Reasoning over multiple human-human interaction activities

Perez, Mauricio Lisboa

2021

Perez, M. L. (2021). Reasoning over multiple human-human interaction activities. Doctoral thesis, Nanyang Technological University, Singapore. <https://hdl.handle.net/10356/151926>

<https://hdl.handle.net/10356/151926>

<https://doi.org/10.32657/10356/151926>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 18:59:48 SGT

Reasoning over Multiple Human-Human Interaction Activities

Mauricio Lisboa Perez

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2021

Reasoning over Multiple Human-Human Interaction Activities

Mauricio Lisboa Perez

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

7 Jul 2021

.....

Date

.....

Mauricio Lisboa Perez

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

7 Jul 2021

.....

Date

.....

Prof. Alex C. Kot

Authorship Attribution Statement

This thesis contains material from 3 paper(s) published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Chapter 3 is published as:

[a] Mauricio Perez, Jun Liu, and Alex C. Kot. Interaction Recognition Through Body Parts Relation Reasoning. IAPR Asian Conference on Pattern Recognition (ACPR), 2019. DOI: 10.1007/978-3-030-41404-7_19.

[b] Mauricio Perez, Jun Liu, and Alex C. Kot. Interaction Relational Network for Mutual Action Recognition. IEEE Transactions on Multimedia (TMM), 2021. DOI: 10.1109/TMM.2021.3050642

The contributions of the co-authors are as follows:

- I designed the initial method. The method was improved by discussions with Prof. Alex C. Kot and Dr. Jun Liu.
- I conducted the experiments.
- I prepared the manuscript drafts. The manuscripts were revised by Prof. Alex C. Kot and Dr. Jun Liu.

Chapter 5 is published as:

[c] Mauricio Perez, Alex C. Kot, and Anderson Rocha. Detection of Real-world Fights in Surveillance Videos. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2019. DOI: 10.1109/ICASSP.2019.8683676

The contributions of the co-authors are as follows:

- I designed the initial method. The method was improved by discussions with Prof. Alex C. Kot and Prof. Anderson Rocha.
- I managed the data collection and annotation.
- I conducted the experiments.
- I prepared the manuscript drafts. The manuscripts were revised by Prof. Alex C. Kot and Prof. Anderson Rocha.

7 Jul 2021

.....

Date

.....

Mauricio Lisboa Perez

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Alex C. Kot, for giving me the opportunity to pursue the Ph.D. in such a renowned institution as NTU, and for providing me important guidance throughout this journey. I am also grateful to Prof. Anderson Rocha for enabling this opportunity by recommending me to Prof. Kot, and also for his support in the early stages of my doctorate studies.

I wish to thank all of those that I have met during my time in ROSE lab, including the Ph.D. colleagues, interns and staff. In special, I want to express my deepest gratitude to Abrar, Kamila, Raman, and Yasin, for their warm welcoming and companionship during my studies. I am thankful as well to Jun, for working together with me in essential projects carried out during my Ph.D. research.

I feel extremely grateful for all of the amazing people I have encountered around NTU and Singapore, which I had the chance to create deep and meaningful connections with. Their support throughout this period, by sharing so many joyful moments and lending me an ear when necessary, was invaluable for me to endure my studies. They are also responsible for many happy memories, that I will carry out with me for the rest of my life.

At last, I would like to thank those that have supported me from so far away. I have a profound gratitude to my dear friends, in Brazil or spread across the globe. Their loyal friendship, despite the long distance, was paramount to give me strength to accomplish this thesis. I am also deeply grateful to my family, for being so understanding with me living geographically so distant during this period. Overcoming the physical barrier, their loving feelings sent through messages, pictures and videos, always warmed my heart and gave me peace of mind to carry on.

“As a well spent day brings happy sleep, so a life well spent brings happy death.”

—da Vinci, Leonardo

To my loving dog Bazinga, forever in my happiest memories

Abstract

Humans are naturally social, engaging in a wide range of interactions with other individuals. These interaction activities can have multiple people involved or just two individuals, and comprise mutual-actions with diverse degrees of complexity and distinct intentions. Therefore, it is still very challenging for machines to autonomously identify them. In this thesis, we explore and propose different techniques to reason about human interactions under distinct scenarios, and exploit multiple types of data.

Person-person mutual action recognition (also referred to as interaction recognition) is an important research branch of human activity analysis. Current methods in the field – mainly dominated by Convolutional Neural Networks (CNNs), Graph Convolutional Networks (GCNs) and Long Short-Term Memory (LSTMs) – often consist of complicated architectures and mechanisms to embed the relationship between the two persons on the architecture itself, to ensure that the interaction patterns can be properly learned. Our main contribution in the first work of this thesis is to propose a simpler yet very powerful architecture, named Interaction Relational Network (IRN), which utilizes minimal prior knowledge about the structure of the human body. We drive the network to identify by itself how to relate the body parts from the individuals interacting. In order to better represent the interaction, we define two types of relationships, leading to specialized architectures and models for each relationship type. These multiple relationship models will then be fused into a single and special architecture, in order to leverage both streams of information for further enhancing the relational reasoning capability. Furthermore we define important structured pair-wise operations to extract additional information from each pair of joints – distance and motion. Ultimately, with the coupling of an LSTM, our proposed method is capable of paramount sequential relational reasoning. These important extensions for our network can also be valuable to other problems that require sophisticated relational reasoning. Our solution is able to achieve state-of-the-art performance on the traditional interaction

recognition datasets SBU Kinect Interaction and UT-Interaction, and also on the mutual actions from the large-scale dataset NTU RGB+D. Furthermore, it obtains competitive performance in the NTU RGB+D 120 dataset interactions subset.

Research on group activity recognition mostly leans on the standard two-stream approach (RGB and Optical Flow) as their input features. Few have explored explicit pose information, with none using it directly to reason about the persons interactions. As our second work in this thesis, we leverage the skeleton information to learn the interactions between the individuals straight from it. With our proposed method Group Interaction Relational Network (GIRN), multiple relationship types are inferred from independent modules, that describe the relations between the body joints pair-by-pair. Additionally to the joints relations, we also experiment with the previously unexplored relationship between individuals and relevant objects (e.g. volleyball). The individuals distinct relations are then merged through an attention mechanism, that gives more importance to those individuals more relevant for distinguishing the group activity. We evaluate our method in the Volleyball dataset, obtaining competitive results to the state-of-the-art, even though using a single modality. Therefore, our experiments demonstrate the potential of skeleton-based approaches for modeling multi-person interactions.

CCTVs have since long been used to enforce safety, for example to detect fights arising from many different situations. But their effectiveness is questionable, because they rely on continuous and specialized human supervision, thus requiring automated solutions for improved efficiency. However, previous works in the field of fight detection are either too superficial (classification of short-clips) or unrealistic (movies, sports, fake fights). Moreover, none has performed detection of actual fights on long duration CCTV recordings. In this thesis' third work, we tackle specifically this problem. Firstly we propose CCTV-Fights, a novel and challenging dataset containing 1,000 videos of real fights, with more than 8 hours of annotated CCTV footage. Then we propose a benchmark pipeline, on which we assess the impact of different feature extractors (Two-stream CNN, 3D CNN and a local interest point descriptor), as well as different classifiers (such as end-to-end CNN, LSTM and SVM). Results confirm how challenging the problem is, and highlight the importance of explicit motion information to improve performance.

Contents

List of Figures	xix
List of Tables	xxi
Acronyms	xxiii
1 Introduction	1
1.1 Background	1
1.2 Challenges and Motivations	4
1.3 Proposed Methods	7
1.4 Thesis Contributions	9
1.5 Outline of the Thesis	10
2 Literature Review	13
2.1 Related Methods	13
2.1.1 Action Recognition	14
2.1.2 Human Interaction Recognition	15
2.1.3 Group Activity Recognition	17
2.1.4 Fight Detection	20
2.1.4.1 Fight Recognition in Trimmed Videos	22
2.1.4.2 Fight Localization in Untrimmed Videos	24
2.1.4.3 Summary Table	25
2.2 Related Datasets	26
2.3 Related Concepts	31
2.3.1 Multilayer Perceptron	31
2.3.2 Relational Network	33
2.3.3 Long Short-Term Memory Network	34
2.3.4 Attention Mechanism	35
2.3.5 Pose Estimation	36
3 Interaction Relational Network for Two-person Action Recognition	39
3.1 Introduction	39
3.2 Interaction Relational Network	44

3.2.1	Relational Network Overview	45
3.2.2	Skeleton Joints as Independent Objects	46
3.2.3	Relationships Architectures	46
3.2.4	Pair-wise Structured Input	49
3.2.5	Sequential Relational Reasoning	50
3.3	Experiments	51
3.3.1	Implementation Details	51
3.3.2	Ablation Experiments	52
3.3.3	State-of-the-art Comparisons	55
3.3.4	Qualitative Analysis on NTU RGB+D 120	59
3.4	Summary	59
4	Skeleton-based Relational Reasoning for Group Activity Analysis	63
4.1	Introduction	63
4.2	Group Interaction Relational Network	66
4.2.1	Overview	67
4.2.2	Relational Modules	69
4.2.3	Auxiliary Individual Modules	71
4.2.4	Attention Mechanisms	72
4.3	Experiments	74
4.3.1	Implementation Details	74
4.3.2	Comparison with Previous Works	76
4.3.3	Ablation Study	78
4.3.4	Evaluation of Inter-Person Connectivity and Number of Joints	80
4.3.5	Robustness Evaluation	82
4.3.6	Qualitative Analysis	84
4.4	Summary	86
5	Detection of Real-world Fights in Surveillance Videos	87
5.1	Introduction	87
5.2	CCTV-Fights Dataset	89
5.3	Methodology	92
5.3.1	Feature Extraction	92
5.3.2	Frame/Snippet Prediction	94
5.3.3	Segment Generation	95
5.3.4	Integrated Pipeline Implementation Details	96
5.4	Experiments	98
5.4.1	Evaluation Metrics	98
5.4.2	Experimental Results	101
5.5	Qualitative Analysis	102
5.6	Summary	103
6	Conclusion and Future Work	107
6.1	Conclusion	107

6.2	Future Work	109
6.2.1	Improved Pair-wise Relational Reasoning	109
6.2.2	Multiple Groups with Independent Interactions	110
6.2.3	Pose Estimation in Surveillance Videos	111
 List of Author's Awards and Publications		 113
Bibliography		115

List of Figures

1.1	Example of dyadic human interaction	2
1.2	Illustration of 3D capturing devices	3
1.3	Example of group activity	3
1.4	Illustration of surveillance systems	4
1.5	Example of shaking hands interaction	5
1.6	Example of skeletons from group activity	6
1.7	Example of human-object interactions	7
1.8	Examples of fights captured by surveillance devices.	8
2.1	Illustration of Bidirectional Recurrent Neural Networks (BRNNs), proposed by Du et al. [1] ©2015 IEEE.	15
2.2	Illustration of Spatial-Temporal Graph Convolutional Network (ST- GCN), proposed by Yan et al. [2] ©2018 AAAI.	15
2.3	Illustration of method proposed by Ji et al. [3] ©2014 IEEE.	16
2.4	Illustration of Actor Relation Graph (ARG), proposed by Wu et al. [4] ©2019 IEEE.	18
2.5	Illustration of PMH heatmaps, proposed by Chen and Lai [5] ©2019 Springer.	19
2.6	Multiple violence and fight examples explored in the literature	21
2.7	Samples from SBU dataset [6] ©2012 IEEE.	27
2.8	Samples from UT-Interaction dataset [7] ©2009 IEEE.	28
2.9	Sample frames with two-persons interactions from NTU RGB+D dataset [8] ©2016 IEEE.	29
2.10	Sample frames with two-persons interactions from NTU RGB+D 120 dataset [9] ©2019 IEEE.	30
2.11	Samples frames from Volleyball dataset [10] ©2016 IEEE.	31
2.12	Illustration of a Multilayer Perceptron (MLP)	32
2.13	Illustration of visual question answering architecture with Relational Network (RN)	33
2.14	Illustration of a Recurrent Neural Network (RNN)	34
2.15	Illustration of an Attention Mechanism proposed by Luong et al.	35
2.16	Illustration of OpenPose method proposed by Cao et al.	36
3.1	Interaction Relational Network (IRN) overview	41
3.2	Illustration of the IRN architecture for human interaction recognition	44

3.3	Simplified view of fusion IRN	48
3.4	Confusion matrices for SBU dataset	54
3.5	Performance per interaction class	60
3.6	Confusion matrices for NTU RGB+D 120 dataset	61
4.1	Summarized Group Interaction Relational Network (GIRN)	65
4.2	Overview of our proposed method GIRN	68
4.3	Attention mechanisms examples. The blue, yellow and red colors of the bounding boxes indicates the contribution weight from lowest to highest, and the white color indicates the reference player. (a) During a spike activity, the group attention mechanism gives more weight to the individual spiking, moderate weight to the blockers, and less weight to the remaining individuals. (b) For an individual spiking, the individual attention mechanism gives more weight to the blockers and less weight to the remaining individuals. (c) For an individual blocking, the individual attention mechanism gives more weight to the opponent spiking, moderate weight to the teammate also blocking, and less weight to the remaining individuals.	73
4.4	Examples of estimated poses and annotated ball coordinates from Volleyball dataset	75
4.5	Subdivision of players positions in the volleyball court. The position of a player is based in the row (Front or Back) and column (Exterior or Middle) he/she is located.	81
4.6	Confusion matrices for our methods variations using pose-only and pose plus volleyball.	84
4.7	Examples of video clips from the Volleyball dataset wrongly classified by $GIRN_{intra+inter+object}$	85
5.1	Frames sampled from CCTV-Fights, showcasing the challenging diversity of its fight scenes and conditions.	88
5.2	Sample frames from existing violence and fight datasets	90
5.3	Summarized Proposed Pipeline for Fight Detection	93
5.4	Illustration of the two-stream approach	94
5.5	Illustration of a joint 3D-CNN and RNN architecture	95
5.6	Illustration of TRoF detector and descriptor main elements	95
5.7	Illustration BoVW-based framework used with TRoF	96
5.9	Example of a video ground truth segment and possible prediction segments	100
5.10	Examples of fight segment predictions generated by 2-tiered Temporal stream	104
5.11	Sample frames from segment cases of correct prediction and false alarm	105

List of Tables

2.1	Summary of approaches on Fight Recognition and Localization. . .	26
3.1	Ablation experiments on SBU dataset.	53
3.2	Comparison of our results with previous work on SBU.	55
3.3	Results from previous approaches and our proposed methods on the UT dataset per set.	56
3.4	Results from our proposed methods on the subset of NTU RGB+D dataset, containing only the mutual action classes.	57
3.5	Results from our proposed methods on the mutual-actions subset of NTU RGB+D 120 dataset.	58
4.1	Comparison of our results with previous works on Volleyball dataset.	77
4.2	Ablation experiments on Volleyball dataset. Accuracy performance of our proposed GIRN with different relationships types and cumulative addition of components.	79
4.3	Evaluation of Inter-Person relationship type with different players connectivity degrees and varying number of joints per subsampling.	82
4.4	Evaluation of robustness against misplaced ball localization. Zero-mean Gaussian generated noise is added to the ball coordinates, with variable standard deviation values indicated by the displacement noise in terms of pixels.	82
4.5	Evaluation of robustness against ball miss-detections. Ball coordinates at each frame is randomly set to zero according to the dropout chances indicated.	82
5.1	Summary of some of the most relevant datasets commonly used for Fight Detection.	91
5.2	Summary of CCTV-Fights statistics. Value in parentheses indicates average number of fight instances per video.	91
5.3	The results for the chosen methods on the CCTV-Fights dataset. First column indicates the feature used and second which classifier was used for training and prediction.	101
5.4	Results for the Two-Stream approach, separately and combined, using directly the CNN output as classifier.	102
5.5	Performance of the Temporal stream, with CNN as classifier, split by training model and type of data source.	102

Acronyms

ACC	Accuracy
BoW	Bag-of-Words
BoVW	Bag-of-Visual-Words
CCTV	Closed-Circuit TeleVision
CNN	Convolutional Neural Network
DTrack	Dense Trajectories
FPS	Frames Per Second
FV	Fisher Vectors
GCN	Graph Convolutional Network
GIRN	Group Interaction Relational Network
IRN	Interaction Relational Network
kNN	k-Nearest Neighbors
LSTM	Long Short-Term Memory
mAP	mean Average Precision
MLP	MultiLayer Perceptron
OF	Optical Flow
PCA	Principal Component Analysis
STIP	Space-Temporal Interest Points
SIFT	Scale-Invariant Feature Transforms
RGB	Red Green Blue image channels
RNN	Recurrent Neural Network
RN	Relational Network
SVM	Support Vector Machine
ViF	Violence Flows

Chapter 1

Introduction

1.1 Background

Humans are highly social beings, thus interacting with each other is a natural part of our daily lives. Consequently, a machine capable of automatically reasoning over human interactions would be an invaluable tool with potential usage over many applications. Some possible applications consists of: to enhance human-computer interaction in robotics; to detect suspicious or inappropriate behaviors for surveillance purposes; to generate statistics and analyses at sports; for retrieval or labeling in content management. However, humans can engage in interactions with a wide range of complexity, going from a simple handshake between two individuals, to chaotic brawls involving multiple persons at once. Unsurprisingly, reasoning over human interactions it is still a very challenging task for machines.

A branch of the existing works targets exclusively at dyadic human interactions, which involve just two individuals (e.g. shaking hands, hugging, exchanging objects). Early proposed techniques [7, 11–14] follow a common framework based on hand-crafted local features and the bag-of-words paradigm. These works often use SVM for the final classification. Subsequent methods [15–18] rely on Convolutional

Neural Networks for feature extraction, in combination with Recurrent Neural Networks for higher-level sequential inference. Majority of these works use RGB videos as input. Sometimes, it couple with motion information from the extracted optical flow [19].

FIGURE 1.1: Example of dyadic human interaction: punching. Sample frames (top row) and poses (bottom row) from SBU dataset [6] ©2012 IEEE.

In the broader field of action recognition (comprising actions performed by a single individual also), many works [1, 20–24] make use of skeleton data as input, instead of RGB videos. Skeleton-based information provides an explicit human pose representation, containing the 2D/3D spatial coordinates of multiple body joints (see Figure 1.1). Human activity analysis using such representation becomes increasingly popular, since it is robust against common complications due to illumination variations, cluttered backgrounds and different viewpoints [1, 21].

The skeleton information is typically extracted from videos containing depth data. These videos are captured with specific devices such as Microsoft Kinect and Azure Kinect (see Figure 1.2). However, due to the positive results of skeleton-based research, there are also reliable solutions [25, 26] for estimating the poses from regular 2D videos. These solutions further extend the applicability of research based on skeleton data. In spite of the advantages and availability from such

¹https://commons.wikimedia.org/wiki/File:Azure_Kinect.jpg

FIGURE 1.2: Illustration of 3D capturing devices, commonly used for extracting skeleton information from the individuals being recorded. Microsoft Kinect v2 (left) and Azure Kinect ¹(right).

representation, human-human interaction reasoning directly over pose information is mostly unexplored by existing works.

FIGURE 1.3: Example of group activity: right volleyball team spiking. In the right team, one of the players is actively spiking the ball, while some of the teammates observe and others get into position for a possible rebound. In contrast, the left team is preparing for defense, with the front players reaching up to block the incoming ball. Frames are sampled from Volleyball dataset [10] ©2016 IEEE.

In parallel to research focusing solely on two individuals interactions, there are also many works [4, 10, 27–35] targeting at recognition of activities performed in groups, such as pedestrians queuing or people playing sports (see Figure 1.3). The great majority of proposed methods for group activity recognition also use RGB videos as input. A few of these methods [5, 36, 37] make use of skeleton data as well. However, in most cases, the pose information is only used as an extra or alternative stream for feature extraction. These methods are not specifically designed to reason about the interactions using directly the skeleton information.

Unfortunately, not all human-human interactions are peaceful. At times, people get into disagreements and disputes, with increasing levels of violence that can eventually lead to physical fights. Although there are many proposed methods [38–42] for the recognition or detection of violent events, previous works fail to consider realistic scenarios. Former datasets [38, 39, 43, 44] contain unrealistic characteristics or limitations such as acted fights, high quality videos, favorable viewpoints, short clips and small scale. Thus, these datasets lack examples from long sequences of real-world fights under adverse recording conditions. Adverse conditions that are commonly found in surveillance footage captured by CCTV devices of monitoring systems (see Figure 1.4).

FIGURE 1.4: Illustration of surveillance systems. CCTV cameras (left) and monitor wall of control room $\textcircled{c}\textcircled{i}\textcircled{d}$ ²(right).

1.2 Challenges and Motivations

Many issues arises when seeking to reason over these distinct sorts of human-human interactions. Here we introduce some of the existing challenges that motivate us to propose solutions to these challenges.

(1) Learning human joints relations during interaction

When two or more individuals are interacting, the movements they make are most likely to be related at some extent. For example, when shaking hands, both individuals reach out with their hands to each other (see Figure 1.5). Moreover the

²https://en.wikipedia.org/wiki/File:CCTV_control_room_monitor_wall.jpg

movement does not have to be mirrored. Such as when punching (see Figure 1.1) one individual swings the fist towards the other person, meanwhile the other individual moves the head back to avoid getting hit. Therefore, solely focusing on the individuals poses separately is insufficient to reason over the whole interaction. Learning also the body joints relations between the individuals is paramount for distinguishing more complex interactions.

FIGURE 1.5: Example of shaking hands interaction. Poses sampled from SBU dataset [6] ©2012 IEEE.

