Machine Learning and Artificial Intelligence
J.-L. Kim (Ed.)

doi:10.3233/FAIA220435

Ball Position Feature Embedded Group Activity Recognition Model for Team Sport Games

Ankhzaya JAMSRANDORJ* [a], Vanyi CHAO* [b], Yin May OO [b], Kyung-Ryoul MUN [c] and Jinwook KIM [c,1]

[a] *Department of Human Computer Interface & Robotics Engineering, University of Science & Technology, South Korea*
[b] *Department of AI Robotics, University of Science & Technology, South Korea*
[c] *Center for Artificial Intelligence, Korea Institute of Science and Technology, South Korea*

Abstract. Most group activity recognition models focus mainly on spatio-temporal features from the players in sports games. Often they do not pay enough attention to the game object, which heavily affects not only individual action but also a group activity. We propose a new group activity recognition model for sports games that incorporates players' motion information and game object positional information. The proposed method uses a transformer encoder for temporal feature extraction and a 'simple' conventional convolutional neural network for extracting spatial features and fusing them with the relative ball position-embedded features. The experimental results show that our model achieved comparable results to state-of-the-art methods on the Volleyball dataset by using only one transformer encoder block and the ball position.

Keywords. Group activity recognition, ball positional encoding, transformer network, deep learning

1. Introduction

Group activity recognition aims to identify an overall movement in a multi-person scene. It has many practical applications, such as surveillance, sports video analysis, and social behavior understanding. To understand the scene of multiple persons, the model needs to describe how the persons' actions change temporally and how they interact with each other. Recognizing the group activity in sports videos is one of the critical topics. However, we can not consider only players' actions and how they interact, but every related object can also play an essential role in recognizing group activity for sports games.

For example, in a volleyball game, when a player on the left side jumps to perform a *spike*. Without the ball position, this can be confused with the player on the right side

[1] Corresponding Author: Principal Researcher, Korea Institute of Science and Technology, Seoul, South Korea; E-mail: jwkim@imrc.kist.re.kr.
* Equal contribution

trying to block the ball. Thus, the model may predict the group activity as a *'right spike'* instead of a *'left spike'*. This can be improved by learning the relations of the ball to the players.

We are inspired by the recent progress of group activity recognition models, specifically, the Actor-Transformers [1], which uses a transformer network that has emerged as a superior method for natural language processing tasks. However, the model does not learn any features of the related object in sports games. We expect to extend the work by utilizing the ball position as the related object in the game scene.

Our method devised a new way to encode the ball position in this work. We conducted the experiments on the well-known Volleyball dataset, and our model achieved comparable results to the state-of-the-art models by using only one transformer encoder with the ball position. Besides, our model is more straightforward and less expensive than the state-of-the-art model. We performed multiple experiments to find a way to combine the ball features with the main stage for group activity recognition. These can be considered our contributions.

2. Background

Multiple research communities have extensively studied group activity recognition due to its wide applications. Earlier methods have primarily relied on a combination of handcrafted features [2,3,4] for each actor with probabilistic graphical models. With the rapid development of deep learning, deep convolutional neural networks (CNNs) [5] and recurrent neural networks (RNNs) [6] have significantly improved group action recognition performance due to the understanding of temporal context and high-level information.

In recent works, graph neural networks (GNNs), which inferred graph-structured data, were applied to learn relations between individuals. [7] built an actor relation graph using a 2D CNN and graph convolutional networks (GCNs) to capture person interactions on a spatio-temporal graph. Later on, several works improved the previous fully-connected graph to a criss-cross one or a dynamic graph when modeling relations and aggregating features to reduce computational costs.

On the other hand, some recent works utilized the self-attention mechanism to selectively highlight actors and group relations without explicitly building any graph. [1] introduced the transformer with an I3D network to capture the temporal evolution and spatial interactions. [8] and [9] designed a cluster attention mechanism for better group informative features with the transformers. However, these methods are designed for general tasks, not only for sports group activity recognition. Therefore, they did not consider the ball positions, which can be an essential clue in group activity recognition for ball sports.

