Temporal Segmentation of Subgroup Activities in Basketball Trainings

Metehan Kaya

Master's thesis

Temporal Segmentation of Subgroup Activities in Basketball Trainings

Metehan Kaya

December 11, 2024

Chair of Data Processing
Technische Universität München

Metehan Kaya. *Temporal Segmentation of Subgroup Activities in Basketball Trainings.* Master's thesis, Technische Universität München, Munich, Germany, 2024.

Supervised by Priv.-Doz. Dr. habil. Hao Shen and Nicolai von Hoyningen-Huene (Ph.D), submitted on December 11, 2024 to the Department of Electrical and Computer Engineering of the Technische Universität München.

Chair of Data Processing, Technische Universität München, 80290 München, Germany, http://www.ldv.ei.tum.de/.

Abstract

Most of the research on Group Activity Recognition (GAR) takes all individuals in the scene into consideration and gives a single prediction per sample.

Who is interacting with whom is not predicted.

There are many cases where people in the scene should be split into different groups, even sub-groups.

Almost all datasets have videos where models take videos, optical flow and human keypoints as input.

Our paper focuses on adapting sub-grouping problem to basketball players in the practice sessions by training GNN-based models using 2D positions of players and shot events.

Define temporal segmentation

Contributions: Different models and datasets

Contents

1 Introduction

1.1 Motivation

Lack of research on GAR and datasets with multiple groups. Mention papers on Group Activity Recognition (GAR) for scenes with a single group. Quote from the paper "GAR Review Paper" claiming that "Real videos can have multiple groups to perform different group activities to detect and track" and "Researchers have already conducted studies on detectingindividual video activity. In contrast, there needs to be more video data sets and practical models for group activity detection problems in group activity research.". Emphasize that most of the datasets have a single group label per sample, like Volleyball dataset.

[Maybe, put an image from a GAR dataset where each sample has only one label]

Lack of datasets on a single sports area. Mention datasets where samples are from different sports areas instead of a single one. Mention a few datasets which focus on only one sports area, but focused on another problem, not GAR.

Lack of datasets with positions of people instead of video itself. Many datasets can be mentioned here.

Mention the challenge of splitting a group of players doing a match into 2 opponent subgroups. It is harder compared to the problem associated with other datasets like Collective Activity Dataset.

1.2 Contribution and thesis structure

This thesis is structured as follows:

Chapter 2: all background information crucial for the experimental methods.

Chapter 3: different types of datasets. a dataset with scenes having at most one group with more than one player. a dataset with scenes that can have multiple groups with more than one player.

Chapter 4: methods developed for both datasets

Chapter 5: experiments on both datasets and results

Chapter 6: conclusion and further research to improve the accuracy.

2 Background

Before presenting our methods, let's take a look at input types, relevant models and spectral clustering.

2.1 PoTion heatmaps

Mention PoTion heatmaps and the way frames from different timestamps are embedded into a single heatmap using color channels. Mention formulas, normalization and stuff.

Put keypoints and corresponding PoTion representation, with different number of color channels.

2.2 Multi-modal models

Multi-modal models: Models with multiple streams. Goal is to take inputs of different types and boost the accuracy.

Mention "SlowFast Networks for Video Recognition", one branch operating at low frame rate, another branch operating at high frame rate.

Mention "Two-Stream Convolutional Networks for Action Recognition in Videos" which takes vide and optical flow in different streams. Then mention "Three-Stream 3D/1D CNN for Fine-Grained Action Classification and Segmentation in Table Tennis"

Put visualization of a model from either "SlowFast Networks" or "Three-Stream 3D/1D CNN".

2.3 Action segmentation

2.3.1 Datasets on action segmentation

COIN, The Breakfast Actions Dataset, 50 Salads Dataset, GTEA Dataset

Put a visualization of a sample from 50 Salads Dataset.

2.3.2 Models on action segmentation

TCN + Transformer based models

Put a visualization of an action segmentation model.

2.4 Group Activity Recognition (GAR)

Emphasize the fact that most of GAR papers focus on models which outputs a single group label per sample. Mention papers using Volleyball and CAD datasets, like "Learning Actor Relation Graphs for Group Activity Recognition". Also say that each individual can have a label but still there is only one group. It is also observed in the combined loss where there is no loss term for interaction between individuals.

Put a visualization of a GAR model.

2.5 GNNs

Mention graphs, nodes, edges and what they represent respectively, based on the area it is used. Give examples from the areas where GNNs are used.

2.5.1 GCNs

Show how convolutional operations are adapted to GCNs. Show some formulas behind conv operations. Mention node classification, edge classification and link prediction.

2.5.2 GATs

Show how attention mechanisms are adapted to GATs to capture relation between nodes. Mention what attention values assigned to edges represent. Don't forget that values are normalized and a matrix of attention values is not necessarily symmetric. Mention the areas that use GATs.

Put a visualization of GCN and GAT next to each other.

2.6 Spectral Clustering

Mention what spectral clustering is and the areas it is used. Indicate the relation between our problem and spectral clustering.

2.6.1 On spectral clustering: analysis and an algorithm

Formulas and stuff. Emphasize that number of clusters is given.

2.6.2 Self-Tuning Spectral Clustering

Improvements on spectral clustering compared to previous paper. Emphasize that a new algorithm is defined in this paper to obtain the optimal number of clusters.

If there is a visualization of comparison based on same input, copy it.