In Chapter 3 we introduce our proposed network capable of learning the different human joints relationships types, and use the inferred relations to perform two-person interaction recognition.

(2) Reasoning over group interaction with skeleton-based relations

As the number of individuals in a single interaction increases, the complexity for reasoning about it also grows. See example in Figure 1.6 showcasing an activity that contains a high number of elements (individuals and body joints). At one side, independently focusing at pairs of individuals is not enough, so reasoning over the relations from all of those involved is required. On the other hand, not all of the relations are equally helpful when reasoning about the interaction (e.g. in spike group activity the relations from players spiking and blocking are more relevant). In addition, the unrestrained inference of relations becomes computationally prohibitive when there are too many individuals and body joints.

We present in Chapter 4 our proposed method that addresses the challenges mentioned above. Thus, our method for group activity recognition is capable to leverage the relations inferred from the individuals skeleton information.

FIGURE 1.6: Example of skeletons from individuals performing group activity involving a ball (red filled circle). In this example there is a high number of individuals and body joints, resulting in many possible relations to infer.

(3) Incorporating human-object relations

Some human interactions can involve objects, such as cash, sport balls and weapons. How the individuals interact with the objects is very important to reason about the ongoing human-human interaction. In fact it is crucial to help distinguish interactions that have similar movements. In volleyball for example, both blocking and setting movements have the players raising their hands (see Figure 1.7), but the ball trajectory and speed has more distinguishable characteristics between these two cases. To include the relations from human-object interactions is therefore paramount to enable reasoning over a wider range of human interactions. Specially when the object plays a major role in the group activity, such as in ball sports.

In Chapter 4 we demonstrate how the human-object relations can be inferred by our network, and then fused with human-human relations to perform group activity recognition.

(4) Detecting real fights in surveillance videos

FIGURE 1.7: Example of human-object interactions. Skeletons of players interacting with a volleyball.

Detection of real fights under the surveillance scenario faces many challenging characteristics (see Figure 1.8). In the wild, fights can have any number of people involved (two or more), and might comprise a very broad and distinct collection of actions (e.g., punching, kicking, pushing, wrestling, headbutting). Moreover, surveillance systems usually rely on cameras with low-resolution, positioned at high places facing downwards, and sometimes covering a large view area.

In Chapter 5, we introduce the large-scale dataset collected and annotated specifically for detection of real fights in untrimmed videos captured by surveillance cameras. We also present our proposed pipeline for tackling this problem.

1.3 Proposed Methods

In this thesis, we present three methods for reasoning over distinct human-human interactions. These methods address one or more of the challenges introduced in the previous section. A brief presentation of each method is provided below.

The first method (Chapter 3) targets at recognition of human interactions involving exclusively two individuals. We propose Interaction Relation Network (IRN), an architecture that infer the relations of body joints in a pair-wise manner. Multiple relationship types found in human interactions are explored, as well as efficient

FIGURE 1.8: Examples of fights captured by surveillance devices.

ways to fuse the relations coming from distinct relationships. Important structured pair-wise operations are inserted at the input of our network, for extraction of paramount distance and motion information. To enable long-term sequential reasoning, the IRN is coupled with LSTMs, thus being capable to learn temporal patterns over the inferred relations. This proposed network is evaluated on four challenging datasets, achieving state-of-the-art performance on three of them and competitive results in the remaining fourth dataset.

The second proposed method (Chapter 4) focuses on recognition of human interactions involving multiple individuals, i.e. group activities. It extends the former method into an architecture named Group Interaction Relational Network (GIRN), which enables pair-wise body joints relational reasoning in the more complex scenario that involves more people. An additional relationship type is proposed for this scenario, covering also interactions between human and objects. Action information from each individual is leveraged by our network for refined inferred

relations, therefore also improving the reasoning about group interactions. Our GIRN is equipped with attention mechanisms that allows our network to focus on the relations incoming from key individuals performing the activity. This method obtains competitive results on a dataset of group activities related to sports.

The third method (Chapter 5) tackles the problem of fight detection in untrimmed surveillance videos. We introduce a benchmark pipeline for our proposed dataset of real-world fights, with many videos captured by CCTVs. The pipeline is comprised by three important steps: 1) Feature extraction; 2) Frame/snippet prediction; and 3) Segment generation. Multiple existing methods for steps 1) and 2) are explored, including different combinations among these methods. It is also proposed a training strategy to specialize a model at fight detection in surveillance footage, by first leveraging on non-surveillance fight examples, to obtain a higher performance than just using surveillance data. Experiments on our proposed dataset, following the introduced benchmark pipeline with the selected methods, provide insightful results that indicate the most promising directions for tackling this problem.

1.4 Thesis Contributions

Our main contributions through this thesis are summarized as follows:

- (i) We introduce a method for skeleton-based interaction recognition through the relational network pair-wise paradigm (Chapter 3). Our proposed method is carefully designed to incorporate important features for handling skeleton data and for interaction recognition. Such design includes, explicit extraction of distance and motion information, inference and fusion of multiple relationship types and sequential relational reasoning. Extensive experiments in four datasets demonstrate the efficacy of our approach.
- (ii) We propose a method, based on skeleton data and pair-wise relational reasoning, for recognition of human interactions with multiple individuals involved

(Chapter 4). This method contains suitable auxiliary individual modules, for enhanced relations inference. Attention mechanisms are also proposed, for selectively focusing on the more important individuals. Our experiments indicate the potential of our proposed method.

- (iii) We explore the utilization of human-object interactions to improve the reasoning about human-human interactions (Chapter 4). Relevant object information is generated, with a format matching the skeleton data euclidean space representation. The inferred relations over this information are incorporated by our method for interaction reasoning with multiple individuals. The valuable complementary between the two types of interactions (human-human and human-object) is demonstrated by our experiments and qualitative analysis.
- (iv) We introduce a novel dataset for detection of real fights over surveillance untrimmed videos (Chapter 5). The dataset contains many recordings of fights in the wild, with varying lengths, and with challenging unconstrained proprieties typical to the surveillance scenario. We temporally annotate all the videos, indicating where the different fight segments begin and end throughout the footage duration. Experiments in this new dataset, with a pipeline we propose and with diverse existing techniques, provide a fundamental benchmark for future research in the field.

1.5 Outline of the Thesis

The remainder of this thesis is organized as follows. In Chapter 2, we review the literature related to human-human interactions. In Chapter 3, we introduce our solution for dyadic interaction recognition using skeleton information, namely IRN. In Chapter 4, we present our proposed Group Interaction Relational Network for recognition of activities involving multiple individuals. In Chapter 5, we introduce

a challenging dataset with CCTV footage of fights, and propose a framework for detecting fights in untrimmed videos under this scenario. Conclusively, in Chapter 6 we present our final remarks and potential future work.

Chapter 2

Literature Review

In this chapter, we review the research literature related to human-human interactions analysis. In the first section, we cover the existing methods for recognition and detection of the different types of human activities. In the second section, we introduce the related datasets, that are used for our experiments in this thesis. In the last section, we briefly give more details for some of the concepts utilized in our works.

2.1 Related Methods

In this thesis, we tackle multiple types of human-human interactions activities. Considering this, we group separately the related works reviewed in this section, based on the type of activity they target. We start with human general action recognition. Then we introduce works related to two-person interaction. Subsequently we present methods for group activity analysis. We finish this section with the literature related to fight detection.

2.1.1 Action Recognition

The field of human-human interaction recognition is related to the more general task of action recognition, which also takes into consideration actions performed by a single person. Previous works in action recognition, with methods using solely RGB videos, usually rely on spatio-temporal local feature description [45] or CNN architectures for processing appearance and motion separately [46–48]. Here we will focus on the previous works that use explicit pose information, and evaluate their methods on the SBU interaction dataset.

Most of the latest works using pose information have moved to solutions involving Deep Learning, specially LSTMs [1, 20, 21, 23]. Du et al. [1] design an hierarchical structure using Bidirectional Recurrent Neural Networks (BRNNs), which initially processes the joints separately by body part, then fuses these parts step-by-step at subsequent layers until the whole body is covered (see Figure 2.1). Zhu et al. [20] introduce a regularization in their LSTM to learn the co-occurrence features from the joints – i.e. learn the discriminative joint connections related to human parts. Liu et al. [21] propose a spatio-temporal LSTM architecture, where the network unrolls in the temporal domain with sequential frames, and also in the spatial domain with a pre-defined traversal order of the joints following the human skeleton structure.

More recently, existing works explore Graph Convolutional Networks (GCNs) [49] to learn the patterns in the human pose during an action [2, 50, 51]. Yan et al. [2] propose the Spatial-Temporal Graph Convolutional Network (ST-GCN), by creating a graph with the same body joints connected in the temporal domain as well. Multiple layers of GCNs are applied over this graph to extract the features used for the action classification (see Figure 2.2). Li et al. [51] further extend [2] method by inserting Actional and Structural links in the graph representation of their proposed Actional-Structural Graph Convolution Network (AS-GCN). The structural links are used for connecting the joints of the same body part, while the actional links connect arbitrary long-range joints.

FIGURE 2.1: Illustration of Bidirectional Recurrent Neural Networks (BRNNs), proposed by Du et al. [1] ©2015 IEEE.

FIGURE 2.2: Illustration of Spatial-Temporal Graph Convolutional Network (ST-GCN), proposed by Yan et al. [2] ©2018 AAAI.

Since these works focus on general actions, which mostly are performed by a single person, they do not include in their proposed methods any strategy specifically designed to model the relationship between persons interacting. Therefore, they disregard the additional information coming from the interaction itself, which our experiments in this thesis demonstrate to be invaluable information when dealing with two-person action recognition.

2.1.2 Human Interaction Recognition

Most of early work on human interaction recognition is based on hand-crafted local features, following a bag-of-words paradigm, and often using SVM for the final classification [7, 11–14]. More recent and successful approaches rely on Convolutional Neural Networks for feature extraction, usually combining it with Recurrent Neural

Networks (e.g., Long-short Term Memory) for more complex temporal reasoning over the interactions [15–18].

Some of these approaches allegedly focus on pose [11, 14], or try to embed human structural knowledge to their methods [18]. Vahdat et al. [11] proposed method requires pre-computed person trajectories, initially they employ a pedestrian detector to obtain bounding boxes around the individuals on the first frame, then subsequently apply a tracker to find the trajectory in the later frames. These bounding boxes are used to retrieve the following input information: spatial location, time of the pose and the cropped image within this region. On the work of Raptis et al. [14], the authors have labeled pre-defined poselets on some of the frames. They create templates from these poselets by extracting local-features on the cropped image around these annotations. Based on these poselets features, their max-margin framework would then select key-frames to model an action.

FIGURE 2.3: Illustration of method proposed by Ji et al. [3] ©2014 IEEE.

Another branch of works do use explicit pose information with the purpose of interaction recognition [3, 6, 52, 53]. Yun et al. [6] propose hand-crafted features, based on pre-defined geometric relations between the joints of two individuals on a series of frames, then feed the computed features to an SVM for classification. Ji et al. [3] group the joints that belong to the same body part to create poselets that describe the interaction of these body parts from both individuals. Subsequently, these poselets are used to generate a dictionary representation that will be fed to an SVM for classification (see Figure 2.3). Wu et al. [53] approach this problem using a sparse group lasso on top of features that are based on the individuals joints distance and motion.

One main drawback with most approaches using pose information, is that they try to learn the body joints relations in the interaction through ad-hoc structures that limit the possible relationships. Differently from these approaches, our proposed method in Chapter 3 is not restricted by the human skeleton structure or hand-crafted features. We do not impose fixed connections in our architecture based on the skeleton edge information and the input is invariant to the joint order, therefore our method learns the existing and important relationships by itself.

Moreover, even though some of these works show good performance for recognition of interactions between two individuals, they lack means for handling the complex scenario of group activity. This scenario contains multiple individuals with different levels of contribution to the activity itself. Our proposed method in Chapter 4 seeks to bridge the necessary gaps so that the interaction descriptive capability of the skeleton information can also be applied to multi-person settings.

2.1.3 Group Activity Recognition

Early literature on this topic uses local-descriptors or hand-crafted features to describe the individuals behaviors in space and time. Choi et al. [27] design their own hand-crafted feature based on tracking the individuals with an Extended Kalman

Filter, describing their poses with HOG and then constructing a histogram per individual that takes into account its neighbors and their poses. To capture person-person and group-person interactions, Lan et al. [28] model the scene with a structured latent network, on top of a crafted action context descriptor which uses histograms of local-features such as HOG. Hajimirsadeghi et al. [29] rely on counting the instances of individual actions throughout the clip to define the overall group activity.

Since the popularization of deep learning and neural networks, research in the field of group activity recognition also explores such techniques. Ibrahim et al. [10] propose an Hierarchical Deep Temporal Model, which consists of an Long Short-Term Memory (LSTM) for the group stacked on top of LSTMs per-individual, that have as input CNN features extracted from each player cropped region, throughout sequential frames. Shu et al. [32] use a similar architecture, but they complement the network with LSTMs for the interactions, and they also propose an energy-based classification layer to replace the softmax layer. Bagautdinov et al. [33] proposed method performs the detection and feature extraction at the same time, with a technique based on feeding the frames to a fully-convolutional network which will detect and describe the individuals. These detections and descriptions are then passed to an RNN for the classification of the group activity in conjunction with the individuals action. Also somewhat related to detection, Azar et al. [54] method is based on generating an intermediate representation that maps directly in the frames the spatial location of the individual actions and the collective activity.

FIGURE 2.4: Illustration of Actor Relation Graph (ARG), proposed by Wu et al. [4] ©2019 IEEE.

More recent works keep on using CNN features as input, but focus on inference of the individuals interactions through complex mechanisms over graphs representations of the group. Ibrahim et al. [34] target at learning the players interactions by describing pair-wise relations from matchings defined by hierarchical graphs, then use the descriptions as mid-level representations for each player, before performing the classification with an LSTM at the top. Wu et al. [4] propose constructing multiple graphs to map the appearance and position relation between players, then use a GCN to extract the players representation used for classification of the individual actions and group activity (see Figure 2.4). Lu et al. [35] build a scene graph by connecting players spatially close to each other, then feed this graph to stacked attention blocks. These blocks use graph convolutions to extract person- and group-level interaction representations that will be passed on to Gated Recurrent Units (GRUs) for temporal inference.

There is a branch of works that tries to go beyond the somewhat standardized input based on CNN features coming from the RGB and optical flow modalities, by incorporating also features coming from pose-based representations. Azar et al. [36] use pose estimation techniques to build a pose heatmap for the scene, consisting on single-channel images where the value of the pixel indicates the presence of any body part. The generated image is fed to a CNN for feature extraction. Lu et al. [37] complement their CNN plus GRU pipeline with pose features extracted using an existing spatio-temporal attention mechanism designed for skeleton information. Chen and Lai [5] also use estimated poses to build a heatmap representation and a CNN for feature learning, but instead of utilizing a single channel for all joints, each joint has its own channel indicating its potential locations (see Figure 2.5).

FIGURE 2.5: Illustration of PMH heatmaps, proposed by Chen and Lai [5]
©2019 Springer.

An issue with these approaches is that the pose information is simply regarded as an extra or alternative input feature. Either as another visual modality that can be fed to an CNN, or through an equivalent feature extractor that feeds on the individual skeleton data. The useful abstraction of describing the pose as a set of joints' spatial coordinates is not directly used for inferring the group interactions. Differently than previous approaches, our proposed method in Chapter 4 explores such abstraction to learn and describe the relationships between the joints from multiples individuals and their surroundings. Then our method uses the relationship descriptions to reason about the group activity.

2.1.4 Fight Detection

Terminology in the literature of fight detection may be ambiguous or imprecise, since sometimes the term Violence is used to refer to Fight specifically and sometimes to General Violence [44]. However, general violence can also include blood, explosions, guns, cold arms, car chases, gory segments, etc. Some examples are presented in Figure 2.6. Because of the focus of the work in Chapter 5, the related literature presented in this section refers to Fights, unless explicitly said otherwise.

Another important characteristic to highlight, is that the majority of works treats acted or sports fights [38, 43, 44] as the same as street or real-world fights. But acted fights (such as in movies, TV shows, researchers pretending to fight), usually are much more artistic and contain artificial movements where there is no intention to injure the opponent. Moreover, in movies and TV shows the fight is the focus in the recording, with video shots that benefit the viewer to see clearly all the strikes and motion of the actors (i.e., centralized and without occlusion). Sport scenes are more realistic, but the fight is also centralized in the recording and the athletes have to follow the rules of each sport, thus the fight actions are constrained to a limited range of actions. Real-world fights usually happen and develop differently from the above scenarios, with strikes and blows more crude. And for certain cases, such as with surveillance footage, the fight may be captured from very distinct points of

FIGURE 2.6: Multiple violence and fight examples explored in the literature. First row contains example of general violence (e.g., explosions and blood). Second row shows acted and sport fights. Last row depicts real-world fights, captured from CCTV cameras or Car cameras.

view and be located at the corner or a small region of the recording. Examples of the different types of fight is depicted in Figure 2.6.

Some of the related works here use techniques that also take advantage of the audio information contained on the videos [42, 56–58]. The audio multimedia information

is not utilized in this thesis because there is no audio available at most surveillance cases, so the discussion here focuses on the visual cues utilized by these works.

2.1.4.1 Fight Recognition in Trimmed Videos

The problem of automated violence detection has been of interest for many years. Early works naturally focus on the simpler task of whole video classification, usually recognizing if short clips or trimmed videos contain a fight [56–59]. Still, there are more recent works that also focus on this sub-problem [40, 60, 61].

Older methods employ hand-crafted features that are based on motion and intuition on what are the particularities of violent scenes. Nam et al. [56] target at measuring the dynamic activity of a video clip, by summing the motion sequence of pixels over the frames, with the assumption that violent scenes contain movements both rapid and intense. Datta et al. [59] rely on silhouettes extraction and limbs information to track an individual and calculate its acceleration, assuming that someone who just got hit would have a sudden change of direction and/or speed. These early works are built over strict constraints that does not generalize to the vast possibilities of fights (e.g. fights on the floor or with more than two people). Moreover, these constraints are also true to some non-fight activities (e.g. sports with intense motion). Besides the fact that these constraints are flawed, better techniques to compute and analyze motion are available since then.

With the progress of computer vision techniques, research has moved from less ad-hoc and constrained methods to methods with greater power of generalization, such as feature extraction of local interest points and bag-of-words [38, 39, 55, 61]. Souza et al. [55] also consider motion information to be paramount to correctly recognize fights, so they employ the local spatio-temporal feature detector and descriptor STIP [62]. They use a traditional Bag-of-Visual-Words strategy to aggregate the STIP low-level descriptors, and an SVM for final classification of the video clip. A similar pipeline is utilized by Bermejo et al. [38], but they use MoSIFT instead of STIP for the low-level feature extraction. MoSIFT is an extension of the traditional

SIFT descriptor into the spatio-temporal domain, that incorporates information from optical flows to describe the motion information.

Hassner et al. [39] target a specific type of fight, the fights that happen in crowds. They collect a dataset of crowded scenes, with and without fights, then develop a video descriptor tailored for this scenario: Violent Flows (ViF). ViF descriptor is based on the flow magnitude (estimated via optical flow) change over time, this change is used to determine if there is a significant motion in that pixel location, where the threshold for magnitude change significance is given by the average change in each frame.

Zhang et al. [60] propose an extension of Weber Local Descriptor (WLD) [63] to incorporate motion information, creating MoWLD. The MoWLD descriptions go through an automated feature selection stage using their proposed Kernel Density Estimation (KDE) method. The remaining features are transformed through Sparse Coding then fed to an SVM for final classification of the videos.

Mohammadi et al. [40] also focus on the crowd fight scenario, but they employ an unconventional approach, based on behavioral heuristics inspired by psychological studies. These heuristics are translated into physical equations, which are estimated from the visual information contained in the video and are summarized in a feature named Visual Information Processing Signature (VIPS). These features are then classified through an SVM.

Concerned also with computational performance (speed and low memory footprint), Moreira et al. [61] extend the traditional fast local detector SURF [64] to the temporal domain. This way they can benefit from motion information, but without requiring too much resources and being able to run in real-time. Their proposed method, namely Temporal Robust Features (TRoF), is used to extract low-level features that are processed by the Fisher Vector technique [65] for mid-level generation. The generated mid-level features are then fed to an SVM for decision making.

Xu et al. [66] propose to first spatially detect the potential active regions of interest (ROI), then perform the fight recognition on these regions. The spatial localization of the active regions is done by two distinct means: using an activation map, to detect high movement areas; and applying an object detector, to obtain human proposals. The authors employ a two-stream based 3D CNN with ROI pooling to perform the fight recognition over the potential spatial regions extracted during localization.

2.1.4.2 Fight Localization in Untrimmed Videos

As the techniques for fight detection advance, and the performances on the simple task of fight recognition improve and reach high values (accuracy above 90%), a harder scenario begins to be explored by the literature: Fight Localization [41, 42, 67, 68]. Under this scenario, the problem proposition is to temporally localize the fight(s) in the timeline of the untrimmed videos, instead of just outputting if that entire video sequence contains a fight or not. This is very important for the real-time detection, and also for the retrieval applications, for example, as a law enforcer looking for fight scenes in a surveillance camera footage.

Rota et al. [67] focus on real-life fights using the dataset they introduce, which contains mainly recordings from car cameras that captured people fighting on the streets. Their method is based on the assumption that the features of interest to detect a fight are mostly in the region in between the persons involved, a region the authors refer to as interpersonal space. The interpersonal space exact location and size are estimated on each frame from the bounding boxes of the subjects of interest. Then, Dense Trajectories [45] features extracted from this region are fed to a BoW and SVM pipeline for classification of that frame.