The following works explored the ball trajectories for identifying group activities in sports videos. [10]. introduced an architecture called Group Interaction Relational Network (GIRN), which operates directly on the joints' spatial coordinates of each actor and the ball positions in the scene, leveraging this representation to infer the interactions among different individuals. The GIRN used a multi perceptron layer to extract relational features from a series of coordinates, such as the joints and the ball positions. Then the attention mechanisms were applied to understand the higher importance of the relations from key individuals with a more significant or distinguishable contribution to

the collective activity. Similarly, [11]. proposed a group activity recognition approach called Pose Only Group Activity Recognition System (POGARS), designed to use only the pose data of each actor and the ball position. POGARS utilized 1D CNNs to learn the spatiotemporal dynamics of actors involved in the group activity. Different from them, we used RGB instead of pose information, which requires an explicit model. Besides, instead of using the ball position coordination directly like in the previous works, we used and embedded the distance between the ball and every actor to learn key actors for group activity recognition.

3. Model Architecture

Our model takes the 1280×720 RGB images, players' tracklets, and ball tracklets as input. It consists of 2 branches, the main branch and the sub-branch, as shown in Figure 1.

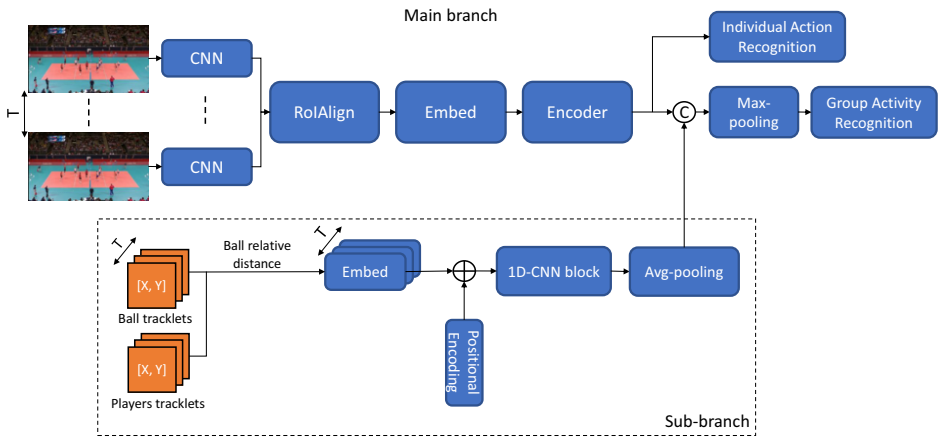

Figure 1. Our network architecture. Ball tracklets and players tracklets represent the position of the ball and center of the player's bounding box, respectively. The Encoder block represents the encoder part of transformer network.

3.1. Main branch

For the main branch, we hypothesize that the RGB stage of the actor-transformer model [1] provided a good enough model that can be used or tweaked. As stated by the author, 3D CNNs are computationally expensive. Thus, instead of 3D CNNs, we used 2D CNN to extract the features of every input image sequence and stacked the output features together before passing them through the RoIAlign [12]. Finally, the rest of the blocks can be successfully used without additional tricks or tweaks. For CNN blocks, we use ResNet-18, which is pre-trained on ImageNet, to extract the features from the input image for the CNN blocks.

3.2. Sub-branch

In our model, we utilize the ball position to improve the accuracy of the group activity recognition. [11] attempted to add the ball features to the network. However, we argue that only the ball coordinates $[x, y]$ in the image coordinate system have no relation to the players. We conduct our experiment with two types of ball positional encoding. 1, we encode the ball position directly [11]. 2, we encode the relative distance between ball tracklets and players' tracklets.

$$d(P_{players}, P_{ball}) = \begin{bmatrix} ||P_1 - P_{ball}|| \\ ||P_2 - P_{ball}|| \\ \vdots \\ ||P_N - P_{ball}|| \end{bmatrix}, \tag{1}$$

where $d(P_{players}, P_{ball}) \in \mathbb{R}^{N\times 1}$ $(N = 12)$ is the euclidean distance of the ball tracklets to players' tracklets. $P_{players} \in \mathbb{R}^{N\times 2}$ is the players' tracklets, and $P_{ball} \in \mathbb{R}^{1\times 2}$ is the ball tracklets.

