labels (total labels is 17).
2. In order to solve the batch size problem, there are two potential approaches: Firstly, I will try to train on a more powerful machine with more batch size. Secondly, I will reduce the number of parameters using dimension reduction methods such as PCA during the training and testing.
3. In the future work, I will implement and improve the *figure* stream by using skeleton sequences rather than single frame's skeleton data, these improvements will bring the network the ability of extracting temporal features and thus improve the recognition efficiency.
4. I will try to train and test the system with more dataset such as volleyball dataset to make the system have more generalization ability.
5. To solve the problem of imbalance of data, I could adopt methods such as Online Hard Example Mining (Shrivastava et al., 2016), SMOTE: Synthetic Minority Over-sampling (Chawla et al., 2002) and few-shot learning methods to improve the system performance.

Critical analysis of the relationship between theory and practical work produced

During the project, I deeply understand the reciprocal relationship between theory and practice of the system. The theory behind the idea of the project is from the psychology term: Figure Ground Perception. It states that humans perceive the world by figure and ground. After I came up with the idea, I did a lot of research on how to transplant the Figure Ground Perception idea onto computer vision and video understanding areas, which includes related work of video understanding network, human body pose estimation, skeleton data recognition, sports event recognition and group activity recognition etc. Then, I started with two general ideas: video features extraction methods and skeleton based action recognition methods corresponding to ground stream and figure stream respectively. In the meanwhile, I also found a suitable dataset, SoccerNet-v2 as my training and testing datasets. However, during the implementation procedure, I realized some problems of my proposed system:

1. Although skeleton based human action recognition methods such as STGCN (Yan et al., 2018), have excellent results on many human action recognition tasks. It barely detects the human skeleton on SoccerNet-v2 dataset due to poor video resolution and motion blur problem. Therefore, I improved the model by using the object detection system first and extracting the skeleton from the detected bounding box.
2. Beside this problem, I also found that skeleton data cannot be input to the figure stream because every scene has a different amount of players but the ConvNets require the same amount of input data. To solve this problem, I implement a skeleton data normalization system that only extracts the same amount of skeleton data and feeds it into the network.

During the training process, I also encountered some problems, such as insufficient GPU memory, slow I/O speed, overfitting, low generalization, and unbalanced labels problems etc.

1. Because processing video data requires machines to have large memory (typically, video has one more dimension), during the training process, it takes around 4 hours to train on one epoch on my machine. This is unexpected and resulted in no extra time for further fine-tuning in the end.
2. Also, slow I/O speed results in slow training time as well. We solve this problem by preprocessing the training data from video to image sequences which highly improve the data processing speed.
3. As for the overfitting problem, I used dropout, data augmentation and such approaches to minimise overfitting.
4. I tried this trained model on another dataset to test the generalization ability of the system, however the results are not ideal, this is another question that I did not consider during the research stage.
5. Furthermore, I realize the unbalanced labels problem in the end and will eliminate it in the future by using techniques such as over-sampling, few-shot learning and SMOTE ,etc.

In conclusion, the proposed two-stream network for group activity recognition can be further improved to make sure the system has more recognition precision, accessibility, robustness and generalizability.

Awareness of Legal, Social Ethical Issues and Sustainability

A fundamental principle for research is to make sure it is legal to use. To avoid this problem, When I requested access to the SoccerNet-v2 dataset (Deli`ege et al., 2021) , I filled out a consent form to ensure that the dataset is only used for research and not for any commercial use. Also, I declared that the dataset and videos are protected by copyrights and I will not share the dataset with anyone else. I agreed and signed the Disclosure Agreement.

Research ethics are important in a research project. It ensures the public can trust the research and the funder can be confident; it supports important moral and social values, for example, not to harm others physically or mentally.

In this project, because I did not collect the data by myself, therefore one of the benefits of it is the availability of data. SoccerNet-v2 is one of the largest soccer game datasets in video understanding and sport understanding research. The datasets have been reviewed several times for legal and social ethical issues. Thus, there are no human participants to conduct any interview,observations, experiment or process any sensitive personal information, and I confirm that the university is not responsible for this study and I am not receiving any funding for this project.

In terms of sustainability, the proposed system in this project is eco-friendly, it does not require any external electronic equipment.

Conclusion

In this essay, the methodology adopted in the project is introduced, the strength and weakness of the project are discussed, the possibilities of the further work are demonstrated, the relationship between theory and practical work are discussed and ethical, legal and sustainability problems are discussed. The proposed two-stream network achieved promising results, however, during the research, implementation and experiment stages, there are also some problems that can be improved, and the proposed system could be more practical.

K. Koffka. Principles of Gestalt psychology. Routledge, 2013.

S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph con- volutional networks for skeleton-based action recognition, 2018.

A. Deli`ege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. V. Droogenbroeck. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos, 2021.

J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional net-works, 2015.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research , 16: 321–357, Jun 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://dx.doi.org/10.1613/jair.953.

A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining, 2016.