Dai et al. [68] work is the winner of MediaEval 2015 task of Violent Scenes Detection [69], which is a challenge for general violence detection in movies. The authors design a system that consists on fusing many visual features, and also auditory,

to localize the violence in the videos of the challenge. The visual features include HoG, HoF, MBH, STIP, Dense Trajectories, CNNs and CNNs with LSTM.

Similarly to [68], Lam et al. [41] also employ a collection of features for general violence detection, although it is evaluated in the MediaEval 2014 challenge instead of 2015. The main difference is that in [41] the impact of each type of feature is studied independently. This way the features individual contribution and complementary with respect to each other are better measured.

Also evaluating on the MediaEval 2014 challenge, Moreira et al. [42] propose other ways to fuse a more limited range of features, and also employ their low-cost local feature detector and descriptor TRoF [61]. The authors manage to achieve similar performance to Lam et al. [41], but without the use of high computational cost methods, such as Dense Trajectories [45] and techniques based in Deep Learning.

Nasaruddin et al. [70] use a background subtraction method to obtain spatial attention regions on which a fight might be occurring, before performing the fight temporal localization. The spatial attention regions are identified using a bilateral background subtraction approach, which is capable to obtain the areas with moving objects and persons. The pixels outside the attention regions are then blurred, in order to highlight the important areas in the frame. The blurred frames are fed to a 3D CNN architecture that will generate the fight score for the input frames sequence.

2.1.4.3 Summary Table

A summary of fight detection related works is provided in Table 2.1. The related works are subdivided in approaches for recognition (i.e., whole video classification) and localization (i.e., temporal localization of fights).

The literature for Fight Localization is notably scarce, with most of the few existing works focusing on general violence and on movies, instead of real fights footage from surveillance cameras. Which leads to a gap for solutions specific for this

TABLE 2.1: Summary of approaches on Fight Recognition and Localization.

	Reference	Year	Visual Information	Classifier
Recognition	Nam et al. [56]	1998	Hand-crafted motion estimator	Ad-hoc threshold
	Datta et al. [59]	2002	Hand-crafted acceleration estimator	Ad-hoc threshold
	Giannakopoulos et al. [57]	2010	Hand-crafted motion features	kNN
	Souza et al. [55]	2010	STIP + BoW	SVM
	Perperis et al. [58]	2011	Motion features + Luminance variance	kNN
	Bermejo et al. [38]	2011	Motion SIFT + BoW	SVM
	Hassner et al. [39]	2011	ViF (Motion descriptor) + BoW	SVM
	Zhang et al. [60]	2015	Motion WLD + Sparse Coding	SVM
	Mohammadi et al. [40]	2016	Hand-crafted physical equations based on behavioral heuristics	SVM
	Moreira et al. [61]	2017	Fast Motion SURF (TRoF) + FV	SVM
	Xu et al. [66]	2019	3D CNN	Softmax layer
	Akti et al. [71]	2019	CNN + LSTM + Attention	Softmax layer
	Lopez and Lien [72]	2020	CNN + Boost Fusion	Sigmoid function
Localization	Rota et al. [67]	2015	DTrack + BoW	SVM
	Dai et al. [68]	2015	Fusion of local descriptors + FV with CNN + LSTM	SVM
	Lam et al. [41]	2017	Fusion of local descriptors + FV with Deep Learning	SVM
	Moreira et al. [42]	2018	Fusion of local descriptors + FV	SVM
	Nasaruddin et al. [70]	2020	3D CNN	Softmax layer

distinct scenario. Our work introduced in Chapter 5, proposes a novel dataset and a pipeline that targets specifically at detection of real-fights in surveillance videos.

2.2 Related Datasets

For the experiments in this thesis, we use multiples datasets that not only differ by their characteristics, but also by their purposes. Firstly, we introduce the datasets used for two-person interaction recognition: SBU, UT-Interaction, NTU RGB+D and NTU RGB+D 120. Subsequently we present the dataset utilized at our experiments for group activity recognition: Volleyball dataset. For our work with Fight detection in untrimmed videos we use a dataset proposed by us, namely CCTV-Fights. The details for this dataset is presented in Chapter 5.

SBU Kinect Interaction [6] is a dataset for two-person interaction recognition, created using Kinect, providing RGBD data and skeleton information for each video (see Figure 2.7). It consists of eight interactions (approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands), seven different participants (pairing up to 21 unique permutations), on a total of 282 short videos

FIGURE 2.7: Samples from SBU dataset [6] ©2012 IEEE.

(around 2-3 seconds each). Recording was done in a single laboratory environment and with a frame rate of 15 frames per second (FPS). Pose information is provided by means of 3D coordinates over 15 joints per person, at each frame. Coordinates are not entirely accurate, containing noise and incorrect tracking at some cases. We follow the 5-fold cross validation protocol defined by the authors, and report the average accuracy over all folds.

UT-Interaction [7] is also a dataset for inter-person interaction (see Figure 2.8), but different from SBU, it only contains RGB information – i.e. no explicit pose information is provided. To overcome this issue, we use OpenPose [26] to extract the skeleton data, estimating the joints coordinates for the actors in the video. This dataset contains six different classes of actions (shake-hands, point, hug, push, kick, and punch). It was recorded under two different set of conditions, therefore it is subdivided into UT-1 and UT-2, comprising a total of 120 videos, half at each set. Set 1 was recorded on a parking lot, with static background and little camera jitter. Meanwhile, Set 2 was taken on a lawn, with partial occlusion from tree branches and with interference from the wind (camera jitter and moving leaves). Videos

FIGURE 2.8: Samples from UT-Interaction dataset [7] ©2009 IEEE.

have about 2-6 seconds each, and a frame rate of 30 FPS. The authors protocol for evaluation consists of a 10-fold cross validation per set.

NTU RGB + D [8] is a dataset with a wide range of general actions (see Figure 2.9). It is not a dataset exclusively for interaction recognition, however it contains more classes of mutual actions than SBU and UT-Interaction, 11 in total (punch/slap, pat on the back, giving something, walking towards, kicking, point finger, touch pocket, walking apart, pushing, hugging, handshaking). Moreover, this subset with interaction-only classes, contains a total of 10,347 videos, which were collected with the more precise Kinect (V2) device and under more challenging conditions. It includes 40 different subjects and a large variation in viewpoints, by recording at the same time with three cameras at different heights and positions. The length of the videos range from 1 to 7 seconds, with a frame rate of

FIGURE 2.9: Sample frames with two-persons interactions from NTU RGB+D dataset [8] ©2016 IEEE.

30 FPS. The dataset contains the 3D coordinates from 25 joints per person for all frames. For evaluation, the authors proposed two protocols: Cross Subject (CS), on which 20 pre-defined actors are used for training, and the remaining for testing; and Cross View (CV), where two cameras are reserved for training. Although this is not a dataset specifically for Human Interaction Recognition, we believe that experimenting over this dataset's mutual-only classes can be highly valuable at validating our methods because of its characteristics: large scale and more challenging conditions.

NTU RGB + D 120 [9] is an extension from the dataset described above, and it is the largest available dataset for human action recognition with RGBD data and skeleton information (see Figure 2.10). It contains 60 additional classes, of which 15 are mutual-action classes (hit with object, wield knife, knock over, grab stuff, shoot with gun, step on foot, high-five, cheers and drink, carry object, take a photo, follow, whisper, exchange things, support somebody, rock-paper-scissors). In total

FIGURE 2.10: Sample frames with two-persons interactions from NTU RGB+D 120 dataset [9] ©2019 IEEE.

there are 24,794 videos from 26 interaction-only classes, captured similarly as the previous version: using Kinect (V2), with many different subjects and a diverse range of viewpoints. The evaluation protocol is slightly different: Cross-Subject, which is the same as before, with a subset of actors being used exclusively for training and another for testing; and Cross-Setup, where instead of splitting the data according to different cameras, pre-defined setups are selected for comprising the training and testing splits.

Volleyball dataset [10] consists of 4,830 snippets (41 frames each) of group activities from professional athletes playing Volleyball (see Figure 2.11). The annotation comprises labels referring to four different activities (set, spike, pass and winpoint),

FIGURE 2.11: Samples frames from Volleyball dataset [10] ©2016 IEEE.

combined with to which side is performing it (left or right), so a total of eight distinct labels. The dataset authors also annotated the individuals actions (nine in total): waiting, setting, digging, falling, spiking, blocking, jumping, moving and standing. Together with the players action labels, they also provide the bounding box locations at the central frame of the snippet. For evaluation, the authors have separated the data into train, validation and test splits, and report the group activity classification accuracy at the test split.

2.3 Related Concepts

In this section, we provide more details to some of the existing concepts that are utilized by the works in this thesis. The concepts presented here are used directly or indirectly by our proposed methods in the following chapters.

2.3.1 Multilayer Perceptron

The multilayer perceptron (MLP) architecture is one of the most basic types of neural networks [73, 74]. Sometimes the MLP is also referred to as “vanilla” neural

networks. A perceptron is a neuron unit that performs a weighted sum of the inputs, then it outputs a value based on a linear or nonlinear activation function (e.g. sigmoid functions, rectified linear unit ReLU). The neurons are organized by layers in the MLP. Each layer can contain distinct number of neurons, where it can be multiple neurons or even just a single one. The neurons from sequential layers are fully (densely) connected with each other, meaning that every neuron at a certain layer is connected to all neurons in the previous layer and also in the next. An MLP has at least three layers, one for the input, one hidden layer and a top layer for the output, as depicted in Figure 2.12. It is possible for an MLP to have more layers, however the architecture is usually kept shallow (i.e. few layers).

FIGURE 2.12: Illustration of a Multilayer Perceptron (MLP) [74] ©2009 Springer. Each circle represents a neuron. The neurons labeled with X , Z and Y represent the neurons from the input, hidden and output layer respectively.

The MLP can learn by adjusting the weights in each neuron. Such training often happens by feeding many batches of data and using backpropagation [75] to update these weights. This architecture is simple, yet very effective at modeling the weights to the input data. Therefore MLPs are useful at many applications.

2.3.2 Relational Network

Relational networks have been introduced by Santoro et al. [76], targeting to improve relational reasoning over a different range of tasks. The first step of this method consists on reducing the input of the task to a set of independent objects, where each object has its own set of information or features. This set of objects is then fed to a simple architecture that models how the objects relate to each other in a pair-wise manner. The inferred relations from each pair is then pooled into a single description that can be used to solve the task being tackled.

FIGURE 2.13: Illustration of visual question answering architecture with Relational Network (RN), proposed by Santoro et al. [76] ©2017 NIPS. In the RN part of the architecture, the CNN features from each object is paired up with the features from the other objects and with the LSTM features from the question. This triplet is fed to the g_{θ} -MLP for relational inference. The inferred relations are sum pooled and fed to the f_{ϕ} -MLP module to generate the answer.

The authors demonstrate the versatility of these networks by applying their method on three tasks: visual question answering (see Figure 2.13); text-based question answering; and complex reasoning about dynamic physical systems. These tasks cover not only distinct purposes (question-answering and physical modeling) but also different types of input data (visual, textual and spatial state-based).

2.3.3 Long Short-Term Memory Network

The Long Short-Term Memory (LSTM) [77] networks are a specialized type of Recurrent Neural Networks (RNN). RNNs have feedback connections that allow them to transfer information between a series of inputs (see Figure 2.14). The advantage of the LSTMs over basic RNNs is that they contain mechanisms for storing and updating information about recent inputs (short-term), while still preserving some information about earlier inputs as well (long-term). This way, LSTMs are ideal for reasoning about sequential information in videos.

FIGURE 2.14: Illustration of a Recurrent Neural Network (RNN) ©¹. Block A represents an RNN cell, and x_t and h_t respectively represent the input and output at timestep t .

LSTMs have been adapted and improved over the years, resulting in many variations. However, they all follow a similar principle. The LSTM cell stores information in a set of variables usually referred to as Cell State. Information in the cell state is used together with the information from the current input to generate the output for this timestep. To update the information in the cell state the LSTM uses mechanisms often called Gates. These gates regulate which information should be added or removed from the cell state, based in the current input and the cell state itself. Usually a gate is implemented as a trainable neural network layer, so the gate can be trained through backpropagation how to regulate the flow of information.

¹<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

2.3.4 Attention Mechanism

Attention Mechanisms are neural network modules utilized to selectively focus on specific parts of the input set. In early works they are mostly applied for machine translation [78–80]. Where the attention mechanism attributes distinct weights to each word in the source sentence. These weights are then used to generate a context vector for the current output word in the target sequence (see Figure 2.15). The context vector is usually a weighted average, computed over the input features of the words in the source sentence. The set of weights, used for computing the average, is inferred by the attention module while taking into consideration the features from the source words in conjunction with the features from the current target state. This means each target word will have its own set of attention weights.

FIGURE 2.15: Illustration of an Attention Mechanism proposed by Luong et al. [79] ©2015 ACL. Where $\bar{h}_s$ are the words features from the source sentence, h_t are the features from the current target state, a_t are the inferred weights for each source word, c_t is the context vector, $\tilde{h}_t$ are the output features for the current state and y_t is the output target word.

An Attention Mechanism can be defined as follows²:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T) V \quad (2.1)$$

²Definition adapted from Equation (1) of [80]

where Q , K and V are matrices containing a set of queries, keys and values respectively. Softmax [81] is an activation function that normalizes the inputs into values between 0 and 1, and the summation after normalization adds up to 1. Basically, the Attention function is for mapping distinct queries to the input vectors of paired keys and values. Such that the compatibility computed between the query and the key vector is used to assign a weight to the corresponding value vector paired to that key. The output is the weighted sum of the values vectors, given the weights assigned by the computed compatibility after the softmax normalization.

2.3.5 Pose Estimation

Given the success of pose-based methods for multiple tasks, including human action recognition, techniques to estimate human poses from a single RGB image have been proposed [25, 26, 82–84]. These techniques do not require depth information, so they are useful to extract skeleton information from videos captured with common devices, which do not provide depth data.

FIGURE 2.16: Illustration of OpenPose method, proposed by Cao et al. [26]
©2019 IEEE.

OpenPose [26] is a readily available tool³ for pose estimation. The method implemented in the tool uses an CNN to jointly detect the body parts and compute an affinity measurement between them. Subsequently the body parts with highest affinity are assembled together into full body poses (see Figure 2.16). The tool can use models with pre-trained weights also made available by the authors. The most accurate model available is *BODY_25*, it contains weights obtained by training

³<https://github.com/CMU-Perceptual-Computing-Lab/openpose>

on a combined dataset of MPII Human Pose dataset [85] with COCO Keypoint Detection dataset [86].

Chapter 3

Interaction Relational Network for Two-person Action Recognition

In this chapter we present our work focusing on the recognition of human interactions involving solely two individuals. We introduce Interaction Relational Network (IRN), which is our proposed architecture to tackle this problem. Our proposal is based on pair-wise relationships of human joints coming from skeleton information.

3.1 Introduction

Recognition of human interaction (mutual actions) in videos is an important computer vision task, which can help us to develop methods for a range of applications, such as surveillance, robotics, human-computer interface, content-based retrieval, and so on. Although there have been many works during the past decades in human interaction recognition [6, 7, 11, 14, 16, 87], it is still a challenging problem, especially when the videos offer unconventional conditions, such as unusual viewpoints and cluttered background.

Most of the previous works focused on mutual action recognition using RGB videos [7, 14, 16, 87]. These methods comprise approaches going from hand-crafted local features to data-driven feature extraction [12, 13, 15–17]. Some techniques try to implicitly learn the information from the poses of the persons interacting [11, 14]. These techniques often operate over the pixel data, making use of manually annotated regions or poselet labels to focus on these specific regions on the frame or to add extra ad-hoc information. However, they fail to explicitly model the relationships between the interacting body parts of the involved persons, crucial information for interaction recognition.

With the development of more advanced capturing devices that can also capture the depth and extract the pose information from the people in the scene (e.g., Microsoft Kinect [88]), a more accurate and valuable description for the human pose can be used directly. Besides, very recently, techniques that can estimate poses from regular RGB videos have improved significantly [26]. This allows us to apply reliable pose-based methods to mutual action recognition in the RGB videos.

The availability of explicit pose representation of the humans in the videos led to a new branch of interaction recognition techniques focusing on this type of data [3, 6, 52, 53, 89, 90]. These methods usually involve hand-crafted features and a classification technique that theoretically incorporates the different body-parts’ relationships, and then generate a rigid architecture highly dependent on what is believed to be the prior knowledge over the human skeleton structure.

There also exists methods, using pose information, that target general action recognition [1, 20–24, 91–93], i.e. they include actions performed by a single individual as well. Majority of the more recent approaches are based on LSTMs and CNNs. These approaches usually try to better model the spatial and temporal dynamics of the poses during the action through ad-hoc network architectures, specific regularizations or special losses. Although these works achieve promising results on general actions, they lack means of leveraging the relationship information between the poses of the two persons when dealing with interactions. Therefore these

FIGURE 3.1: Interaction Relational Network (IRN) overview. Joints of both individuals are separately paired up, and then fed to independent relation modules. The pair-wise inferred relationships are then aggregated into a global description used for, at last, perform human interaction recognition.

methods underperform in the interaction case.

Recurrent models (RNN/LSTM) and convolutional models (CNN/GCN) with carefully designed and complicated architectures have dominated the field of pose-based (skeleton-based) action recognition for several years. Instead of using recurrent/-convolutional architectures for feature learning, in this work we propose to view the mutual action recognition task as a body-parts relation inferring and learning problem, and thus design a new architecture to automatically handle this task. Our contribution is based on the powerful Relational Network (RN) [76] for implicitly reasoning over the joints relationships present at human interactions. Although simple, RNs proved to be highly efficient at learning what are the important relationships between pair of objects. Santoro et al. [76] proposed RNs as a simple architecture to deal with problems that require relational reasoning, obtaining state-of-the-art performance on tasks such as visual and text-based question answering, as well as understanding the dynamic of physical systems that included motion data. To the best of our knowledge, no one before have developed a method based on Relational Network with the purpose of interaction recognition, nor with explicit body information.

On this work we propose Interaction Relational Network (IRN), summarized in

Fig. 3.1. Since our method is the first attempt inspired by the Relational Networks (RNs) for skeleton-based interaction recognition, our first contribution lies on how to logically rearrange our problem components and characteristics in terms that match the high-level concept behind the original RNs. In other words, on how to use the pose information as input, identifying the existing objects (joints) and their properties (coordinates through time), and how to adequately pair these objects, representing the potential relations relevant for interaction recognition. In our new architecture, instead of using pre-defined joints dependency structures as previous methods, we enable our network to infer by itself the existing relations among specific relationships types. We identified two types of relationships, namely, *intra* relations of body joints from the same person and *inter* relations of the joints from different persons, thus the complex relations of different body parts in the person-person interaction sequence are well represented. The relation modules from the same relationship mapping share the same weights, and will infer from the consciously paired up joints which relations exists between them, and which are important for interaction recognition. The relationship mapping, which defines how the joints are paired up, will determine the type of relationships being inferred by that Relation Module (*intra* or *inter*). The description generated from all the relation modules will then be pooled and fed to a module, which will learn how to interpret this arrangement of relations, and classify the interaction accordingly.

Given the specificity of our task and data, we propose two distinct types of relations, however the original RNs contains a single model regardless of relationship type. Thus another contribution in our work is the design of multi-relationship architectures, merging different relation models into a single end-to-end network and leveraging both types of information, therefore leading to a more accurate recognition. We demonstrate the importance of appropriately choosing a fusion architecture, and initializing it with the right models prior to training. Furthermore, we enrich our method by equipping our new network with structured operations over the object-pair, based on domain knowledge, to extract valuable joint distance and motion information before reasoning about their relationships. In

contrast, the original RNs only contain pair independent operations previously to the relational module. At last, we incorporate an LSTM to our architecture, in order to enable the IRN reasoning over the interactions relationships over a longer duration sequence as well.

We validate our method through extensive experiments on four different datasets. Two of these are traditional datasets for human interaction recognition: SBU [6] and UT [7]. On which we obtain the state-of-the-art performance, close to perfect accuracy. Seeking to further validate our IRN with more challenging data, we propose to use the large-scale datasets NTU RGB + D [8] and NTU RGB + D 120 [9]. Although these datasets were initially created for benchmarking general human action recognition, they contain many classes of mutual actions, in fact more than SBU and UT. Additionally, they have much more samples per class and adverse evaluation protocols. Even though the NTU RGB + D datasets are more difficult, the IRN still outperforms the previous works.

Our contributions in this chapter can be summarized in the following manner:

- A novel approach to Human Interaction Recognition, using the different body parts from the pose information as independent objects, and modeling their relationship pairwise.
- We design effective ways to fuse different types of relationships, leveraging their complementary for an improved performance.
- We augment the input information for our Relation Modules by embedding their architecture with pair-wise mechanisms to explicitly extract relevant extra information from the input pair of joints.
- A new and efficient paradigm for interaction recognition, based on relational networks, as an alternative to the CNN/RNN/GCN architectures that currently dominate the field.

These contributions constitute the novel architecture Interaction Relational Network, which is simple yet effective. We validate on four important datasets that

contains human interactions under different conditions, achieving state-of-the-art performance in three of them and competitive results in the fourth, demonstrating the strength of our technique.