We calculate the ball relative distance by using the euclidean distance shown in Eq. (1). $d(P_{players}, P_{ball})$ of each frame is fed into each linear layer accordingly to generate a descriptive feature embedding for each frame which will stack together (dimension: $\mathbb{R}^{T\times 64}$). We add the positional encoding of the frame position to the feature embedding before feeding it to the 1D CNN Blocks.

The 1D CNN block takes $\mathbb{R}^{T\times 64}$ as input to produce a temporally convolved feature representation $\mathbb{R}^{T\times 1024}$ followed by an average pooling to remove the temporal dimension $\mathbb{R}^{1024}$. The 1D CNN block contains three 1D convolutional layers followed by batch normalization and ReLU with skip connection, explained in [11]. We use the stack of four such a block with a fully connected layer at the end. We concatenate the features from RGB stage $\mathbb{R}^{N\times 1024}$ with the features from the ball stage $\mathbb{R}^{1024}$ to produce the features whose dimension is $\mathbb{R}^{(N+1)\times 1024}$.

Our model performs multi-task learning by predicting group activity and individual action labels. We attach the individual action block after the transformer encoder block and the group activity block after the max-pooling of the concatenated features. We use a linear layer for both individual action recognition and group activity recognition block.

4. Experiments

4.1. Dataset

We use the Volleyball dataset introduced by [13]. The Volleyball dataset was created from publicly available YouTube videos. The 4830 video clips were manually trimmed from the collected 55 video matches and belong to eight group activity classes, including *set, spike, pass and win point (with left and right)*. Each video clip contains 41 frames, where the middle frame is labeled with group activity and individual action labels. There are nine individual classes such as *waiting, setting, digging, falling, blocking, jumping, standing*, and *moving*.

For our experiment, we also use the dataset split technique suggested by [13]. In addition, we use the extended version of the dataset introduced by [10]. They manually annotated the 2D ball location for every frame. We use the data augmentation technique to reduce model overfitting by horizontally flipping the training samples and the labels accordingly.

4.2. Implementation details

4.2.1. Experimental setup

The experiment is carried out by evaluating our model on the well-known Volleyball dataset [13]. The 2D coordinates of the ball were manually annotated and obtained from [10]. We choose nine frames as the input, four before and four after the middle frame. The input frames have a resolution of 1280×720 and three channels (RGB). The middle frame contains group activity labels and individual action labels. All input frames include ball coordinates $[x, y]$ in the image coordinate system.

4.2.2. Multi-task loss function

We use the cross entropy loss function for group activity recognition and individual action recognition. We use the weighted loss technique to deal with the imbalanced dataset of individual action classes by assigning weights to each loss function of individual classes. We set the weights to $[1, 1, 2, 3, 1, 2, 2, 0.2, 1]$ for *blocking, digging, falling, jumping, moving, setting, spiking, standing and waiting*, respectively.

We also perform multi-task learning for group activity recognition and individual action as described in [11]. To train our model, the losses of the group activity and individual action recognition tasks are optimized simultaneously, as shown in Eq. (2).

$$\mathscr{L}_{total} = \mathscr{L}_{GA} + \alpha \mathscr{L}_{IA}, \tag{2}$$

where $\mathscr{L}_{GA}$ and $\mathscr{L}_{IA}$ represent group activity loss and individual action loss, respectively. The weight α was set to 1.2 to obtain the optimum accuracy for group activity recognition.

4.2.3. Training

We implemented the experiment using the PyTorch framework. During our experiment, the multi-task loss function and Adam optimizer were used by setting the initial learning rate to 0.0001, which will be reduced by half when the validation loss remains plateau for three epochs. We trained our model for 40 epochs with a batch size of 8.