2.7 GAR with social grouping (or call it Social GAR)

Focus on the paper "Joint Learning of Social Groups, Individuals Action and Sub-group Activities in Videos". Emphasize that the author extends the dataset CAD to social-CAD to have multiple group labels. Explain combined training loss and its components. Emphasize how training and test phases differ by mentioning attention values in GAT and spectral clustering.

Put a visualization of the GAT-based model implemented in the paper.

3 Datasets

Mention that there are two datasets consisting of sessions from two different tenants and explain the difference between: maximum number of non-singleton groups.

3.1 Data pipeline

Emphasize that positions are not given directly. A tracking algorithm is run on sessions and projection on camera coordinates are applied to get real world coordinates where the origin is center of the court. Also mention that results of tracking algorithm is not fully correct, a player can have multiple track IDs and a track ID can correspond to multiple players.

1. Prepare unlabeled data
    a) Prepare positions from tracks captured by a single camera
    b) Merge positions and shot events for a single camera
    c) Merge unlabeled data coming from different cameras, if there are
2. Prepare labeled data
    a) Prepare annotation data having player, subgroup, activity phase, frame interval relation
    b) Prepare annotation data having player, track ID, frame interval relation
    c) Merge unlabeled data with annotations
    d) Normalize the position coordinates to the range [0, 1]

Show all this process with a figure.

3.2 Heimstetten sessions

Scenes from Heimstetten sessions can have at most one non-singleton group.
Show labels with some pictures.

3.3 X sessions

Sessions from tenant X have scenes with multiple groups. Claim that the models based on this dataset will be more complex.

Show labels with some pictures.

4 Models

Before going deep into the models, explain how heatmap inputs are constructed first. Categorize heatmaps in different ways like grayscale vs potion heatmap or heatmap showing single player vs multiple players.

Mention that there are different configurations for both grayscale and potion heatmaps. Also say shooting players can be visualized in different way compared to non-shooting players.

Put a scene of single player and corresponding PoTion heatmap. Put a scene of multiple players and corresponding PoTion heatmap.

4.1 Models on Heimstetten sessions

Repeat that Heimstetten sessions have at most one non-singleton group and therefore there is no need to a stage which handles splitting players into more than one non-singleton groups.

4.1.1 Double-stage model

The first stage of the double-stage model

The first stage is for phase classification and therefore it is CNN-based multi-class classification model where the loss is a simple cross-entropy loss. Also mention optimizer used in training, etc.

Plot the model by showing different streams it can take.

The second stage of the double-stage model

The second stage is for social grouping to split a non-singleton group into two subgroups which are opponent teams playing game-phase in our case.

The model is GNN-based social grouping model where the loss is calculated by using ground-truth relation between players and weights assigned to edges in GNN where each node corresponds to a player. Also mention optimizer used in training, etc.

Since social groping is a harder problem compared to classification, it needs a longer history to capture the information. So, heatmap inputs of second stage is slightly different from the first stage.

There are two types of GNNs used: GAT and GCN. Weights assigned to edges are attention values for GATs whereas those weights correspond to link prediction values for GCNs. Since we have no information about players, we define a complete graph for each GNN type.

For GAT, emphasize that training and testing phases are slightly different. Aim of the training phase is to adjust learnable parameters of GATs in the model so that attention values assigned to edges between nodes which correspond to players in the same sub-group will have higher value whereas attention values assigned to edges between nodes which correspond to players in different sub-groups will have less value. When it comes to testing phase, focus on spectral clustering and specify that it is totally okay to set the number of clusters to 2.

For GCN, link prediction problem is considered during training phase and results of those values are used in testing phase by applying spectral clustering again.

Again, it is crucial to note that spectral clustering is only used in the testing phase and there are different algorithms to do that.

Plot the model for each type.

4.1.2 Single-stage model

Input of the single-stage model is similar to input of the second stage of double-stage model. Since the model has to output both phase prediction and information (attention value in GAT or score in link prediction), there is a need to define combined loss.

Unlike the second stage of double-stage model, we don't know the number of groups now. Therefore, all spectral algorithms are not applicable here. This issue make our problem more challenging.

Again, it is crucial to note that spectral clustering is only used in the testing phase.

Plot the model by showing different streams it can take.

4.2 Models on X sessions

Repeat that X sessions can have multiple non-singleton groups because now there can be multiple courts or players in game-phase using only half of the court.

4.2.1 Triple-stage model

The first stage of the triple-stage model

Classification model (all players)

The second stage of the triple-stage model

Grouping model (only game-phase, etc.)

The third stage of the triple-stage model

Sub-grouping model (only one group)

5 Experiments

5.1 Experiments on Heimstetten sessions

We have results of a few experiments but the most important fact to mention is results of double-stage model is better than single-stage model and the reasons behind it: spectral clustering into only two subgroups (more algorithms are applicable) and no need to adjust alpha coefficient that is in front of term for social grouping loss in the combined loss.

5.2 Experiments on X sessions

No experiment yet but I don't expect it to have higher accuracy than experiments on Heimstetten sessions.

6 Conclusions

Background information is given, etc.

Methodologies are listed, etc.

Experiments and results are shown, etc.

Comments on the experiments...

Future study: better spectral clustering algorithm, a model where there is no need to do spectal clustering using weights assigned to edges in GNN, a model which captures more complex information like opponent sub-groups.