3.2 Interaction Relational Network

FIGURE 3.2: Illustration of the IRN architecture for human interaction recognition. First, we extract the information across frames from each joint separately (j_n). Then we use the set of joints from both persons (P_p) as input to our different architectures, IRN_{inter} and IRN_{intra} . Each architecture models different relationships between the joints, and can be used independently to predict the action. Furthermore, the models can be fused as $IRN_{inter+intra}$, this way using the knowledge from both types of relationships for a more accurate prediction.

We designed an architecture, namely Interaction Relational Network, tailored for human interaction recognition using pose information, inspired by the relational network paradigm of reasoning over pair-wise relationships. Fig. 3.2 contains a visual representation of the proposed method. First we define how to logically re-arrange the skeleton information so it can be used as input to an RN based architecture, then we formulate important relationships mapping as to comprise all relevant relations pertaining to our problem. In order to fully capture the range of relationships types present in the human interaction problem, it was necessary to extend the RN proposal to handle multiple relations models and to fuse them properly. Also, we further extend the architecture with domain-knowledge pair-wise operations over the input at each relation module, allowing the network to explicitly extract information known to be valuable when dealing with pose data:

distance and motion. Finally, we couple our architecture with an LSTM, allowing our method to reason over the interactions during the whole video sequence.

3.2.1 Relational Network Overview

The Relational Network (RN) proposed by Santoro et al. [76] can be simplified through the following equation:

$$RN(O) = f_{\phi} \left(\sum_{i,k} g_{\theta}(o_i, o_k) \right) \quad (3.1)$$

Where O is a set of objects $\{o_1, o_2, \dots, o_N\}$, on which any i^{th} object is represented by an arbitrary $\mathbb{R}^m$ vector containing the properties of that object. The function g , with learnable parameters θ , is responsible by modeling the relationships for each pair of input objects independently, therefore being also referred to as Relational Module. Meanwhile function f , with trainable parameters ϕ , is in charge for reasoning from the merger of the relationships inferred by g_{θ} . It is a versatile formulation which, as long as the objects are fed in pairs to the relational module, can have as input different types of data describing the objects, such as CNN/LSTM features or even physical states.

However, the original RN does not cover the case when well-distinguished relationships can be defined due to the nature of the objects and the chosen pairing up rule, leading to distinct specialized models for each type. Missing therefore good ways to train these models and properly fusing them afterward. Also, in its base formulation the Relational Network only contains an unstructured function for reasoning over the input pair, represented by g_{θ} , lacking a pair-wise structured function that can guide the general relational module with important domain-knowledge information. In the process of tailoring the RN paradigm to our domain, we design important novel extensions to address these issues on our proposed Interaction Relational Network, detailed next.

3.2.2 Skeleton Joints as Independent Objects

The power of the relational networks lies on breaking down the problem into separate interacting objects, and learning specific parameters for describing how each object relate to each other, before merging this information for the classification of the whole. In the case of pose information and action recognition, we can define each joint j_i as an object, using as their low-level feature its coordinates along the frames, together with integers to identify to which joint and body part does it belong: $j_i = (x_1, y_1, x_2, y_2, \dots, x_T, y_T, i, b)$, where x_t and y_t are the 2D coordinates of the joint i belonging to the body part b at the frame t , and T is the desired sampling of frames to be used. As previous work have done [1, 8, 23], we considered five body parts: torso, left hand, right hand, left leg, and right leg. Each person p will therefore have a set of joints for each video, which can be defined as: $P_p = \{j_1^p, j_2^p, \dots, j_N^p\}$, where N is the total number of joints provided by the pose data.

3.2.3 Relationships Architectures

The simplest approach would be to mix and match all the joints, learning a single model with all pairs combinations. However, besides being computationally inefficient, this naive strategy disregards the fact that the joints might come from other individuals, hence the relationship which each other carries different meaning than when the joints are from the same individual. To allow our method to distinguish between the relevant relationships, and therefore better represent human interactions, we propose exclusive relationships mappings, which will be later merged into a single architecture.

Inter-person Relationships. Since we are dealing with interaction recognition, it is desired to map the relation between the joints inter-person. Therefore, all joints from one individual will be paired with all joints from the other individual, bi-directionally, so that it can be order-invariant (in the case of active/passive

actions such as kicking). For that purpose, inspired by the formulation (3.1), we derive the following equation:

$$IRN_{inter}(P_1, P_2) = f_\phi \left(\sum_{i,k} g_\theta (j_i^1, j_k^2) \oplus \sum_{i,k} g_\theta (j_i^2, j_k^1) \right) \quad (3.2)$$

Where f_ϕ and g_θ can be MultiLayer Perceptrons (MLPs), with learnable parameters ϕ and θ respectively. In theory $\sum$ and $\oplus$ can be any pooling operation, such as sum, max, average or concatenate, but from our experiments, we decided to use average because it gives the best results.

Intra-person Relationships. Since the intra-person relationship of the joints can be highly informative as well, we also propose another architecture, where the joints from each person will be paired with the other joints from the same person. For this case there is no need to pair bi-directionally, since the paired joints are from the same person, what would only add unnecessary redundancy to our model – in fact our preliminary experiments demonstrated that it can lead to overfitting in some cases. The pooled output from each individual is concatenated ($\frown$) before going through function f with trainable parameters ϕ' .

$$IRN_{intra}(P_1, P_2) = f_{\phi'} \left(\sum_{i=1}^N \sum_{k=i+1}^N g_{\Theta} (j_i^1, j_k^1) \frown \sum_{i=1}^N \sum_{k=i+1}^N g_{\Theta} (j_i^2, j_k^2) \right) \quad (3.3)$$

Fusing Relationships. Conclusively, we propose an architecture that fuses both types of relationships under the same function f (parameters ϕ''), by concatenating the pooled information from each function g , each with its own parameters θ and Θ :

FIGURE 3.3: Simplified view of $IRN^{fc1}_{inter+intra}$, an alternative way from fusing both types of relationships. On this case, the fusion occurs after the first fully-connected layer ($fc1$) from f_{ϕ} and $f_{\phi'}$, which provides a higher level description for the fusion $f_{\phi''}$.

$$IRN_{inter+intra}(P_1, P_2) = f_{\phi''} \left(\sum_{i,k} g_{\theta}(j_i^1, j_k^2) \oplus \sum_{i,k} g_{\theta}(j_i^2, j_k^1) \curvearrowright \sum_{i=1}^N \sum_{k=i+1}^N g_{\theta}(j_i^1, j_k^1) \curvearrowright \sum_{i=1}^N \sum_{k=i+1}^N g_{\theta}(j_i^2, j_k^2) \right) \quad (3.4)$$

Alternatively, we also propose another fusion architecture: $IRN^{fc1}_{inter+intra}$ (Fig. 3.3). At this design we concatenate the output from f first fully-connected layer ($fc1$), of *inter* and *intra* models (ϕ and ϕ' parameters respectively), before feeding into the fusion $f_{\phi''}$. The motivation behind this design is to allow us to transfer knowledge not only from the relational modules g_{θ} and g_{Θ} , but also from their exclusive f modules. The modules f_{ϕ} and $f_{\phi'}$ should be specialized on the pooled descriptor for its respective type of relationships, hence contributing with a higher level description for the fusion.

3.2.4 Pair-wise Structured Input

One of the advantages of RNs highlighted by Santoro et al.[76] is the flexibility it has to reason over relatively unstructured inputs, for example embeddings from CNNs and LSTMs, as well as it can handle raw data such as state descriptors and, in our case, joints coordinates. Nonetheless, when dealing with structured data it is advantageous to design the architecture with mechanisms to leverage the intrinsic information contained in the data structure. Although this is the case with CNNs and LSTMs used for extracting the embeddings used as input to the RN, they are applied considering only the structure per-object independently. In the case of joints coordinates, the input pair itself is organized in a manner that there is a well-defined structure among them, i.e. coordinates at the same location in the array represents the spatial location of the objects at the same particular time in the video sequence. Therefore it can also be designed architectures to take leverage of the pair-wise structure of the input.

In order to enable the RN to allow this desired feature, but also keeping its flexibility, we propose an extension to function g_θ . We incorporate a function h , which can be carefully designed to model important domain-knowledge pair-wise information. So we can write the improved g_θ , used by IRN' , as:

$$g'_\theta(j_i, j_k) = g_\theta(j_i, j_k, h(j_i, j_k)) \quad (3.5)$$

Inspired by previous work that demonstrated the usefulness of explicit distance and motion features for pose-based action recognition [6, 53], we decided to augment our relational network with modules that can automatically extract this information from each pair of joints. In other words, our h function, to be also used as extra input by g_θ , will explicitly compute vectors of distances ($D(j_i, j_k)$) and motions ($M(j_i, j_k)$) from the input object-pair:

$$h(j_i, j_k) = (D(j_i, j_k) \cup M(j_i, j_k)) \quad (3.6)$$

Considering c_t^i the coordinates of joint i at the frame t , the vector of distances is basically the euclidean distance between the input joints at each frame.

$$D(j_i, j_k) = (\|c_1^i - c_1^k\|, \|c_2^i - c_2^k\|, \dots, \|c_T^i - c_T^k\|) \quad (3.7)$$

And the motion vector is defined as the distance between the joints, but at sequential frames:

$$M(j_i, j_k) = (\|c_1^i - c_2^k\|, \|c_2^i - c_3^k\|, \dots, \|c_{T-1}^i - c_T^k\|) \quad (3.8)$$

3.2.5 Sequential Relational Reasoning

Until this point, our proposed IRN could only reason on the interactions over a short period of time. Defined by the number of frames (T) sampled when assembling the joints information (j_i) for input to our network. In other words, the hyper-parameter T would be analogous to a temporal receptive field of our IRN. Preliminary experiments showed that using a too large value of T is detrimental to performance. It was more effective to skip some of the frames by some step size, as in dilated convolutions [94], in order to increase the temporal range of the input. However, this workaround still does not allow the IRN to use the information available in the whole video sequence, and moreover, it has no means to model the intrinsic order of movements contained in an interaction.

Seeking to enable our method to use information from all frames and reason over the evolution of the interactions over time, we incorporate an LSTM to our architecture. The LSTM module is at the end of our architecture, after f_ϕ and before predicting the action. This architecture, hereinafter referenced as LSTM-IRN, can then leverage the information from the whole interaction sequence before classifying the video. Resulting in an architecture also suitable to model over time the relationships present in human interactions.

3.3 Experiments

3.3.1 Implementation Details

MLPs Configuration. Hyperparameters detailed here were tuned during preliminary experiments. The IRN is implemented as an MLP, where g_θ consists of four fully-connected layers, the first three with 1000 units and the last with 500, and f_ϕ contains three fully-connected layers with 500, 250 and 250 units respectively, with dropout rate of 0.25. The activation function used is the rectified linear unit (ReLU). The LSTM module contains 256 units and also a dropout of 0.25. A softmax layer is placed at the end of the architecture to generate the interactions prediction scores. Training was performed with Adam optimizer, learning rate value of 1e-4 and weight initialization following a truncated normal distribution with zero mean and 0.045 standard deviation for SBU and UT, and 0.09 standard deviation for NTU RGB+D and NTU RGB + D 120. A different standard deviation value is used for the NTU RGB+D datasets because the skeleton information is normalized, as described in [8].

The current implementation of our architecture, in its most complete form (both relationship types and with LSTM), has around 7M network parameters. In contrast, the image-based 2D CNN architectures utilized by Shu et al. [95] (namely AlexNet) and Shi et al. [17] (namely VGG16) have 62M and 138M parameters respectively. And the recent skeleton-based methods employing graphs, ST-GCN [2] and AS-GCN [51], have 3M and 7M approximately.

Training Procedure. To improve generalization during training, we randomly swap the input order between the persons' joints set ($P_1 \rightleftharpoons P_2$). This was significantly beneficial for the IRN_{intra} architecture to avoid bias on the order of the concatenated feature generated after g_θ . $IRN_{inter+intra}$ parameters θ and Θ are fine-tuned from the weights obtained previously by training IRN_{inter} and IRN_{intra} separately, meanwhile ϕ is randomly initialized. We employ this strategy since fine-tuning $IRN_{inter+intra}$ leads to a better performance than training from scratch,

as demonstrated by the ablation experiments in the following subsection. For $IRN_{inter+intra}^{fc1}$, in addition to θ and Θ , we also use the $fc1$ layer and weights from previously trained IRN_{inter} and IRN_{intra} , for initializing the model before training.

Pose Estimation. OpenPose [26] is a readily available toolbox for extracting pose information from RGB videos. We ran it with its default options, which were not the most accurate, but it was faster and sufficiently precise for our experiments. However, because the output is frame-based, we had to apply some post-processing steps to guarantee the consistency of the bodies from Person 1 and 2 throughout the video. It consisted mainly on correctly assigning the pose at each frame to its respective body, based on the distance of the joints between the current frame and previous frames, assuming the overall distance should not be too different between consecutive frames. This is important when there are more than two poses in the frame, due to noise or because of passersby, or when the order of the actors' pose change.

Joints and Frames Subsampling. Although the poses for UT (estimated) and NTU RGB+D (provided) contains 25 joints, we sampled only 15 of them, analogous to what is provided by SBU data. As for the parameter T , regarding the temporal receptive field, we use different values for each dataset. For SBU, since the videos are shorter, we use 8 consecutive frames as a sampling for our input feature. Since UT, NTU RGB+D and NTU RGB+D 120 have longer videos, we use 32 frames. For the experiments without an LSTM we have chosen to sample the central frames, because most likely they contain the more relevant parts of the interaction. For LSTM-IRN we build our input sequence with overlapping frames at each timestep, by sampling T frames of the video with a step of $T/2$ frames, i.e. 8 frames input for every 4 frames step for SBU and 32 frames every 16 for the other datasets.

3.3.2 Ablation Experiments

First, to better evaluate the impact of each part from our proposed methodology, we separately show our results on the SBU dataset, reported in Table 3.1.

TABLE 3.1: Ablation experiments on SBU dataset.

Experiment	Acc
IRN_{inter} (Baseline)	88.7%
IRN'_{inter} (Self-Augmented Input)	93.6%
LSTM- IRN'_{inter}	94.6%
IRN_{intra} (Baseline)	95.4%
IRN'_{intra} (Self-Augmented Input)	95.8%
LSTM- IRN'_{intra}	95.2%
Naive- $IRN'_{inter+intra}$	92.0%
Averaging scores	94.3%
Random- $IRN'_{inter+intra}$	92.5%
$IRN'_{inter+intra}$	96.1%
$IRN'^{fc1}_{inter+intra}$	96.7%
$IRN'^{fc2}_{inter+intra}$	94.9%
$IRN'^{fc3}_{inter+intra}$	95.3%
LSTM- $IRN'^{fc1}_{inter+intra}$	98.2%

Our baseline, using only the central frames coordinates and joint indexing information, already obtains 88.7% of accuracy with IRN_{inter} and 95.4% of accuracy with IRN_{intra} architecture. These results indicates that our approach is able to successfully map the different types of relationships present in the problem of interaction recognition. Incorporating our IRN with mechanisms to take leverage of the pair-wise structure of the input, what allows the network to extract explicit distance and motion information, proved to be highly valuable to IRN_{inter} and it was also helpful to IRN_{intra} , increasing the performance on both cases.

The Naive- $IRN'_{inter+intra}$ experiment consists on simply merging all the relationships mapping (*inter* and *intra*) into a single model relational network. This means only one set of parameters for g and pooling the outputs indiscriminately into the same global descriptor before f . The Naive- $IRN'_{inter+intra}$ approach underperforms, obtaining only 92.0% of accuracy, lower than the specialized IRN'

models. Averaging the scores from these specialized models is better than naively merging the relationships, but it is still worse than using IRN'_{intra} alone.

We were able to truly take leverage of the complementary between the relationships models only through our proposed fused architecture $IRN'_{inter+intra}$. This method maintains the specialized models not only on its design, but also by initially training them separately – random initialization of weights (Random- $IRN'_{inter+intra}$) also underperforms. Furthermore, our $IRN'^{fc1}_{inter+intra}$ approach seems to further correlate and benefit from the complementary of the relationships, obtaining a more significant improvement. However, our attempts to fuse at even higher-levels of the MLPs (fc2 and fc3 layers) was not beneficial, due to higher overfit. Thus fc1 is the optimal layer for fusing the relationships.

Ultimately, we extend the relational reasoning to all available frames and enable reasoning over the interaction sequence by coupling our IRN with the LSTM module. Although leading to higher over-fitting for the *intra* relationship type on this small-scale dataset, which obtains a slightly lower performance than the baseline, the LSTM addition is advantageous for the *inter* variation and, more importantly, when fusing the relationships. With LSTM- $IRN'^{fc1}_{inter+intra}$, our method achieves the best performance in this dataset: 98.2%.

FIGURE 3.4: Confusion matrices for SBU dataset with methods: (a) LSTM- IRN'_{inter} , (b) LSTM- IRN'_{intra} and (c) LSTM- $IRN'^{fc1}_{inter+intra}$.

Fig. 3.4 contains the confusion matrices for all LSTM-IRN architectures. It is interesting to notice how the two relationships models have confusion on different interactions, for example *Inter* have some confusion between *Punching* and

Exchanging while *Intra* does not, and how this confusion is greatly reduced at *Inter + Intra*. More importantly, interaction classes getting confused by both models, such as *Pushing* and *ShakingHands*, are almost entirely distinguished with *Inter + Intra*. This qualitative analysis is a good indication on the capability of our proposed architecture to keep the strength of both types of relationships models, and also to leverage their complementary for distinguishing between even harder cases.

3.3.3 State-of-the-art Comparisons

TABLE 3.2: Comparison of our results with previous work on SBU.

Method	Acc
Yun et al. [6]	80.3%
Ji et al. [3]	86.9%
HBRNN [1] (reported by [20])	80.4%
CHARM [96]	83.9%
CFDM [52]	89.4%
Co-occurrence LSTM [20]	90.4%
Deep LSTM (reported by [20])	86.0%
ST-LSTM [21]	93.3%
VA-LSTM [97]	97.2%
Wu et al. [53]	91.0%
Two-stream GCA-LSTM [23]	94.9%
LSTM- IRN'_{inter}	94.6%
LSTM- IRN'_{intra}	95.2%
LSTM- $IRN'^{fc1}_{inter+intra}$	98.2%

SBU. We report our best results for SBU dataset in Table 3.2, alongside previous work results. Our specialized relationship architectures, LSTM- IRN'_{inter} and LSTM- IRN'_{intra} , can outperform or be comparable to almost all of previous work. After fusion through architecture LSTM- $IRN'^{fc1}_{inter+intra}$ we obtain state-of-the-art

performance, with an advantage over 1 percentage point over the second best: 97.2% with VA-LSTM [97].

TABLE 3.3: Results from previous approaches and our proposed methods on the UT dataset per set.

Method	Acc	
	UT-1	UT-2
Ryoo et al. [7]	70.8%	-
Raptis et al. [14]	93.3%	-
Donahue et al. [15]	85.0%	-
Kong and Fu [98]	93.3%	91.7%
Wang and Ji [87]	95.0%	-
Aliakbarian et al. [16]	90.0%	-
Ke et al. [18]	93.3%	91.7%
Shi et al. [17]	97.0%	-
Shu et al. [95]	98.3%	-
LSTM- IRN'_{inter}	93.3%	96.7%
LSTM- IRN'_{intra}	96.7%	91.7%
LSTM- $IRN'^{fc1}_{inter+intra}$	98.3%	96.7%

UT-Interaction. Table 3.3 shows our results on the UT dataset, together with the results from previous works. The LSTM- $IRN'^{fc1}_{inter+intra}$ obtains a performance equivalent to the current state-of-the-art on the set UT-1, an almost saturated accuracy of 98.3%. On UT-2 set, our architectures outperform the previous methods, and set the new state-of-the-art to 96.7%, which is also very high. For UT-2 there was no improvement from using LSTM- $IRN'^{fc1}_{inter+intra}$, when in comparison to the specialized LSTM- IRN'_{inter} model, it might be the case that for this dataset the interactions are not that complex, therefore not benefiting so much from more advanced relationships modeling. To the best of our knowledge, our work is the first using explicit pose information at UT-Interaction dataset, with the previous works being based on RGB information only. Therefore, our experiments demonstrate that it is also possible to perform high performance human interaction recognition in RGB videos using estimated poses.

TABLE 3.4: Results from our proposed methods on the subset of NTU RGB+D dataset, containing only the mutual action classes.

Method	Acc Mutual Actions	
	Cross-Subject	Cross-View
ST-LSTM [21]	83.0%	87.3%
GCA-LSTM [99]	85.9%	89.0%
Two-stream GCA-LSTM [23]	87.2%	89.9%
FSNET [100]	74.0%	80.5%
ST-GCN [2]	83.3%	87.1%
AS-GCN [51]	89.3%	93.0%
LSTM- IRN'_{inter}	89.5%	92.8%
LSTM- IRN'_{intra}	87.3%	91.7%
LSTM- $IRN'^{fc1}_{inter+intra}$	90.5%	93.5%

NTU RGB+D. Our experiments results over the interaction classes of this dataset are presented in Table 3.4. Both of the specialized architectures are able to match or outperform the majority of the compared methods on the two protocols, including the graph-based method ST-GCN [2]. LSTM- IRN'_{inter} by itself obtains a performance very similar to another recent graph-based method AS-GCN [51], demonstrating the importance of considering the inter-person joints relationships when dealing with mutual actions. Our LSTM- $IRN'^{fc1}_{inter+intra}$ architecture can obtain an even higher performance, going beyond 90% accuracy not only on the Cross-View protocol, but also on the more challenging Cross-Subject, ultimately outmatching all the compared methods in this dataset.