4.3. Experiment Results

We first performed an ablation study of our approach on the Volleyball dataset to validate the effectiveness of various schemes of ball position encoding for group activity recognition. Two different ball encoding schemes, which were introduced in Section 3.2, were investigated. As shown in Table 1, both variations of ball position encoding outperformed the base model by 0.3% and 0.4% for individual action, 0.4% and 1% for group activity recognition. This demonstrated that the ball position could be an essential indicator of

identifying group activities in sports videos. We achieved the best performance with the relative ball position encoding scheme.

Next, we compared our best model with the state-of-the-art approaches on the Volleyball dataset in Table 2 using accuracy metrics for individual action and group activity recognition. Compared with [1] trained with only RGB, the performance of our method is increased by 1.9%, up to 93.3% by simply adding ball position encoding. Our model achieved promising performance using only RGB frames, ResNet-18 backbone, and the ball position compared to some state-of-the-art methods with computationally high backbones and additional optical flow and pose input.

Table 1. Our model comparison result with different settings, evaluated on Volleyball dataset.

Method	Backbone	Individual Action	Grouop Activity
Our - base model	ResNet-18	83.0	92.3
Our - with ball positions	ResNet-18	83.3	92.7
Our - with ball relative distance	ResNet-18	**83.4**	**93.3**

Table 2. Comparison of our model against different baselines and state-of-the-art method.

Models	Backbone	Modalities	Individual Action	Group Activity
HDTM [13]	AlexNet	RGB	∼	81.9
CERN [14]	VGG-16	RGB	69.1	83.3
stagNet [15]	VGG-16	RGB	∼	89.3
SSU [16]	Inception-v3	RGB	81.8	90.6
AFormer [1]	I3D	RGB	∼	91.4
ARG [7]	Inception-v3	RGB	83.0	92.3
DIN [17]	ResNet-18	RGB	∼	93.1
GIRN [10]	∼	Pose	∼	88.4
AFormer [1]	∼	Pose	∼	92.3
POGARS [11]	∼	Pose	∼	93.2
AFormer [1]	I3D	RGB+Flow	83.7	93.0
JLSG [6]	I3D	RGB+Flow	83.3	93.1
Gformer [8]	I3D	RGB+Flow	84.0	94.9
Dual-AI [9]	Inception-v3	RGB+Flow	**85.3**	**95.4**
CRM [18]	I3D	Pose+RGB+Flow	∼	93.0
AFormer [1]	I3D	Pose+RGB+Flow	85.9	94.4
GIRN [10]	∼	Pose+Ball tracklets	∼	92.2
POGARS [11]	∼	Pose+Ball tracklets	∼	93.9
Ours	ResNet-18	RGB+Ball tracklets	83.4	93.3

The confusion matrix of our model on the Volleyball dataset is shown in Figure 2. Our model achieved an accuracy of over 90% for every group activity except the *'right set'* class. The model mostly confused between *'set'* and *'pass'* activities, which can be due to the actors' involvement in the activity, because the ball is passed between actors in both of these activities.

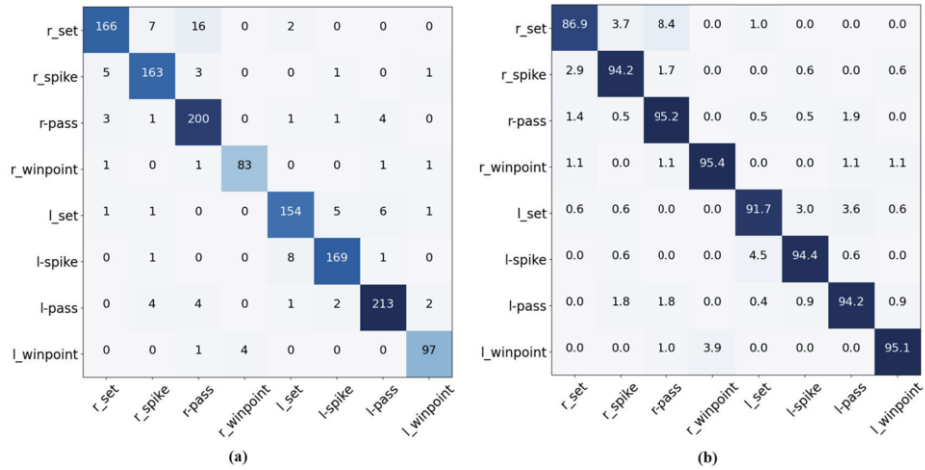

Figure 2. Confusion matrices for group activity recognition. (a) is the confusion matrix of sample, and (b) is the confusion matrix of accuracy.