NTU RGB+D 120. The new version of the NTU RGB+D dataset [9] is even more challenging, as shown in Table 3.5 containing our results together with previous work methods. Our proposed LSTM- IRN architectures can still outmatch most of the compared methods, demonstrating they are at some extent scalable with respect to the number of classes. Moreover, as the complexity of the interaction recognition problem grows, more benefit can be obtained with the multi-relationship model LSTM- $IRN'^{fc1}_{inter+intra}$. Claim supported by the fact that at NTU

TABLE 3.5: Results from our proposed methods on the mutual-actions subset of NTU RGB+D 120 dataset.

Method	Acc Mutual Actions	
	Cross-Subject	Cross-Setup
ST-LSTM [21]	63.0%	66.6%
GCA-LSTM [99]	70.6%	73.7%
Two-stream GCA-LSTM [23]	73.0%	73.3%
FSNET [100]	61.2%	69.7%
ST-GCN [2]	78.9%	76.1%
AS-GCN [51]	82.9%	83.7%
LSTM- IRN'_{inter}	74.3%	75.6%
LSTM- IRN'_{intra}	73.6%	75.2%
LSTM- $IRN'^{fc1}_{inter+intra}$	77.7%	79.6%

RGB+D the gain in performance obtained was approximately of 1 percentage point over LSTM- IRN'_{inter} , meanwhile at NTU RGB+D 120 this gain was between 3 and 4 percentage points. Our LSTM- $IRN'^{fc1}_{inter+intra}$ method obtains competitive results to ST-GCN [2], being slightly inferior at the Cross-Subject protocol, but superior at Cross-Setup. The graph-based method AS-GCN [51] have a higher scalability than ours LSTM- IRN , however our better performance in NTU RGB+D [8] indicates the compactness of our LSTM- $IRN'^{fc1}_{inter+intra}$ method make it more suitable when handling less classes. This might be related to the adoption of Graph Convolution (GCN) blocks by AS-GCN, instead of the Multi-Layer Perceptron (MLP) blocks adopted by our approach. Usually MLPs are better at specializing, meanwhile GCN have a higher capacity to generalize. Which characteristic is preferable over the other will be determined by the scenario being targeted (i.e. few or many classes).

3.3.4 Qualitative Analysis on NTU RGB+D 120

A more detailed examination of the performance of LSTM- $IRN_{inter+intra}^{fc1}$, on NTU RGB+D 120 dataset, is shown in Fig. 3.5, which contains the accuracy per interaction class for both protocols. Some interactions are significantly more challenging than others, with a recognition rate much inferior than the average (approx. 78%), e.g. *HitWithObject*, *WieldKnife*, *ShootWithGun*, and *GiveSomething*.

Through the confusion matrices present in Fig. 3.6 it can be seen that, at least for the most severe cases (around or below 70% accuracy), both protocols have confusion mostly between the same interactions. A few notable differences are the confusion between *SupportSomebody* and *KnockOver* for Cross-Subject, and *TouchingPocket* with *PattingOnBack* for Cross-Setup. On both protocols there is a lot of confusion between *HitWithObject* and *WieldKnife*, and these two with *Punch/slapping*. Also, there is some confusion between *GiveSomething* with *Handshaking* and *ExchangeThings*, as well as between *ShootWithGun* and *TakePhoto*. These confusions are reasonable, since we are using only the body parts coordinates information, and these interactions should have similar human movements (e.g. extending the hands to each other), being mainly distinguishable by the object (or its absence) each individual holds. We believe this is a strong evidence that for more advanced interaction recognition, RGB visual information should also be used, so that features about the objects could also be extracted.

3.4 Summary

In this chapter, we proposed the novel Interaction Relational Network for recognition of mutual actions, through an architecture focusing at relational reasoning over the different relationships between the human joints during interactions. Current techniques dominating the field of pose-based interaction recognition are based on CNNs/LSTMs/GCNs and consist of fixed structures determining how to relate each joint. Therefore, given our IRN capacity to learn from the data itself how the

FIGURE 3.5: Performance per interaction class of method $LSTM-IRN'_{inter+intra}^{fcl}$ on the NTU RGB+D 120 dataset, for both protocols (Cross-Subject and Cross-Setup). Total average accuracy of 78.7%.

human body parts relate to each other while two individuals interact, it stands as a highly valuable alternative to these techniques.

Our proposed method obtains state-of-the-art performance on the traditional interaction datasets SBU and UT, and it also obtains the highest performance on the mutual actions subset of the NTU RGB+D dataset, surpassing even recent graph-based methods. Meanwhile achieving competitive results in the interactions subset of NTU RGB+D 120, what indicates it can fairly scale to more classes while being compact. To accomplish such deed, important extensions had to be made over the original Relational Network proposition in order to tailor it for our problem and data. Extensions such as: multiple relationships models; fusion of these multiple relationship models; domain knowledge pair-wise structured operation over the input; and sequential relational reasoning. We believe these extensions can also be

FIGURE 3.6: Confusion matrices for NTU RGB+D 120 dataset with method $LSTM-IRN_{inter+intra}^{fc1}$ for protocols Cross-Subject and Cross-Setup .

valuable to other problems, with specific data and relationship particularities that the original RN it is not capable to suitably model without the proposed extensions.

At its current form, the proposed method operates under the assumption that there is a single action happening in the input video. However, through a simple sliding-window strategy, our method can be applied without modification to classify actions with unknown length in untrimmed videos. Nonetheless, using segmented videos would lead to a better performance, and these can be obtained by employing action proposal methods [\[101–103\]](#).

Chapter 4

Skeleton-based Relational Reasoning for Group Activity Analysis

In our previous chapter we focused on interactions with two individuals only, in this work we move to more complex interactions with multiple persons involved. To tackle this problem we propose Group Interaction Relational Network (GIRN), an architecture also based in pair-wise relationships via skeleton, but with important modifications to handle the higher complexity in recognition of group activities.

4.1 Introduction

Advances in the field of computer vision and machine learning have naturally led researchers to seek solutions for problems with increasing complexity. Starting from recognition of simple actions performed by a single individual [104, 105] (e.g., walking, hand-waving, bending) and subsequently moving to mutual-actions being executed by two persons [6, 7, 106] (e.g., shaking hands, hugging, punching). Then finally targeting at more advanced activities that encompass many actors at

once [10, 27, 30], such as pedestrians queuing or players practicing sports. Reasoning over scenarios with multiple individuals interacting (or not), can be very useful at many applications, for example surveillance, robot-human interaction and sports analysis.

Research in this field has mainly made use of image visual information – i.e. RGB and Optical Flow (OF) – to describe each individual behavior and the interactions among them. Early works propose methods based on local descriptors and hand-craft features [27–29]. Subsequent proposals move to deep-learning based approaches, using Convolutional Neural Networks (CNNs) to extract the low-level features and coupling Recurrent Neural Networks (RNNs) for temporal inference [10, 31–33]. More recent works tackle this problem from a graph-based perspective [4, 34, 35]. They use the graph representation to model the group individuals and connections, then apply techniques such as Graph Convolutional Networks (GCNs) to infer the individuals relationships and interactions. Some of these works do not focus in the interactions at all, while others reason about the interactions at a higher-level in their framework. Therefore, they lack means to learn the interaction relations at a lower-level in the architecture.

Certain authors aspired to complement or replace the commonly used image visual features with explicit pose information from the individuals [5, 36, 37]. However, these works do not directly leverage the pose information when reasoning about the individuals interactions. They basically use the pose as an extra/alternative stream for feature extraction. Even at the two-person scenario there are not many works that explore human skeleton representation with methods specifically tailored for human interactions [3, 6, 53]. Besides, since these proposed methods assume there are only two individuals interacting directly, they lack means to handle the more complex scenario of group activities. Where there are multiple individuals taking part on a distinguishable collective activity, but performing separate actions which can be similar and related to each other or completely different and independent.

FIGURE 4.1: Summarized Group Interaction Relational Network (GIRN) pipeline. Skeleton information from all individuals pass through multiple relational modules, each specialized in a different relationship type. The inferred relations per individual then pass through our attention mechanism, so they can be merged with different weights into a single descriptor for the group relations. Finally, this descriptor is used by our inference module for recognition of the group activity in the scene.

In our work, we propose a method based on skeletons that addresses the aforementioned issues. Our approach leverages the skeleton information for reasoning over multi-persons interactions also. A summarized overview of our method is presented in Fig. 4.1. We introduce an architecture called Group Interaction Relational Network (GIRN), that operates directly on the joints spatial coordinates, leveraging this representation to infer the interactions among different individuals. Through a pair-wise relational network, our architecture can learn different relationships relevant for distinguishing the group activity taking place. By consciously choosing which pairs will our relational modules be exclusively fed with, the modules are guided into specializing for specific types of relationships. One module focuses on learning the relations between joints from a single individual, while another focuses on the relations of joints from different persons. We also try another possible module, which is specialized in the relationship between the human joints and an object of interest (e.g. volleyball). To further improve the quality of the relations inferred by our modules, we couple our architecture with auxiliary models that allows the knowledge from the individuals actions to be distilled into the network, improving also the group activity recognition. The GIRN is also equipped with attention

mechanisms, which allows our inference modules to attribute higher importance to the relations coming from key individuals. These individuals have a greater or more distinguishable contribution to the collective activity.

We validate the effectiveness of our proposed method, for the task of group activity recognition, through comprehensive experiments in the Volleyball dataset [10]. Thus we demonstrate that, when properly leveraged, euclidean space representations can be directly used to reason about the interactions between multiple individuals and also between individuals and objects. Our method obtains a performance better than several multi-modalities works, and comparable to the state-of-the-art.

4.2 Group Interaction Relational Network

Our proposed Group Interaction Relational Network (GIRN) is tailored for the recognition of activities involving multiple persons, on which the persons are performing individual actions that can be either similar or distinct to each other. Inspired by [90], which showed promising results on two-person interaction recognition using skeleton information with a method based on relational networks, we build a novel architecture that can encompass any number of individuals. This architecture equips important modules necessary to tackle the more complex task of group activity recognition.

In this section, we begin by giving an overview of our architecture, explaining the basic concepts behind it. Subsequently we go deeper into the details on how the relational modules operates, and define the different relationship types these modules will be specialized on. We move on to describe the auxiliary modules, which incorporate the individual actions knowledge into the network. Finally we explain the attention mechanisms utilized by our architecture to focus on the relations from the more important individuals.

4.2.1 Overview

Our method, depicted in Fig. 4.2, runs directly over the skeleton information, in other words, human joints spatial coordinates. Therefore the input features for each person in the scene are the collection of its body joints coordinates, throughout the duration of the activity. The coordinates are grouped by joint, hence the features of person p is defined as: $P_p = \{j_1^p, j_2^p, \dots, j_J^p\}$, where J is the total number of joints. Each joint j_i^p is a flat array containing joint i coordinates: $j_i = (x_1, y_1, x_2, y_2, \dots, x_T, y_T)$, where x_t and y_t are the 2D coordinates of the joint i at the frame t , and T is the number of frames. Finally, the complete group input can be written as: $G = \{P_1, P_2, \dots, P_N\}$, for a group with N persons in total.

FIGURE 4.2: Overview of our proposed method. Skeleton information is fed into multiple Relational Modules, each specialized in a different type of relationship. The specialization happens because of the different pairing strategies (illustrated with the joints/objects connections). Relations from the same type are averaged pooled (Σ), then concatenated with the others to generate each individual relations. These relations are then passed on to our Attention mechanism, which will pool them into a single Group Relations descriptor ($\bar{R}$), at last used by our Group Activity Inference module (f_G) to classify the collective activity. The auxiliary individual modules runs in parallel to the group branch, and have a similar architecture. Here the input of the Individual Action Inference module (f_I) is the individual relations concatenated to the attention pooled relations from the other individuals ($\bar{\mathcal{R}}^p$).

The joints of each individual is paired in different ways at our relational module, targeting at inferring multiple relationships types for each individual. Subsequently we pool the individual relationships as indicated below:

$$\overline{\mathcal{R}} = \mathcal{R}_1 \oplus \mathcal{R}_2 \oplus \cdots \oplus \mathcal{R}_N \quad (4.1)$$

where $\overline{\mathcal{R}}$ is the final vector with the pooled individual relationships, $\mathcal{R}_p$ is the relationships vectors for person p and $\oplus$ represents any pooling operation (e.g. max, concatenation). Average pooling is used as the default pooling operation.

The relationships pooled for the group are then fed to our group activity inference module (f_G):

$$GIRN(G) = f_G(\overline{\mathcal{R}}) \quad (4.2)$$

where in our experiments f_G is implemented as a Multilayer Perceptron (MLP).

The output of f_G is connected to a softmax layer to obtain the group labels predictions, so our architecture can be trained through back-propagation using a cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N_G} \sum_g \hat{\mathbf{y}}_{G,g} \log \mathbf{y}_{G,g} \quad (4.3)$$

where $\hat{\mathbf{y}}_{G,g}$ are the ground-truth values for the group activity label, and $\mathbf{y}_{G,g}$ are the predictions computed by the softmax layers. N_G is the total number of group activities.

4.2.2 Relational Modules

The individual relationships used for group activity are extracted with the relational modules described here. These models have the same basic structure g , which follows the conceptual idea of relational networks [76], meaning that each structure has a single pair of objects as input: $g_\theta(o_i, o_k)$, where θ represents the parameters set and (o_i, o_k) is the input pair. For our experiments, g is implemented as a multilayer perceptron (thus θ represents the learnable weights) and the input is

a series of coordinates, such as the joints array j_i . Moreover, to tailor our relational module to our problem, we explicitly extract the distance and the motion information between the input-pair, and concatenate the extracted information with the coordinates before feeding the input to the first fully-connected layer.

To better encompass different relationships types relevant to our problem, distinct pairing strategies are used, with every module having its own set of trainable weights: Θ , ϕ and ω . This way each module can specialize on identifying and describing a specific type of relationship. The inferred pairs relations are then averaged pooled (represented by $\sum$) per relationship type and individual. The strategies adopted (depicted in Fig. 4.2) are defined as follows:

- **Intra-Person:** Relations of the joints intra-person, i.e. pairs the joints from a single individual with each other.

$$\mathcal{R}_p^{intra} = \sum_{i=1}^J \sum_{k=i+1}^J g_{\Theta}(j_i^p, j_k^p) \quad (4.4)$$

- **Inter-Person:** Relations between the inter-person joints, i.e. pairs the joints from a specific individual to the joints from the other individuals.

$$\mathcal{R}_p^{inter} = \sum_{q \in \mathbb{C}_p} \sum_{i=1}^J \sum_{k=1}^J g_{\phi}(j_i^p, j_k^q) \quad (4.5)$$

where $\mathbb{C}_p$ is the set of individuals connected to the person p . Different connection strategies can be applied based on the problem.

- **Person-Object:** Relations between a person joints and an object (e.g. volleyball), i.e. pair the joints of an individual to the coordinates of a specific object.

$$\mathcal{R}_p^{object} = \sum_{i=1}^J g_{\omega}(j_i^p, o) \quad (4.6)$$

where the object coordinates are defined analogously as the joints: $o = (x_1, y_1, x_2, y_2, \dots, x_T, y_T)$.

To use a combination of two or more of the relationships types just described, the relations output vectors are concatenated to generate a new descriptor containing multiple relationships:

$$\mathcal{R}_p = [\mathcal{R}_p^{intra}; \mathcal{R}_p^{inter}; \mathcal{R}_p^{object}] \quad (4.7)$$

4.2.3 Auxiliary Individual Modules

In case the GIRN is being applied on a scenario with well-defined individual actions, where annotations are available during training, our architecture can accommodate the individual label information in order to improve the relationship feature learning. With some modifications to Eq. (4.2), the individual module can be defined as:

$$GIRN_{indiv}^p(G) = f_I([\mathcal{R}_p; \overline{\mathcal{R}^p}]) \quad (4.8)$$

where f_I is implemented as an MLP similar to f_G , but with its own set of parameters that are shared between individuals, and $\overline{\mathcal{R}^p}$ is the pooled relationships from the individuals other than p (i.e. all except the person p).

To couple the auxiliary individual modules into the training of our architecture, the loss function in Eq. (4.3) is redefined as:

$$\mathcal{L} = -\frac{1}{N_G} \sum_g \hat{\mathbf{y}}_{G,g} \log \mathbf{y}_{G,g} - \frac{2}{N_I \cdot N} \sum_i \hat{\mathbf{y}}_{I,i} \log \mathbf{y}_{I,i} \quad (4.9)$$

such that $\hat{\mathbf{y}}_{G,g}$ and $\hat{\mathbf{y}}_{I,i}$ are the ground-truth values for the group activity and individual action labels, and $\mathbf{y}_{G,g}$ and $\mathbf{y}_{I,i}$ are the predictions computed by the softmax layers, respectively. N_G and N_I are the numbers of classes for group activity and individual action respectively. The loss component coming from the individual actions is divided by the total number of individuals (N), so that the group activity recognition be prioritized during the learning process. However, we multiply this component by a factor of 2 to improve the performance.

4.2.4 Attention Mechanisms

Although participating on the same group activity, each individual can have different roles during the activity execution itself. Some of these roles can be more important, or simply more discriminating, for determining which activity is in fact happening (see Figure 4.3 (a)). Therefore, instead of naively average pooling the relations from all the players, an attention mechanism can be used so that more weight can be given to those potentially key individuals.

We opt to equip our network with a dot-product ($\cdot$) based attention [79], hence the attention score equation can be defined as follows:

$$a_{G,p} = W_{GQ} \cdot (\tanh(W_{GK}\mathcal{R}_p)) \quad (4.10)$$

where W_{GQ} and W_{GK} are trainable weights, respectively responsible for the query and key terms of the dot-product, and $a_{G,p}$ is the group attention score for person p . These scores will be normalized with the softmax function, then be used as weights for a weighted average of the individuals relations being pooled. The attention pooled relations ($\overline{\mathcal{R}}$) will then be used by the group activity inference module defined in Eq. (4.2).

Intuitively, an attention mechanism can also be coupled to the auxiliary individual modules. A person action should as well be influenced with different degrees by the other persons in the scene. For example it might be a reaction to another individual action, or even be a joint action. See Figure 4.3 (b) and (c) for visual illustrations. However, since the person importance is relative to a reference person, the individual attention mechanism should not generate a single score for each target person. It is necessary a more complex mechanism, such as the multi-head attention [80], which allows computing the targets attention scores considering multiple queries. Where in our case the query will change according to the reference individual. To

(a) Group Attention

(b) Individual Attention for Spiker

(c) Individual Attention for Blocker

FIGURE 4.3: Attention mechanisms examples. The blue, yellow and red colors of the bounding boxes indicates the contribution weight from lowest to highest, and the white color indicates the reference player. (a) During a spiking activity, the group attention mechanism gives more weight to the individual spiking, moderate weight to the blockers, and less weight to the remaining individuals. (b) For an individual spiking, the individual attention mechanism gives more weight to the blockers and less weight to the remaining individuals. (c) For an individual blocking, the individual attention mechanism gives more weight to the opponent spiking, moderate weight to the teammate also blocking, and less weight to the remaining individuals.

fulfill this requirement, the individual attention mechanism is defined as follows:

$$key_p = \tanh(W_{IK}\mathcal{R}_p) \quad (4.11)$$

$$query_p = W_{IQ}key_p \quad (4.12)$$

$$a_{I,p,q} = query_q \cdot key_p \quad (4.13)$$

such that $a_{I,p,q}$ is the individual attention score for person p with reference to individual q , and the trainable parameters are W_{IQ} and W_{IK} . Analogously to the group attention, the scores here are softmax normalized and used for weight-average pooling the relations present in Eq. (4.8), as a replacement for the average pooling strategy to generate $\overline{\mathcal{R}^p}$.

4.3 Experiments

4.3.1 Implementation Details

For our experiments, the skeleton information is extracted using the readily available tool OpenPose [26], with the *BODY_25* model and net resolution of 1312×736 . Examples of extracted poses are presented in Figure 4.4. We apply basic ad-hoc heuristics to attribute the extracted poses to each individual, and impute short-gap missing detections with interpolation. The skeletons are then normalized by removing the estimated camera motion between frames, and changing the origin to the coordinate central to all poses. For Intra-Person and Person-Object relationship types we use the following seven joints: Nose, Neck, Mid Hip, Left/Right Wrists and Left/Right Ankles. For Inter-Person we restrict to: Left/Right Wrists and Left/Right Ankles. We select this joints subsampling through preliminary experiments. The poses are randomly mirrored during training (by inverting the x-coordinates). This works as a good data-augmentation technique to help reduce the impact of over fitting. The group label is also inverted (left/right) when the skeleton is mirrored.

Although the poses are estimated for all 41 frames, not all of them are used when building the input for our method. We alternately sample the frames, since that leads to better performance for our models. In other words, the joints' arrays j_i are constructed by using the coordinates from 21 of the 41 frames available, by picking one and skipping the next.

FIGURE 4.4: Examples of estimated poses and annotated ball coordinates from Volleyball dataset. Contains examples from four distinct group activities: (a) Right Set; (b) Left Pass; (c) Right Spike; and (d) Right Winpoint.

Unfortunately, for the estimation of the ball coordinates there is no reliable tracker yet, to the best of our knowledge. Current research in ball tracking still faces difficulties generalizing to in-the-wild scenarios, where there is no information about the camera or any type of calibration [107]. To overcome this issue, and be able to assess the potential of our Person-Object relationship type, we use manually annotated ball coordinates¹. Samples of annotated ball coordinates are depicted in Figure 4.4.

As it was mentioned previously, our relational and inference modules are implemented as MLPs. The relational module (g) consists of four fully-connected layers, first three with 1000 units and the fourth with 500. For the inference (f) module, we use three fully-connected layers with 500, 250 and 250 units respectively. The attention mechanisms parameters (W_{GQ} , W_{GK} , W_{IQ} and W_{IK}) have the same dimensions as the input relations vectors (R_p). Our whole network is trained

¹Ball annotation: <https://drive.google.com/file/d/1urZpZiiepC85JD1u3VeURgUpztRgIOy1>

through back-propagation with Adam optimizer and learning rate value of $1e-4$. Weight initialization coming from a truncated normal distribution with 0 mean and 0.045 standard deviation. We use a dropout rate of 0.25 for the f module and the attention parameters, no dropout at g . Additionally to the individual loss weighting factor defined in Eq. (4.9), we also apply weights specific to class, since the number of instances for each of the individual actions is highly imbalanced (Standing represents almost 70%, Jumping only 0.6%). The relationship types are first trained separately, then their specific set of g parameters (Θ , ϕ and ω) is fine-tuned while the upper layers (attention mechanisms and f modules) are trained from scratch with the concatenated relations.

The individuals connection strategy used by the inter-person relationship type, responsible for defining $\mathbb{C}_p$ in Eq. (4.5), is tailored for the players structure in volleyball matches. Individuals are first split into two teams, then into two sub-groups: front and back. The *front* sub-group consists of the three players positioned closer to the volleyball net and the *back* sub-group are the remaining three, positioned furthest from the net. The players at the *back* are only connected to those on the same team, meanwhile the players in the *front* are also connected to the opponent players positioned in the other team's *front*. In case there are players missing, we leave unfilled spots in the *back* sub-groups. Connecting all the players would be counter-effective because the features from the important interactions could vanish due to the average operation with so many other interactions, possibly less relevant. Moreover, the number of pairs grows geometrically with the number of individuals connections and the number of joints used, becoming computationally prohibitive at some cases. In Subsection 4.3.4 we provide empirical results addressing these issues, that further support the connection strategy adopted.

4.3.2 Comparison with Previous Works

Here we compare previous techniques to our complete proposed method, using multiple relationships and including all components, the auxiliary individual modules

and the attention mechanisms. We chose to report for comparison the results from two of our multiple relations models, one using only the estimated poses and the other also including the annotated volleyball information, namely $GIRN_{intra+inter}$ and $GIRN_{intra+inter+object}$ respectively. The results for comparison are presented in Table 4.1.

TABLE 4.1: Comparison of our results with previous works on Volleyball dataset.

Method	Accuracy
HDTM [10]	81.9%
CERN [32]	83.3%
SSU [33]	90.6%
HRN [34]	89.5%
Multi-stream CNN [36]	90.5%
PC-TDM [108]	87.7%
Multimodal Attention [37]	91.7%
ARG [4]	92.6%
CRM [54]	93.0%
PMH [5]	87.7%
$GIRN_{intra+inter}$	88.4%
$GIRN_{intra+inter+object}$	92.2%

The $GIRN_{intra+inter}$ form of our proposed method is already capable of outperforming many of the previous works (i.e. HDTM [10], CERN [32], PC-TDM [108] and PMH [5]), even though it does not incorporate any CNN-based input as they do (e.g. RGB, optical flow, heatmap). Additionally, if we compare our method to reported ablation results not using RGB and OF, $GIRN_{intra+inter}$ also outperforms Multi-stream CNN [36] (82.6%) and is more closely comparable to CRM [54] (90.8%). We believe CRM [54] achieves a higher performance because it makes use of diverse visual information in a complex pipeline that incorporates a 3D CNN architecture and multiple steps of refinement.

With regard to $GIRN_{intra+inter+object}$, a direct comparison with previous work would be unfair because this variation makes use of manually annotated ball coordinates, data not automatically extracted from the videos and that does not contain errors. It is a good indication, however, on the potential of such type of information, as can be seen that the explicit incorporation of this input by the relational modules is able to boost our method performance to an accuracy close to the best results obtained by the compared methods. Moreover, this result is achieved while still keeping our input modality restricted to euclidean space representations, in other words, not incorporating visual spatial and motion information from RGB and optical flow. In addition, our results related to robustness evaluation in Tables 4.4 and 4.5 shows our method can leverage the ball location information even if it is noisy or incomplete, and can still outperform many of the previous work under these conditions.

Although not actively targeting at it, it is possible that the CNN image-based approaches are already learning patterns related to the ball location and movement during the activity. Therefore, this source of information should be somehow explicitly incorporated to pose-based approaches, for them to be able to suitably reason about group activity on sports. Besides, even the non-pose approaches could possibly improve their performance if the ball information was deliberately explored in their architecture. Naturally this should also hold true to different scenarios other than sports, whenever there are relevant objects that the individuals are directly or indirectly interacting with.

4.3.3 Ablation Study

Here we introduce the experiments conducted to evaluate the efficiency of the relationship types independently, and their positive complementary when fused. In these experiments, we also incrementally assess the impact of the different parts of our architecture. The results are shown in Table 4.2.

TABLE 4.2: Ablation experiments on Volleyball dataset. Accuracy performance of our proposed GIRN with different relationships types and cumulative addition of components.

Relationships	Baseline	+ Indiv. Modules	+ Attention
Intra-Person	86.7%	87.6%	88.0%
Inter-Person	83.5%	84.4%	84.7%
Intra+Inter	87.2%	88.0%	88.4%
Person-Object	85.5%	87.1%	87.6%
Intra+Inter+Object	90.1%	91.8%	92.2%

The baseline for our proposed method makes use only of the group activity labels during training, and the individual relations are average pooled with equal weights before being fed to the group activity inference module. Among the pose-only relationships, Intra-Person obtains a notable better performance than Inter-Person. However, their complementary consistently leads to an improvement when fused (Intra+Inter). The quality of these relations are meaningfully enhanced when the auxiliary individual modules are added to the baseline, allowing the group activity module to obtain an increase of almost 1 percentage point of accuracy. Subsequently, replacing the naive average pooling with our defined attention mechanisms further increase the performance, with 88.4% of accuracy being the highest result from our GIRN method when using only the pose information.

Aggregating the ball location information to our method allows the Person-Object relationship type to be also used. The inferred relations coming from this module show promising discriminating properties, obtaining a performance close to the Intra-Person relationship. Moreover, its high complementary to the previous two relationship types leads to a significant boost in performance, with 92.2% of group activity recognition accuracy.

4.3.4 Evaluation of Inter-Person Connectivity and Number of Joints

The performance for the inter-person relationship type is dependent on the predefined connectivity among the individuals and the number of joints selected. The results for the ablation experiments above follow the configuration detailed in Subsection 4.3.1, with a subsampling of four joints and a densely connection strategy (although not fully-connected). Here we explore the impact on the performance of different connection strategies and joints subsampling in conjunction.

For the joints subsampling, we range from two to seven joints, starting from the upper limbs (Wrists) and incrementally adding joints from the spine (Neck, MidHip and Nose) and lower limbs (Ankles). The connection strategies are based in the player position in the volleyball court. In a volleyball game, there are six players per team, arranged facing the net in two rows and three columns. For simplicity, we refer to the players positions based if they are in the front or back row, and if they are in the middle column or in the exterior columns (see Figure 4.5). The explored connection strategies are:

- **Full Connectivity:** Every player is connected to all other players.
- **Dense Connectivity:** All players are connected to their teammates, and players in the front row are also connected to the opponent team front row players.
- **Moderate Connectivity:** Similar to Dense, but players at the exterior columns have less connections (they are not connected to the players in the opposite exterior column).
- **Sparse Connectivity:** Similar to Moderate, but players at the middle column also have less connections (they are not connected to the players at their diagonals, i.e. the players in the exterior columns that are not on the same row as the player).

FIGURE 4.5: Subdivision of players positions in the volleyball court. The position of a player is based in the row (Front or Back) and column (Exterior or Middle) he/she is located.

The results are reported in Table 4.3. For some combinations of connectivity and joints subsampling there are so many input pairs that it becomes computationally infeasible to run. Thus there are absent values in the higher connectivity and joints cells (e.g. Full Connectivity with 5-joints, Dense Connectivity with 7-joints). Nonetheless, it can be seen that the performance does not always improve with the increase of connectivity or the number of joints. For example, using 5 or more joints seems to be detrimental to the performance with Dense connectivity. Moreover, the optimal number of joints depends on the level of connectivity. Four joints seems to be ideal for Dense, but not ideal for the other connectivity levels. In general there seems to be a reverse performance correlation between the degree of connectivity and the number of joints, i.e. the optimal number of joints increases as the degree of connectivity decreases. Overall, the dense connectivity degree obtains the highest accuracy results, with 4-joints being its optimal subsampling.

TABLE 4.3: Evaluation of Inter-Person relationship type with different players connectivity degrees and varying number of joints per subsampling.

Connectivity	2-joints	3-joints	4-joints	5-joints	6-joints	7-joints
Full	83.4%	83.8%	82.9%	-	-	-
Dense	84.1%	84.4%	84.7%	83.7%	83.8%	-
Moderate	83.1%	83.4%	82.9%	83.3%	83.3%	82.7%
Sparse	81.9%	82.3%	82.9%	82.1%	81.9%	83.5%

2-joints: Left/Right Wrists; 3-joints: 2-joints + Neck; 4-joints: 2-joints + Left/Right Ankles;
5-joints: 4-joints + Neck; 6-joints: 5-joints + MidHip; 7-joints: 6-joints + Nose.

4.3.5 Robustness Evaluation

To analyze the impact on performance of using inaccurate ball coordinates, as if they were coming from ball-tracking algorithms, we conduct experiments adding different types of noise to this information and evaluating our complete method over it: Intra+Inter+Object relationships with individual modules and attention mechanisms. This investigation is important to assess how robust our method is with regard to the ball information, i.e. how precise this information has to be such that GIRN leverages it when performing group activity recognition. The results from these experiments are presented at Table 4.4 and Table 4.5. Further details and discussion are provided subsequently.

TABLE 4.4: Evaluation of robustness against misplaced ball localization. Zero-mean Gaussian generated noise is added to the ball coordinates, with variable standard deviation values indicated by the displacement noise in terms of pixels.

Noise Std Dev	0px	20px	40px	60px	80px	100px	120px
Accuracy*	92.2%	92.0%	91.4%	90.8%	89.3%	89.1%	87.6%

* Accuracy averaged over five repetitions per parameter.

TABLE 4.5: Evaluation of robustness against ball miss-detections. Ball coordinates at each frame is randomly set to zero according to the dropout chances indicated.

Dropout Chance	0%	5%	10%	15%	20%	25%
Accuracy*	92.2%	91.2%	90.3%	89.6%	88.8%	87.7%

* Accuracy averaged over five repetitions per parameter.

We start by studying the impact of misplaced localization, through addition of Gaussian noise to the coordinates. The generated noise has zero mean and variable standard deviation parameters, where the standard deviation values represent the displacement in terms of pixels from the ball actual coordinates. The majority of the videos have a resolution of 1280×720 , but a few have 1920×1080 . To keep the added displacement proportional, we multiply the indicated pixel value by 1.5 times when generating the noise for the latter case. As it can be seen at Table 4.4, our method is considerably robust to inaccurate ball coordinates, it can still obtain more than 90% of accuracy even with a displacement noise of standard deviation set to 60 pixels (approximately $1/20th$ or $1/12th$ of the frame width and height respectively)s. The use of the ball information stops being advantageous (same performance as Intra+Inter) only when the standard deviation value reaches 120 pixels, which is a very high displacement noise (approximately $1/10th \times 1/6th$ of the frame resolution).

Our next experiment, reported at Table 4.5, targets at measuring the impact of ball missed detection at some frames, therefore having an input with coordinates equal to zero at different timesteps. To accomplish this, we randomly set the ball coordinates to zero according to distinct dropout chances and feed it to our method for testing. Our method is more sensitive to this type of noise, however it remains capable of achieving more than 90% accuracy regardless of 10% of the coordinates being set to zero. In fact, a quarter of the sequence has to be missing for our method to stop obtaining a performance better than using only the skeletons. Moreover, taking into consideration that the previous experiment validates the robustness of our method to inaccurate coordinates, the miss-detections can be imputed with interpolated values from the acquired detections, reducing the negative impact to performance.

In conclusion, the results from these experiments indicate that our method is sufficiently robust to noise in the ball coordinates, not requiring a precise ball detection in all frames. Supported by the results here, we expect that our method can obtain

a satisfactory performance even with incomplete ball trajectories, by using a simple interpolation as a work around to fill in the missing gaps.

4.3.6 Qualitative Analysis

For a more in depth analysis of the performance of our method, and also to visualize where are the differences in the results from our variations with and without using the volleyball, we produce the confusion matrices present in Fig. 4.6.

FIGURE 4.6: Confusion matrices for our methods variations using pose-only ($GIRN_{intra+inter}$) and pose plus volleyball ($GIRN_{intra+inter+object}$).

Focusing first in the confusion matrix for $GIRN_{intra+inter}$, it can be seen a consistent confusion between Right and Left sides for the same type of activities (anti-diagonal values). The degree of confusion changes slightly for some of the activities (higher with Pass) and the sides (higher for Right). Moving on to $GIRN_{intra+inter+object}$ matrix, it is evident that the addition of the ball information greatly helps at distinguishing between sides. There are much less confusion in the anti-diagonal cells and in its containing quadrants, only a few exceptions remains and with low-values mostly. However for the Winpoint activity, the confusion is still high between sides. Such result is not surprising, since for this activity the side

label is not straightforwardly related to the location of the ball, in many cases the ball is not even present in the scene. For this activity there are instances which are challenging even for humans to distinguish, requiring a longer term inference based on which team scored the point, or a finer analysis per individual to identify if their response is cheerful or upset. See Figure 4.7 for an example of the aforementioned failure case, together with examples from other cases.

a) Left Winpoint wrongly classified as Right Winpoint

b) Right Pass wrongly classified as Right Set

c) Left Set wrongly classified as Left Pass

FIGURE 4.7: Examples of video clips from the Volleyball dataset wrongly classified by $GIRN_{intra+inter+object}$.

With regard to the activities directly involving the volleyball, Spike seems to be more distinguishable, obtaining the highest accuracy results in average. Pass and

Set in the other hand have a greater confusion between them, and also some miss-labeling for Set as Spike. A possible cause for these confusions are that the pose estimation might have some errors – due to model precision or occlusion – in the hands and arms joints, which are more important for identifying these activities. Such issues might be automatically sorted out when more precise pose estimation models are available. Another reason might be temporally related. Since these activities often happen in a sequence (i.e. Pass $\Rightarrow$ Set $\Rightarrow$ Spike), and the temporal length of our input comprises the whole snippet duration, it is possible our method is labeling the scene with the previous or next activity instead of the central one. This could be solved by adding a mechanism to learn the temporal dynamics and causality between actions, or by giving more weight to the central activity.

4.4 Summary

In this chapter, we tackled the problem of group activity recognition by a different perspective than previous works: using the skeleton information for reasoning over the interactions between multiple individuals and also between individuals and objects. We demonstrated, through our GIRN, that such pose information can be leveraged for solving this problem by using multiple pair-wise relational modules that will specialize in different relationships types based on the nature of the pairs being fed to them. With the support of individual actions knowledge to refine the relations descriptions, and attention mechanisms to give more importance to key persons, our method obtain promising results. Although being exclusively based in a single modality for input, our method is able to achieve competitive performance with respect to the state-of-the-art, which commonly uses at least RGB and Optical Flow modalities. Our experiments also demonstrate how explicitly accommodating domain specific objects of interest (e.g. volleyball) in the architecture can lead to significant improvements to performance. To the best of our knowledge, this is the first work to actively explore such type of information for the task of group activity recognition.

Chapter 5

Detection of Real-world Fights in Surveillance Videos

In this chapter we further increase the complexity of the human interactions being addressed, we focus on detection of fights in untrimmed videos. Firstly we introduce a novel dataset, with challenging footage of real-world fights captured by surveillance cameras (CCTVs). Then we propose a general framework for fight detection, and evaluate how some of the existing methods perform in this scenario.

5.1 Introduction

A common technological device for increasing security is Closed-Circuit TeleVision (CCTV), which consists of a system for video surveillance, usually covering different locations, such as public places, schools, shopping malls, residential or commercial areas. In spite of the effort, CCTV effectiveness is questionable [109], specially because they require sufficient trained supervisors, and human attention capability in viewing videos for a long period of time by itself is limited.

Among the different events that might be relevant for detection and prevention using CCTV, fight is a common event of interest, and that might arise from different

FIGURE 5.1: Frames sampled from CCTV-Fights, showcasing the challenging diversity of its fight scenes and conditions.

situations (e.g., heated discussions, burglary, hate crimes). Consequently, it is a paramount event to have its detection automated. Nonetheless, this problem has been mainly neglected by previous literature. Most of the previous works, although using the surveillance scenario as motivation, focus their methods on detection of general activities [16, 110–112]. Moreover, the few works that focus on fight detection do not base their techniques on realistic fights and/or surveillance footage [38, 40–42], with the great majority targeting short-clips, instead of long, untrimmed, videos.

A possible reason for this research gap might be related to the fact that there is no available dataset, which comprises all these characteristics: real-world fights from surveillance cameras. For example, existing datasets comprise video data from general activity [113], Hollywood movies [44] and faked actions [43]. For that

reason, it is paramount to propose a new dataset with these features, that could enable us to design and evaluate an approach better suited for this scenario.

In this regard, our contributions in this paper are twofold: First we introduce the new dataset CCTV-Fights, consisting only of real-world fights, and containing more than 8 hours of CCTV footage temporally annotated. Then we propose an initial methodology to tackle the fight detection problem and that serves as a baseline for future research in the field. We evaluate different feature extraction methods ranging from Deep Learning to Local Interest Points, and also different classifiers – including end-to-end CNN, LSTM and SVM.

5.2 CCTV-Fights Dataset

Existing datasets for fight detection have a diverse range of characteristics: different nature, duration of the videos, number of videos, purpose, recording source, staged or not, and so on. Table 5.1 contains a summary of existing datasets in prior art, and sample frames are provided in Figure 5.2. Yet, none of them can properly cover the scenario of realistic fights recording coming from CCTV surveillance cameras, which is mandatory to evaluate a method on the paramount problem of fight detection under these circumstances. To overcome this issue, a new dataset was collected, containing 1,000 videos picturing real-world fights, recorded from CCTVs or mobile cameras: **CCTV-Fights**.

The dataset videos were collected from YouTube, searching with keywords like: CCTV Fight, Mugging, Violence, Surveillance, Physical violence, etc. The fights can contain a diverse range of actions and attributes, for example: punching, kicking, pushing, wrestling, with two persons or more, etc. It was discarded videos that did not come directly from a CCTV recording (e.g., footage made with a mobile camera recording a screen), as well as videos with heavy special effects (e.g., shaded borders, slow-motion). Figure 5.1 depicts some examples from the collected videos.

Hockey Fight

[38]©2011 IEEE

Movies

[38]©2011 IEEE

Violent-Flows

[39]©2012 IEEE

VSD

[44]©2015 Springer

RE-DID

[67]©2015 IEEE

BEHAVE

[43]©2010 BMVA

CCTV-Fights (Proposed dataset)

FIGURE 5.2: Sample frames from existing violence and fight datasets. Very distinct types of fights, definitions of violence, and footage characteristics.

TABLE 5.1: Summary of some of the most relevant datasets commonly used for Fight Detection.

	Name	Size	Characteristics
Trimmed	Hockey Fight [38]	1,000 clips	Hockey players
	Movies [38]	200 clips	Trimmed action movies
	Violent-Flows [39]	246 clips	Crowd violence
Untrimmed	VSD [44]	25 movies	Complete Hollywood movies
	RE-DiD [67]	30 videos	Urban fights + Cars/Mobiles
	BEHAVE [43]	4 videos	Acted fights + CCTVs
	CCTV-Fights	1,000 videos	Urban fights + CCTVs/Mobiles

Approx. duration: Clips (2-5 secs) - Videos (20 secs - 5 mins) - Movies (1.5 - 2 hours)

TABLE 5.2: Summary of CCTV-Fights statistics. Value in parentheses indicates average number of fight instances per video.

	Videos	Duration (hours)	Fight Instances
All	1,000	17.68	2,414 (2.41)
CCTV	280	8.54	747 (2.67)
Non-CCTV	720	9.13	1,667 (2.32)

This way we managed to acquire 280 CCTV videos containing different types of fights, ranging from 5 seconds to 12 minutes, with an average length of 2 minutes. By itself, this set is a bigger corpus than any of the existing datasets in prior art. Furthermore, we collected additional 720 videos of real fights from other sources (hereinafter referred to as Non-CCTV), mainly from mobile cameras, but a few from car cameras (dash-cams) and drones or helicopters. These videos are shorter, 3 seconds to 7 minutes, with an average length of 45 seconds, but still some have multiple instances of fight and can help the model to generalize better. Table 5.2 presents a summary of the dataset statistics.

Videos were annotated at frame-level, i.e., each fight instance segment in the video was labeled with its exact start and end points. Although the occurrence or not of a fight considering a complete scene or a few seconds split is usually a consensus among most human viewers, the specific begin and end points of the fight is more

subjective, being prone to discussion according to different points of view. The following good practices were adopted to overcome this issue:

- Annotation should contain a few seconds extra at the edges of the event;
- Short moments during the fight without strikes or hits, but with the perpetrators still in fighting instance, should still be labeled as positive;
- Long breaks should be labeled as negative, with following fights being considered different instances.

On top of the aforementioned good practices, it is expected that the evaluation metrics should not be so strict over the precise start and end of the fight predictions, introducing flexibility at some degree for the temporal localization. For the experiments, we randomly selected 50% of the videos for training, 25% for validation and another 25% for testing.