5. Conclusion

We proposed a straightforward architecture with a single transformer and the ball position encoder for group activity recognition in ball sports. Using only RGB frames and the ball tracklets, our model achieved competitive results on the Volleyball dataset. Experiments have shown that the ball position could be influential in recognizing sports group activity. In the future, we plan to explore more efficient ways of ball position encoding for group activity recognition in ball sports.

Acknowledgement

This work was supported by Athletes' training/matches data management and AI-based performance enhancement solution technology Development Project (No.1375027374).

References

[1] Gavrilyuk K, Sanford R, Javan M, Snoek CG. Actor-transformers for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 839-848); 2020.

[2] Amer MR, Todorovic S. Sum product networks for activity recognition. 2015, IEEE transactions on pattern analysis and machine intelligence, 38(4), 800-813.

[3] Amer MR, Lei P, Todorovic S. Hirf: Hierarchical random field for collective activity recognition in videos. In European Conference on Computer Vision (pp. 572-585). 2014 September, Springer, Cham.

[4] Amer MR, Todorovic S, Fern A, Zhu SC. Monte carlo tree search for scheduling activity recognition. In Proceedings of the IEEE international conference on computer vision (pp. 1353-1360); 2013.

[5] Li K, Wang Y, Zhang J, Gao P, Song G, Liu Y, Qiao Y. Uniformer: Unifying convolution and self-attention for visual recognition. 2022, arXiv preprint arXiv:2201.09450.

[6] Ehsanpour M, Abedin A, Saleh F, Shi J, Reid I, Rezatofighi H. Joint learning of social groups, individuals action and sub-group activities in videos. In European Conference on Computer Vision (pp. 177-195). 2020 August, Springer, Cham.

[7] Wu J, Wang L, Wang L, Guo J, Wu G. Learning actor relation graphs for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9964-9974); 2019.

[8] Li S, Cao Q, Liu L, Yang K, Liu S, Hou J, Yi S. Groupformer: Group activity recognition with clustered spatial-temporal transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 13668-13677); 2021.

[9] Han M, Zhang DJ, Wang Y, Yan R, Yao L, Chang X, Qiao Y. Dual-AI: Dual-path Actor Interaction Learning for Group Activity Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2990-2999); 2022.

[10] Perez M, Liu J, Kot AC. Skeleton-based relational reasoning for group activity analysis. 2022, Pattern Recognition, 122, 108360.

[11] Thilakarathne H, Nibali A, He Z, Morgan S. Pose is all you need: The pose only group activity recognition system (POGARS). 2021, arXiv preprint arXiv:2108.04186.

[12] He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969); 2017.

[13] Ibrahim MS, Muralidharan S, Deng Z, Vahdat A, Mori G. A hierarchical deep temporal model for group activity recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1971-1980); 2016.

[14] Shu T, Todorovic S, Zhu SC. CERN: confidence-energy recurrent network for group activity recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5523-5531); 2017.

[15] Qi M, Qin J, Li A, Wang Y, Luo J, Van Gool L. stagnet: An attentive semantic rnn for group activity recognition. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 101-117); 2018.

[16] Bagautdinov T, Alahi A, Fleuret F, Fua P, Savarese S. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4315-4324); 2017.

[17] Yuan H, Ni D, Wang M. Spatio-temporal dynamic inference network for group activity recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7476-7485); 2021.

[18] Azar SM, Atigh MG, Nickabadi A, Alahi A. Convolutional relational machine for group activity recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7892-7901); 2019.