5.3 Methodology

We propose the following pipeline, depicted in Figure 5.3, as a benchmark for this dataset, to represent our first take at solving the CCTV fight detection problem. The pipeline can be split into three specific steps, which will be further discussed bellow, as well as the specific methods utilized at each one of these steps. Regarding the chosen methods, we rely upon Two-Stream [46], 3D CNN [114] and local interest-points [42].

5.3.1 Feature Extraction

The first step of the pipeline, **Feature Extraction**, consists on using the RGB information from the frames to extract meaningful features for the task at hand.

FIGURE 5.3: Proposed Pipeline. At the first step, it is extracted the features from the video frames. Second step consists on predicting the frame/snippet based on these features, resulting in confidence scores for each frame/snippet. In the last step, fight segments predictions are generated by merging continuous frames with sufficiently high confidence scores. The output is represented by a timeline bar, on which the red parts indicate the fight segments predictions (starting and ending time in seconds).

These features are meaningful if they are discriminative enough for a decision-making method to correctly classify that feature as coming from a fight or not. Depending on how these features are generated, it can be used to describe a single frame or a small snippet of the video (e.g., a few sequential frames).

The feature extraction for the two-stream [46] based approach (see Figure 5.4), is performed by using a 2D-CNN architecture for generating two different models, one for the spatial stream of the videos (RGB data of Frames) and another for the temporal stream (Stack of Optical Flows). We aggregate this information in the end by average pooling the scores or by concatenating the features from the last fully-connected layer before feeding it to a classifier.

The 3D-CNN method [114] consists of a convolutional neural network architecture that enables convolution on three dimensions (see Figure 5.5). This way, the convolution can be explored not only for the spatial correlation within a single frame, but also for the temporal correlation between a short sequence of frames. The 3D-CNN is applied over a stack of sequential frames only, no optical flow information is used.

For the local interest-points approach, we chose to base on Moreira et al. [61], which is a recent paper related to Fight Localization (although only evaluated in general violence on movies). In their work, it is used their proposed local features detector

FIGURE 5.4: Illustration of the two-stream approach [46] ©2014 NIPS. Both streams, Spatial and Temporal, use the same 2D-CNN (ConvNet) architecture (detailed in the box sequence). However, each stream ConvNet is trained with different types of data: Spatial stream is trained with the video frame directly; and Temporal stream with a stack of optical flow representations. The scores from both streams are fused to obtain the final class score for the video frame.

and descriptor, namely Temporal Robust Features (TRoF). TRoF is an extension to the temporal domain of the traditional local detector SURF. To achieve that, the authors propose 3D detection boxes and description blocks (see Figure 5.6).

5.3.2 Frame/Snippet Prediction

The next step, **Frame/Snippet Prediction**, denotes the moment in which the classifier of choice will determine whether if the feature comes from a positive case (fight) or a negative case, according to what it has learned before from the training and split of the data. Predictions produced in this step are at the frame or snippet level, represented by a confidence score.

At this stage, each of the chosen features above were paired up with a different technique for prediction, based on their reference works. The two-stream method was applied end-to-end, with the frame/snippet prediction coming from the CNN classification layer (see Figure 5.4). For the 3D CNN, the features from the last fully-convolutional layer before the classification is extracted, then fed to an LSTM for prediction (see Figure 5.5). Finally, the snippet classification of TRoF features (after Fisher Vectors) is done by a linear SVM (see Figure 5.7).

FIGURE 5.5: Illustration of a joint 3D-CNN and RNN architecture [114] ©2016 arXiv. The video clips being fed to the C3D are stacks of sequential frames, which can overlap or not. The C3D features from each video clip is then fed to an RNN for temporal inference and for activity classification of that snippet of the video. LSTM is the type of RNN used in [114] and also in our pipeline.

FIGURE 5.6: Illustration of TRoF detector and descriptor main elements [61] ©2017 IEEE. (a) depicts the values of the 3D box filters used to detect the interest-points in the video. (b) represents the three description planes $[x, y]$, $[x, t]$ and $[y, t]$ of the block around the interest-point in position $P(x, y, t)$.

5.3.3 Segment Generation

The last step, **Segment Generation**, is responsible for aggregating predictions from the previous step to produce well-defined temporal segments for the predicted fight instances. Here, some higher-level intuition on the continuity of an event can

FIGURE 5.7: Illustration BoVW-based framework used with TRoF [61] ©2017 IEEE. Low level: TRoF local feature detection and description. Mid level: Fisher Vectors coding and pooling. High level: SVM frame label prediction.

be used to smooth the punctual predictions from the frames/snippets and achieve more realistic segments than directly using the scores independently.

Similarly to previous work [42, 114, 115], no specialized method was employed to generate the final segments. A straightforward strategy of smoothing then aggregating was used. The smoothing is a traditional mean filter applied to reduce the impact of noisy prediction scores by using the score information from the neighbors snippets. After smoothing the scores, continuous predictions that satisfy a pre-determined score threshold are merged to create the final segments.

There are recent and advanced methods for action localization [101–103] that could be employed for segment generation and potentially perform well. However, we chose not to employ a specialized method in order to keep consistent with previous works in fight localization [41, 42, 115]. Moreover, this way we could focus on evaluating the different methods for the feature extraction and frame prediction steps of our pipeline.

5.3.4 Integrated Pipeline Implementation Details

The overall pipeline, with the selected methods aforementioned, is depicted in Figure 5.8. For the Two-Stream approach, it is used the 2D-CNN architecture VGG16 [116] for frame and flow feature extraction and prediction, fine-tuning

FIGURE 5.8: Detailed proposed pipeline, showing the different techniques integrated by our approach to detect the fights in the video.

with pre-trained weights from ImageNet and UCF101, for the spatial and temporal streams respectively. As suggested in Simonyan and Zisserman [46], ten consecutive flows are stacked and used as input to the temporal stream. The optical flow displacement fields are computed using Brox et al. [117] proposed method. With regard to the 3D CNN method, the C3D architecture is utilized as a feature extractor, applying directly the weights learned from Sports-1M dataset [118], and the output from the layer “fc7” as the extracted feature vector. With respect to TRoF, it is used to extract low-level spatio-temporal features from the videos, then the features undergo PCA transformation for whitening and reducing their dimensionality by half. Subsequently, Fisher Vectors (FV) [119] is applied for mid-level

generation of features for each snippet of the video.

The features generated through the C3D and TRoF methods are used to describe short snippets of the video, exactly 16 frames, with a stride of 8 frames between the center frame of the snippets – which leads to overlapping snippets. For the TRoF framework, that means that the low-level features pooled during the Fisher Vectors stage will come from the 16 frames comprised in the current snippet. For the C3D architecture, the 16 frames-sized snippet translates into the temporal input size of the network.

For the snippet prediction stage, the LSTM architecture and training hyperparameters are picked by grid-searching and using a validation split for measuring the performance. The SVM hyperparameter is chosen through grid search and cross-validation during the training phase.

To make a more fair comparison between different methods for feature extraction and snippet prediction, all the permutation possibilities are evaluated.

5.4 Experiments

5.4.1 Evaluation Metrics

To evaluate the performance from the chosen methods, we use the metrics **mean Average Precision (mAP)**, for assessing the detection of the fight segments, and the **F-Measure**, for assessing fight detection at the frame level. Qualitatively speaking, the mAP is a coarse-grained metric, since it appraises how many fight segments are correctly detected per video, penalizing the measurement in case of wrong prediction segments. Meanwhile, the F-Measure is a more fine-grained manner to assess the predictions, since it is based at the frame level, i.e. at the actual time length which is correct or wrong within a segment prediction.

Firstly, the mAP metric requires the computation of the video average precision (AveP), defined below as:

$$\text{AveP}(v) = \frac{\sum_{k=1}^n (P(k) \times \text{rel}(k))}{\text{number of fight segments}} \quad (5.1)$$

where n is the number of prediction segments, k is the rank in the sequence of predictions ranked in descending order by their confidence score, $P(k)$ is the precision considering the first k predictions, and $\text{rel}(k)$ is 1 if the k -th prediction is a hit or 0 otherwise (a miss). This way, the mAP value, for a certain number V of test videos, can be computed as follows:

$$\text{mAP} = \frac{\sum_{v=1}^V \text{AveP}(v)}{V} \quad (5.2)$$

For the mAP metric, it is necessary to determine if an entire fight prediction segment is a hit or a miss (i.e. is correct or wrong). However, as we are dealing with localization, determining whether a prediction segment is a hit or miss is not straightforward. If we considered only perfect matches, that would be too strict and could not assess properly the prediction quality. Also, it has to be taken into account that, the exact moment when a fight begins or ends can be subjective, so the evaluation metric should not be so rigid. Therefore, similarly to MediaEval protocol [120], to deem a prediction as a hit (i.e., to contain a fight), it must have at least 50% of its length overlapping with a ground-truth segment. Also, if many (small) segments satisfy this requirement for a same ground-truth segment, only one will be considered as a true positive, and the rest will be disregarded (neither true nor false positive). Examples of fight prediction segments, hits and misses, are presented in Figure 5.9.

The aforementioned criteria to define a true positive, based on MediaEval protocol [120], penalizes predictions that are too long or misaligned with respect to the ground-truth. However, it favors short predictions that are within a ground-truth segment, even if it covers just small portion of the ground-truth total length. In order to provide a stricter metric to this regard, we also report the mAP with the

FIGURE 5.9: Example of a video ground truth segment and possible prediction segments. For a prediction segment to be considered a hit (i.e. correct), at least 50% of its length has to overlap with the ground truth segment.

intersection-over-union (IoU) criteria. This criteria is commonly used in action localization benchmarks [113, 121]. By this criteria, for a prediction to be considered a true positive, the intersection-over-union ratio between the prediction segment and the ground-truth segment has to be above a certain threshold. We choose the threshold value to be 0.2, as it penalizes very short detections (e.g. a single frame), while remaining flexible enough so that partially correct detections are not dismissed.

The F-Measure (also known as F-score or F_1), is a useful metric for binary classification, which balances out precision and recall while computing a single measurement. This metric can be summarily defined as follows:

$$\text{F-Measure} = \frac{tp}{tp + \frac{1}{2}(fp + fn)} \quad (5.3)$$

where tp , fp and fn is respectively the number of true positives, false positives and false negatives. For computing the F-Measure, it is considered the prediction at the frame level, in contrast to evaluating the entire segment.

5.4.2 Experimental Results

Table 5.3 contains the results for the previously described methods, in the CCTV-Fights dataset testing split. The Two-Stream approach is significantly better than the others, regardless of the classifier used, and for both metrics. Using directly the CNN scores output leads to the highest mAP, but the slightest highest F-Measure comes from using the LSTM.

TABLE 5.3: The results for the chosen methods on the CCTV-Fights dataset. First column indicates the feature used and second which classifier was used for training and prediction.

Features	Classifier	mAP	F-Measure	mAP (IoU)
Two-Stream	CNN	79.5%	75.0%	47.5%
	SVM	76.6%	72.8%	50.4%
	LSTM	76.0%	75.9%	51.9%
C3D	SVM	64.5%	58.6%	28.3%
	LSTM	61.0%	58.1%	30.3%
TRoF	SVM	69.2%	63.3%	36.3%
	LSTM	63.8%	63.5%	32.4%

The explanation on why the two-stream approach has higher performance than the others should be related to the use of the explicit motion information. As can be seen in Table 5.4, when we look at the streams individual performance, the Spatial stream has a much lower performance than the Temporal stream, being equivalent to TRoF. In fact, its fusion with the Temporal information even lowered the performance of the latter by itself.

To look closer at the results, we report Table 5.5 with the Temporal stream performance split by type of data source: All, Non-CCTV and CCTV. We also target at specializing the CNN model by training it with the CCTV data in two different manners: 1-tiered, by training only with these videos; 2-tiered, by first training with all data, then fine tuning using only the CCTV source. As expected, surveillance footage is much more challenging, having a significantly lower mAP and

TABLE 5.4: Results for the Two-Stream approach, separately and combined, using directly the CNN output as classifier.

Features	mAP	F-Measure	mAP (IoU)
Spatial	68.6%	61.0%	38.2%
Temporal	80.8%	75.3%	49.0%
Two-Stream	79.5%	75.0%	47.5%

F-measure performance than Non-CCTV. Results also show that using data from multiple sources helps generalize the model better, and that a specialization in a two-stage training leads to a better performance than training solely on the CCTV domain.

TABLE 5.5: Performance of the Temporal stream, with CNN as classifier, split by training model and type of data source.

Model	Source	mAP	F-Measure	mAP (IoU)
All	All	80.8%	75.3%	49.0%
	Non-CCTV	85.9%	79.6%	50.7%
	CCTV	73.7%	66.7%	48.8%
1-tiered	CCTV	72.1%	63.5%	44.3%
2-tiered	CCTV	75.6%	67.7%	48.8%

5.5 Qualitative Analysis

To further analyze the performance of the 2-tiered Temporal stream, we sample some examples of generated segment predictions (Figure 5.10), and also show sampled frames from cases of correct prediction and false alarm (Figure 5.11).

We show in Figure 5.10 the aligned Ground truth fight segments and Prediction fight segments generated by the 2-tiered Temporal stream method, for multiple videos with distinct duration and number of fight segments. In a) fight_0017,

the first generated prediction matches closely to the ground truth and the short break between the fights is preserved, however, the second prediction segment extends for a longer duration than the annotated fight segment. The second example (fight_0066) contains a longer duration video (almost 8 minutes), nonetheless the generated predictions can detect the fight segments, although there are some short false alarms with low confidence score (represented by the color intensity within the segment). For video c) fight_0069, although the last prediction segment has the highest confidence and matches a ground truth segment, there are many false alarms throughout the video, including a wrong prediction with high confidence at the beginning of the video. Video d) fight_0206 consists entirely of a single fight segment, which was correctly detected by the 2-tiered Temporal stream method. On the other hand, in video e) fight_0226, no prediction segment was generated, thus missing the short and only fight segment in the video.

For exemplification purposes, we present in Figure 5.11 frames sampled from two different sequences, one on which the 2-tiered Two stream method has correctly detected the fight segment, and another where the generated fight segment is a false alarm (i.e. there is no fight in the sequence). In the first case, the method is able to correctly detect the fight even though the fighting is happening far from the camera (top right corner in the frames), and there is even a lamp post occluding partially the scene. In the false alarm case, the individuals seem closer to the camera and are mostly located at the center of the frame, however the method wrongly classifies this sequence as a fight, despite the favorable recording conditions. The wrong detection possibly happens because of the high motion in the scene, since there is an individual running towards another person.

5.6 Summary

In this chapter, a new and challenging dataset was created for the scenario of real-world fights on surveillance cameras, namely CCTV-Fights. Our dataset is valuable not only because of the high volume of data obtained and the important

FIGURE 5.10: Examples of fight segment predictions generated by 2-tiered Temporal stream. The dark parts in the Ground truth and Prediction bars represents, respectively, the annotated and the generated prediction fight segments in the video. For the Prediction bar, the color intensity of the segment indicates the method confidence in the prediction, i.e. darker segments means higher confidence score. The numbers below the Prediction bar shows the video timestamp (in seconds) at that location (omitted for fight_0066 due to visual overlapping).

FIGURE 5.11: Sample frames from segment cases of correct prediction (left column) and false alarm (right column), generated by 2-tier Temporal stream. At the example sequence in the left column, the fight can be correctly detected although being far from the camera and with partial occlusion at some moments. In the right column sequence, the method wrongly predicts the segment as a fight (false alarm), even though the individuals are located at the center of the footage.

temporal-level annotations, but also because there was no standard dataset that completely covered this scenario in previous work. This way, our dataset stands as a challenging benchmark to assess the robustness of fight detection methods in the aforementioned scenario. Moreover, it encourages the development of more advanced and reliable methods in this area. Subsequently we proposed a pipeline for fight detection and localization. Our results showed that the use of explicit motion information (e.g., Optical Flows) has a major positive impact on performance, being significantly superior than the RGB-only methods. Moreover, it is possible to leverage the information coming from Non-CCTV fights, through a 2-tiered model that better generalizes for the CCTV source.

Chapter 6

Conclusion and Future Work

In this chapter, we provide the conclusion to our works in this thesis, highlighting our achievements and contributions with each one of them. Subsequently, it is presented possible future research works, based on the findings of our works introduced in the previous chapters.

6.1 Conclusion

In this thesis, we have proposed novel methods and a new dataset for reasoning over multiple human-human interaction activities. We tackled distinct scenarios of human activities, each with its own set of characteristics, constraints and requirements. Moreover, we explored different types of input data, ranging from RGB and Optical Flow, to human skeleton information.

In our first work (Chapter 3) presented in this thesis, we introduced a novel method for two-person human interaction recognition, named IRN. It consists of an architecture that focuses at relational reasoning over the different relationships between the human joints during interactions. Our proposed method IRN, is capable to learn from the data itself how the human body parts relate to each other during

the interaction, thus standing as a highly valuable alternative to the usual techniques based on CNNs/LSTMs/GCNs. Important extensions are made over the original Relational Network proposition in order to tailor it for our problem and data, human interaction using skeleton information. Extensions such as: model specialization at multiple relationship types; fusion of the multiple types of relationships into a single model; pair-wise structured operation over the input to extract distance and motion information; and addition of LSTM for sequential relational reasoning. The IRN obtains state-of-the-art performance on two traditional interaction datasets, SBU and UT, and on the NTU RGB+D dataset mutual actions subset. Our method also achieves competitive results in the interactions subset of NTU RGB+D 120. These experiments validate the effectiveness of our proposed method.

In our second work (Chapter 4), we focused on the task of group activity recognition. We proposed a novel method (GIRN), which uses the skeleton information for reasoning over the interactions between multiple individuals and also between individuals and objects of interest. Our GIRN leverages pose information for group activity recognition, through multiple pair-wise relational modules, that specializes in distinct relationships types based on the nature of the pairs being fed to them. Our proposed method is enhanced with improved relations learning, by also incorporating individual actions information to our network. Moreover, group and individual attention mechanisms are added, which enables our method to give more weight to the relations with higher importance, further improving the recognition performance. Extensive experiments, in a challenging dataset with sport group activities, demonstrate the effectiveness of our method. The proposed method obtains a performance competitive with the state-of-the-art, in spite of our method relying solely in the euclidean space modality as input, in contrast to RGB and Optical Flow modalities. The experimental results also show that explicitly exploring human-object interaction leads to a significant improvement in performance.

Finally, in our third work (Chapter 5), we moved on to the even more challenging

scenario of fight detection on surveillance footage, which comprises complex human-human interactions captured under adverse recording conditions. We started by introducing our new and challenging CCTV-Fights dataset, which contains many long-duration videos with the aforementioned characteristics: real-world fights captured by CCTVs. This novel dataset is comprised by 1,000 videos, with a total duration over 17 hours, where each video is temporally annotated indicating when a fight segment begins and ends. In this work, we also proposed a pipeline for fight detection and localization, providing a benchmark approach and baseline assessment for the newly proposed dataset. Experiments following our pipeline, and exploring multiple existing methods, indicates the most promising techniques and which type of information is more relevant for this task. Moreover, a 2-tiered training strategy was also proposed, for leveraging the non-surveillance data before specializing the model at fight detection on surveillance footage.

6.2 Future Work

In this section, we briefly discuss some of the potential future research directions related to the works presented in this thesis and to the field of human-human interactions analysis.

6.2.1 Improved Pair-wise Relational Reasoning

Our works, presented in Chapter 3 and Chapter 4, demonstrated the effectiveness of handling human-human interactions through a pair-wise relational reasoning approach. However, improvements can be made to further enhance the reasoning capability of our methods.

Currently, the network module responsible for the pair-wise input (g function) is simple yet effective, since it is implemented as a traditional MLP. However, a more sophisticated architecture can lead to higher performance and potentially

demand less computational resources. Such as an architecture that takes into account the structure of the input data (in our case the body joint coordinates throughout sequential frames). For example, a pair-wise network with an CNN-based architecture can be designed, such that the input pair coordinates do not need to be flattened and concatenated at input, keeping their temporal (frames sequence) and spatial (x and y coordinates) dimensions separated.

Another possible improvement, would be to design a multi-modality architecture, that can fuse information from distinct types of data before inferring the pair-wise relations. Such architecture could then use data coming from multiple sources as input, such as estimated poses, RGB video and computed Optical Flow. Therefore it would have access to rich and complementary information when learning the input pair relationships and reasoning about the human-human interaction.

6.2.2 Multiple Groups with Independent Interactions

Current works in the field of group activity analysis have the implicit assumption that there is a single, or a main, group activity happening in the video clip. However, this is not always the case in the real-world, with unconstrained scenarios. It is possible that there are multiple groups of individuals in the same scene, where the persons of a group are interacting with each other, but there is no interaction between groups, thus the interactions are independent. For example, in a surveillance camera recording of a wide area such as a park, there might be distinct groups of individuals engaging at all sorts of independent activities, and with variable number of members (e.g. talking, having a picnic, playing Frisbee, walking a dog).

Research in this scenario would have to deal with new and challenging issues. For example, how to cluster the individuals and determine which activity they are performing. Ultimately, such type of scenario is important to push research forward into understanding and reasoning of human-human interactions with high complexity.

6.2.3 Pose Estimation in Surveillance Videos

Reasoning over human-human interactions captured by surveillance cameras is currently quite challenging, thus it could benefit largely by incorporating skeleton data as an extra or alternative source of information. The main issue is that obtaining reliable skeleton data from usual surveillance footage is currently not possible. Existing pose estimation methods do not work well on most videos coming from surveillance systems, because these systems commonly make use of 2D capturing devices with low-resolution, under various illumination conditions and from adverse viewpoints. In order to leverage skeleton data in the surveillance scenario it would be first necessary to design new pose estimation methods, robust to the aforementioned conditions encountered in surveillance footage. These methods might need to incorporate some kind of depth perspective awareness, to accurately estimate poses from individuals at different distances and angles from the camera.

With skeleton data from reliable estimated poses, more advanced methods for detection and recognition of human interactions under surveillance videos can be developed. Opening a new and promising frontier for research in the field of surveillance video analysis.

List of Author's Awards and Publications

Awards

- **Second Best Poster**, “Graduates’ Research Showcase,” *NTU EEE*.

Journal Articles

- **Mauricio Perez**, Jun Liu and Alex C. Kot, “Skeleton-based Relational Reasoning for Group Activity Analysis,” Under review in *Elsevier Pattern Recognition (PR)*, 2021.
- **Mauricio Perez**, Jun Liu and Alex C. Kot, “Interaction Relational Network for Mutual Action Recognition,” in *IEEE Transactions on Multimedia (TMM)*, 2021.
- Jun Liu, Amir Shahroudy, **Mauricio Perez**, Gang Wang, Ling-Yu Duan and Alex C. Kot, “NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

Conference Proceedings

- **Mauricio Perez**, Jun Liu and Alex C. Kot, “Interaction Recognition Through Body Parts Relation Reasoning,” in *IAPR Asian Conference on Pattern Recognition (ACPR)*, 2019.
- **Mauricio Perez**, Alex C. Kot and Anderson Rocha, “Detection of Real-world Fights in Surveillance Videos,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

Bibliography

- [1] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015. xix, 2, 14, 15, 40, 46, 55
- [2] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. In *AAAI Conference on Artificial Intelligence*, 2018. xix, 14, 15, 51, 57, 58
- [3] Yanli Ji, Guo Ye, and Hong Cheng. Interactive body part contrast mining for human interaction recognition. In *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2014. xix, 16, 17, 40, 55, 64
- [4] Jianchao Wu, Limin Wang, Li Wang, Jie Guo, and Gangshan Wu. Learning Actor Relation Graphs for Group Activity Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9956–9966, 2019. xix, 3, 18, 19, 64, 77
- [5] Hsing-yu Chen and Shang-hong Lai B. Group Activity Recognition via Computing Human Pose Motion History and Collective Map from Video. In *IAPR Asian Conference on Pattern Recognition (ACPR)*, volume 12047 of *Lecture Notes in Computer Science*, pages 705–718, 2020. xix, 3, 19, 64, 77
- [6] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, and Dimitris Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 28–35, 2012. xix, 2, 5, 17, 26, 27, 39, 40, 43, 49, 55, 63, 64
- [7] Michael S Ryoo and Jake K Aggarwal. Spatio-Temporal Relationship Match : Video Structure Comparison for Recognition of Complex Human Activities.

- In *IEEE International Conference on Computer Vision (ICCV)*, pages 1593–1600, 2009. xix, 1, 15, 27, 28, 39, 40, 43, 56, 63
- [8] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. xix, 28, 29, 43, 46, 51, 58
- [9] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2684–2701, 2019. xix, 29, 30, 43, 57
- [10] Moustafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A Hierarchical Deep Temporal Model for Group Activity Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1971–1980, 2016. xix, 3, 18, 30, 31, 64, 66, 77
- [11] Arash Vahdat, Bo Gao, Mani Ranjbar, and Greg Mori. A discriminative key pose sequence model for recognizing human interactions. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1729–1736, 2011. 1, 15, 16, 39, 40
- [12] M S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming video. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1036–1043, 2011. 40
- [13] Yimeng Zhang, Xiaoming Liu, Ming-Ching Chang, Weina Ge, and Tsuhan Chen. Spatio-Temporal Phrases for Activity Recognition. In *Springer European Conference on Computer Vision (ECCV)*, pages 707–721, 2012. 40
- [14] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2650–2657, 2013. 1, 15, 16, 39, 40, 56
- [15] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015. 1, 16, 40, 56
- [16] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging LSTMs

- to Anticipate Actions Very Early. In *IEEE International Conference on Computer Vision (ICCV)*, pages 280–289, 2017. 39, 40, 56, 88
- [17] Yuge Shi, Basura Fernando, and Richard Hartley. Action Anticipation with RBF Kernelized Feature Mapping RNN. In *Springer European Conference on Computer Vision (ECCV)*, pages 305–322, 2018. 40, 51, 56
- [18] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Leveraging Structural Context Models and Ranking Score Fusion for Human Interaction Prediction. *IEEE Transactions on Multimedia (TMM)*, 20(7):1712–1723, 2018. 1, 16, 56
- [19] Berthold K. P. Horn and Brian G. Schunck. Determining Optical Flow. In *SPIE Technical Symposium East*, pages 319–331, 1981. 2
- [20] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks. In *AAAI*, pages 3697–3703, 2016. 2, 14, 40, 55
- [21] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In *Springer European Conference on Computer Vision (ECCV)*, pages 816–833, 2016. 2, 14, 55, 57, 58
- [22] Yanhua Yang, Cheng Deng, Shangqian Gao, Wei Liu, Dapeng Tao, and Xinbo Gao. Discriminative Multi-instance Multitask Learning for 3D Action Recognition. *IEEE Transactions on Multimedia (TMM)*, 19(3):519–529, 2017.
- [23] Jun Liu, Gang Wang, Ling Yu Duan, Kamila Abdiyeva, and Alex C. Kot. Skeleton-Based Human Action Recognition with Global Context-Aware Attention LSTM Networks. *IEEE Transactions on Image Processing (TIP)*, 27(4):1586–1599, 2018. 14, 46, 55, 57, 58
- [24] Qiuhong Ke, Mohammed Bennamoun, Hossein Rahmani, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning Latent Global Network for Skeleton-based Action Prediction. *IEEE Transactions on Image Processing (TIP)*, 29:959 – 970, 2019. 2, 40
- [25] Hao-shu Fang, Shuqin Xie, Yu-wing Tai, and Cewu Lu. RMPE: Regional Multi-person Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, 2017. 2, 36

- [26] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(1):172–186, 2019. 2, 27, 36, 40, 52, 74
- [27] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1282–1289, 2009. 3, 17, 64
- [28] Tian Lan, Yang Wang, Weilong Yang, Stephen N. Robinovitch, and Greg Mori. Discriminative Latent Models for Recognizing Contextual Group Activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, 2012. 18
- [29] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2596–2605, 2015. 18, 64
- [30] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorbun, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3043–3053, 2016. 64
- [31] Xin Li and Mooi Choo Chuah. SBGAR: Semantics Based Group Activity Recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2895–2904, 2017. 64
- [32] Tianmin Shu, Sinisa Todorovic, and Song Chun Zhu. CERN: Confidence-energy recurrent network for group activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4255–4263, 2017. 18, 77
- [33] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3425–3434, 2017. 18, 64, 77
- [34] Mostafa S. Ibrahim and Greg Mori. Hierarchical Relational Networks for Group Activity Recognition and Retrieval. In *Springer European Conference on Computer Vision (ECCV)*, pages 721–736, 2018. 19, 64, 77

- [35] Lihua Lu, Yao Lu, Ruizhe Yu, Huijun Di, Lin Zhang, and Shunzhou Wang. GAIM: Graph Attention Interaction Model for Collective Activity Recognition. *IEEE Transactions on Multimedia (TMM)*, 22(2):524–539, 2020. 3, 19, 64
- [36] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, and Ahmad Nickabadi. A Multi-Stream Convolutional Neural Network Framework for Group Activity Recognition. *arXiv preprint arXiv:1812.10328*, 2018. 3, 19, 64, 77
- [37] Lihua Lu, Huijun Di, Yao Lu, Lin Zhang, and Shunzhou Wang. Spatio-temporal attention mechanisms based model for collective activity recognition. *Signal Processing: Image Communication*, 74:162–174, 2019. 3, 19, 64, 77
- [38] Enrique Nievas Bermejo, Oscar Deniz Suarez, Gloria Bueno García, and Rahul Sukthankar. Violence detection in video using computer vision techniques. *Computer Analysis of Images and Patterns (CAIP)*, 6855:332–339, 2011. 4, 20, 22, 26, 88, 90, 91
- [39] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–6, 2012. 4, 22, 23, 26, 90, 91
- [40] Sadegh Mohammadi, Alessandro Perina, Hamed Kiani, and Vittorio Murino. Angry Crowds : Detecting Violent Events in Videos. In *Springer European Conference on Computer Vision (ECCV)*, pages 1–16, 2016. 22, 23, 26, 88
- [41] Vu Lam, Sang Phan, Duy Dinh Le, Duc Anh Duong, and Shin’ichi Satoh. Evaluation of multiple features for violent scenes detection. *Springer Multimedia Tools and Applications (MTAP)*, 76(5):7041–7065, 2017. 24, 25, 26, 96
- [42] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Multi-modal Data Fusion for Sensitive Scene Localization. *Information Fusion*, pages 1–39, mar 2018. 4, 21, 24, 25, 26, 88, 92, 96
- [43] S Blunsden and R. B. Fisher. The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(4): 1–11, 2010. 4, 20, 88, 90, 91
- [44] Claire Hélène Demarty, Cédric Penet, Mohammad Soleymani, and Guillaume Gravier. VSD, a public dataset for the detection of violent scenes in movies:

- design, annotation, analysis and evaluation. *Multimedia Tools and Applications*, 74(17):7379–7404, 2015. 4, 20, 21, 88, 90, 91
- [45] Heng Wang and Cordelia Schmid. Action Recognition with Improved Trajectories. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3551–3558, 2013. 14, 24, 25
- [46] Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014. 14, 92, 93, 94, 97
- [47] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Springer European Conference on Computer Vision (ECCV)*, pages 20–36, 2016.
- [48] Yemin Shi, Yonghong Tian, Yaowei Wang, and Tiejun Huang. Sequential Deep Trajectory Descriptor for Action Recognition With Three-Stream CNN. *IEEE Transactions on Multimedia (TMM)*, 19(7):1510–1520, 2017. 14
- [49] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*, pages 1–14, 2017. 14
- [50] Chenyang Si, Ya Jing, Wei Wang, Liang Wang, and Tieniu Tan. Skeleton-Based Action Recognition with Spatial Reasoning and Temporal Stack Learning. In *Springer European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 14
- [51] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-Structural Graph Convolutional Networks for Skeleton-Based Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3590–3598, 2019. 14, 51, 57, 58
- [52] Yanli Ji, Hong Cheng, Yali Zheng, and Haoxin Li. Learning contrastive feature distribution model for interaction recognition. *Journal of Visual Communication and Image Representation*, 33:340–349, 2015. 17, 40, 55
- [53] Huimin Wu, Jie Shao, Xing Xu, Yanli Ji, Fumin Shen, and Heng Tao Shen. Recognition and Detection of Two-Person Interactive Actions Using Automatically Selected Skeleton Features. *IEEE Transactions on Human-Machine Systems*, 48(3):304–310, 2018. 17, 40, 49, 55, 64

- [54] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. Convolutional Relational Machine for Group Activity Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7892–7901, 2019. 18, 77
- [55] Fillipe Souza, Guillermo C. Chávez, Eduardo A. Valle, and Arnaldo de A. Araújo. Violence detection in video using spatio-temporal features. In *SBC Graphics, Patterns and Images (SIBGRAPI)*, pages 224–230, 2010. 21, 22, 26
- [56] Jeho Nam, Masoud Alghoniemy, and Ahmed H Tewfik. Audio-visual content-based violent scene characterization. In *IEEE International Conference on Image Processing (ICIP)*, volume 1, pages 353–357, 1998. 21, 22, 26
- [57] Theodoros Giannakopoulos, Alexandros Makris, Dimitrios Kosmopoulos, Stavros Perantonis, and Sergios Theodoridis. Audio-visual fusion for detecting violent scenes in videos. In *Springer Artificial Intelligence: Theories, Models and Applications*, volume 6040, pages 91–100, 2010. 26
- [58] Thanassis Perperis, Theodoros Giannakopoulos, Alexandros Makris, Dimitrios I. Kosmopoulos, Sofia Tsekeridou, Stavros J. Perantonis, and Sergios Theodoridis. Multimodal and ontology-based fusion approaches of audio and visual processing for violence detection in movies. *Expert Systems with Applications*, 38(11):14102–14116, 2011. 21, 26
- [59] Ankur Datta, Mubarak Shah, and Niels Da Vitoria Lobo. Person-on-person violence detection in video data. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 433–438, 2002. 22, 26
- [60] Tao Zhang, Wenjing Jia, Baoqing Yang, Jie Yang, Xiangjian He, and Zhonglong Zheng. MoWLD: a robust motion image descriptor for violence detection. *Springer Multimedia Tools and Applications (MTAP)*, 76(1):1419–1438, 2015. 22, 23, 26
- [61] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, and Anderson Rocha. Temporal Robust Features for Violence Detection. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2017. 22, 23, 25, 26, 93, 95, 96
- [62] Ivan Laptev. On space-time interest points. *Springer International Journal of Computer Vision (IJCV)*, 64(2-3):107–123, 2005. 22

- [63] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, Matti Pietikainen, Xilin Chen, and Wen Gao. Wld: A robust local image descriptor. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 32(9):1705–1720, 2010. 23
- [64] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-Up Robust Features (SURF). *Springer Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008. 23
- [65] Florent Perronnin and Christopher Dance. Fisher Kernels on Visual Vocabularies for Image Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 23
- [66] Qichao Xu, John See, and Weiyao Lin. Localization Guided Fight Action Detection in Surveillance Videos. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 568–573, 2019. 24, 26
- [67] Paolo Rota, Nicola Conci, Nicu Sebe, and James M. Rehg. Real-life violent social interaction detection. In *IEEE International Conference on Image Processing (ICIP)*, pages 3456–3460, 2015. 24, 26, 90, 91
- [68] Qi Dai, Rui-wei Zhao, Zuxuan Wu, Xi Wang, Zichen Gu, Wenhai Wu, and Yu-gang Jiang. Fudan-Huawei at MediaEval 2015 : Detecting Violent Scenes and Affective Impact in Movies with Deep Learning. In *Proceedings of the MediaEval 2015 Workshop*, pages 5–7, 2015. 24, 25, 26
- [69] Mats Sjöberg, Yoann Baveye, Hanli Wang, Vu Lam Quand, Bogdan Ionescu, Emmanuel Dellandréa, Markus Schedl, Claire-hélène C.-H. Demarty, and Liming Chen. The MediaEval 2015 Affective Impact of Movies Task. In *MediaEval 2015 Workshop*, pages 1–3, 2015. 24
- [70] Nasaruddin Nasaruddin, Kahlil Muchtar, Afdhal Afdhal, and Alvin Prayuda Juniarta Dwiyanoro. Deep anomaly detection through visual attention in surveillance videos. *Springer Journal of Big Data*, 7(1):87, 2020. 25, 26
- [71] Seymanur Akti, Gozde Ayse Tataroglu, and Hazim Kemal Ekenel. Vision-based Fight Detection from Surveillance Cameras. In *IEEE International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6, 2019. 26

- [72] Dylan Josh Coming Lopez and Cheng-chang Lien. Real-Time Human Violent Activity Recognition Using Complex Action Decomposition. In *IEEE International Computer Symposium (ICS)*, pages 360–364, 2020. 26
- [73] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958. 31
- [74] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009. 31, 32
- [75] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. 32
- [76] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4967–4976, 2017. 33, 41, 45, 49, 69
- [77] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997. 34
- [78] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*, pages 1–15, 2015. 35
- [79] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015. 35, 72
- [80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017. 35, 72
- [81] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 6.2.2.3 Softmax Units for Multinoulli Output Distributions. In *Deep Learning*, pages 180–184. 2016. 36
- [82] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2016. 36

- [83] Jun Liu, Henghui Ding, Amir Shahroudy, Ling-Yu Duan, Xudong Jiang, Gang Wang, and Alex C. Kot. Feature Boosting Network For 3D Pose Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 14(8):1–8, 2019.
- [84] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep High-Resolution Representation Learning for Human Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5686–5696, 2019. 36
- [85] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3686–3693, 2014. 37
- [86] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Springer European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 37
- [87] Xiaoyang Wang and Qiang Ji. Hierarchical context modeling for video event recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(9):1770–1782, 2017. 39, 40, 56
- [88] Zhengyou Zhang. Microsoft Kinect Sensor and Its Effect. *IEEE Multimedia*, 19(2):4–10, 2012. 40
- [89] Meng Li and Howard Leung. Multiview Skeletal Interaction Recognition Using Active Joint Interaction Graph. *IEEE Transactions on Multimedia (TMM)*, 18(11):2293–2302, 2016. 40
- [90] Mauricio Perez, Jun Liu, and Alex C Kot. Interaction Recognition Through Body Parts Relation Reasoning. In *IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 268–280, 2020. 40, 66
- [91] Xingyang Cai, Wengang Zhou, Lei Wu, Jiebo Luo, and Houqiang Li. Effective Active Skeleton Representation for Low Latency Human Action Recognition. *IEEE Transactions on Multimedia (TMM)*, 18(2):141–154, 2016. 40
- [92] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid. Learning Clip Representations for Skeleton-Based 3D Action Recognition. *IEEE Transactions on Image Processing (TIP)*, 27(6):2842–2855, 2018.

- [93] Jian Liu, Hossein Rahmani, Naveed Akhtar, and Ajmal Mian. Learning Human Pose Models from Synthesized Data for Robust RGB-D Action Recognition. *Springer International Journal of Computer Vision (IJCV)*, 127(10): 1545–1564, 2019. 40
- [94] Fisher Yu and Vladlen Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations (ICLR)*, pages 1–13, 2016. 50
- [95] Xiangbo Shu, Jinhui Tang, Guo-Jun Qi, Wei Liu, and Jian Yang. Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 1811.00270:1–1, 2019. 51, 56
- [96] Wenbo Li, Longyin Wen, Mooi Choo Chuah, and Siwei Lyu. Category-Blind Human Action Recognition: A Practical Recognition System. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4444–4452, 2015. 55
- [97] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2136–2145, 2017. 55, 56
- [98] Yu Kong and Yun Fu. Close human interaction recognition using patch-aware models. *IEEE Transactions on Image Processing (TIP)*, 25(1):167–178, 2016. 56
- [99] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C. Kot. Global Context-Aware Attention LSTM Networks for 3D Action Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1647–1656, 2017. 57, 58
- [100] Jun Liu, Amir Shahroudy, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Skeleton-Based Online Action Prediction Using Scale Selection Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(6):1453–1467, 2019. 57, 58
- [101] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-Matching Network for Temporal Action Proposal Generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3888–3897, 2019. 62, 96

- [102] Shuming Liu, Xu Zhao, Haisheng Su, and Zhilan Hu. TSI: Temporal Scale Invariant Network for Action Proposal Generation. In *Springer Asian Conference on Computer Vision (ACCV)*, volume 12626 LNCS, pages 530–546, 2021.
- [103] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast Learning of Temporal Action Proposal via Dense Boundary Generator. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 11499–11506, 2020. 62, 96
- [104] C. Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local SVM approach. In *IEEE International Conference on Pattern Recognition (ICPR)*, pages 32–36, 2004. 63
- [105] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1402, 2005. 63
- [106] Yu Kong, Yunde Jia, and Yun Fu. Learning human interaction by interactive phrases. In *Springer European Conference on Computer Vision (ECCV)*, pages 300–313, 2012. 63
- [107] Paresh R. Kamble, Avinash G. Keskar, and Kishor M. Bhurchandi. Ball tracking in sports: a survey. *Springer Artificial Intelligence Review*, 52(3): 1655–1705, 2019. 75
- [108] Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. Participation-Contributed Temporal Dynamic Model for Group Activity Recognition. In *ACM International Conference on Multimedia*, pages 1292–1300, 2018. 77
- [109] Hina Uttam Keval. *Effective design, configuration, and use of digital CCTV*. Phd, University College London, 2009. 87
- [110] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1914–1923, 2016. 88
- [111] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1942–1950, 2016.

- [112] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5783–5792, 2017. 88
- [113] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 88, 100
- [114] Alberto Montes, Amaia Salvador, Santiago Pascual, and Xavier Giro-i Nieto. Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks. *arXiv preprint arXiv:1608.08128*, pages 1–5, 2016. 92, 93, 95, 96
- [115] Qi Dai, Zuxuan Wu, Yu Gang Jiang, Xiangyang Xue, and Jinhui Tangz. Fudan-Njust at mediaeval 2014: Violent Scenes Detection using Deep Neural Networks. In *Proceedings of the MediaEval 2014 Workshop*, pages 1–2, 2014. 96
- [116] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*, pages 1–10, 2014. 96
- [117] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *Springer European Conference on Computer Vision (ECCV)*, volume 3024, pages 25–36, 2004. 97
- [118] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 97
- [119] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *Springer European Conference on Computer Vision (ECCV)*, pages 143–156, 2010. 97
- [120] Markus Schedi, Mats Sjöberg, Ionut Mironica, Bogdan Ionescu, Vu Lam Quang, Yu-Gang Jiang, and Claire-Helene Demarty. VSD2014: A dataset for violent scenes detection in hollywood movies and web videos. In *IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6, 2015. 99

- [121] Alexander Gorban, Haroon Idrees, Yu-Gang Jiang, Amir R. Zamir, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. THUMOS Challenge: Action Recognition with a Large Number of Classes. <http://www.thumos.info/>, 2015. 100