RELATIONAL LEARNING USING BODY PARTS FOR HUMAN ACTIVITY RECOGNITION IN VIDEOS

Keerthy Kusumam
MSc. Advanced Computer Science
2011 -2012

The candidate confirms that the work submitted is her own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material, which is obtained from another source, may be considered as plagiarism.

(Signature)

Acknowledgment

I would like to extend my sincere gratitude to Prof. Anthony G. Cohn for giving me an opportunity to work on this project. I am indebted to him for his invaluable suggestions and constructive discussions which provided a great source of inspiration while doing this project.

I express my gratefulness to Dr. MuraliKrishna Sridhar who gave me constant support throughout the project work. I would also like to thank Feng Gu for his suggestions and help.

I would like heartfelt gratitude to Ms. Elaine Duffin for her overwhelming support in reading my work and giving comments that gave me some of the crucial insights on the project work.

I also thank Aryana Tavanai for his support and encouragement which aided me to complete this project on time.

I would also thank my friends and colleagues for motivating me and providing support throughout the course of this project, without which I would not have completed this work.

Summary

Human activity recognition is a prominent area of computer vision research today. It has various applications especially in fields like video surveillance. The main idea of this project is to model human activities from videos using body parts of the interacting person and apply a graph-based relational representation and learning to it. This project investigates if such a fine grained representation of human activities would be beneficial to describe SOME verbs.

Human interactions are described as changes in qualitative spatio-temporal relationships between the body parts of a person and an interacting object. These changes in spatio-temporal relationships would have similar patterns with respect to an event class, which is learned by a relational event model. Graphs are used to represent and identify these changes in qualitative spatio-temporal relationships between the objects.

By implementing a framework designed on the basis of this hypothesis, it was observed that using body parts for modelling some interactions were better than a baseline approach which considered a person as a single object. The project also investigates different ways these interactions can be described using body parts.

Table of Contents

Chapter 1

Introduction

1.1. Overview

Human activity recognition is a prominent area of interest in computer vision today (Lavee, Rivlin, & Rudzsky, 2009). Recognition systems capable of detecting human activities analyse an input video and identify the events that involve human participation. This can lead to a variety of applications that include smart surveillance systems, patient monitoring systems, service robotics, intelligent environments etc.

Human activities can be broadly categorised into events that represent actions and interactions. Actions are activities involving a single person. These are composed of simple gestures (stretching hand, raising leg) occurring in a temporal order, like 'walking', 'running', 'punching' (Lavee et al., 2009). Interactions model activities that involve more than one person or objects. For example, activities like 'throw', 'kick', 'hit' are interactions between a person and an object while activities like 'give', 'take' or 'exchange' are interactions between two persons in a scene.

This project investigates how an interaction can be modelled using the body parts of a person and an object. This follows from the intuitive idea that some interactions are better described when the part of the body involved in it is specified. For instance, 'kick' is an event where a person interacts with a ball and 'throw' is also an event where the person interacts with the ball. One can distinguish between these events if 'kick' is described as an interaction where the ball moves away from the leg of a person and 'throw' as an interaction where the ball moves away from the hand of a person. The concept of 'qualitative spatio-temporal relationships' proposed by Sridhar et al. (2010) is used to represent the spatio-temporal relationships between the bounding regions of interacting objects. This project seeks to identify the role of using a fine grained representation of humans in order to model their interactions with other objects.

The following sections describe the main aims and objectives laid out in this project, the research methodology adopted and the basic framework followed to implement an activity recognition system. Some of the questions which motivate the research undertaken by the project and an overview of the chapters explained in the report are also included in this chapter.

1.2 Aim and Objectives

The aim of this project is to investigate the role of spatial relationships between the different body parts of a person (or between body parts of a person and an object) in modelling human activities in a video. The framework consists of applying Computer Vision and Knowledge Representation and Machine Learning techniques to a set of video data for human activity recognition. In order to build the framework for activity recognition and fulfil the aim, the following steps are necessary:

1. Choose a set of verbs that represent human actions and interactions like ‘walk’, ‘run’, ‘lift’, ‘throw’, ‘catch’, ‘kick’, ‘bounce’ etc.
2. Choose a set of videos that represent these verbs.
3. Apply pre-processing of the video data to obtain the detections of persons and other objects interacting in each frame. Manual tracks are used initially.
4. Obtain the body parts of the person within detection window using a part based model proposed by Yang and Ramanan(2011).
5. Choose a set of qualitative spatial relationships, like RCC8 or QTC6 [6] to represent the spatial relationships between extracted body parts.
6. Create spatio-temporal relationship graphs pertaining to the body parts (and objects) obtained in each frame.
7. Mine the obtained set of interaction graphs to extract sub graphs for feature representation using ‘Bag-of-Graphlets’ approach.
8. Apply supervised learning on the obtained feature sets to learn models to detect the human activities.

1.3 Minimum Requirements

The following minimum requirements are set for the project,

1. Build a supervised model of human activities by considering relationships between body parts of person in the chosen dataset.
2. Build a software system which is able to use the above model for detecting human actions and interactions.
3. Evaluate the system including investigation of sensitivity parameter adjustments and dataset variety.

1.4. Research Methodology

The methodology is 'evolutionary prototyping' where the approach is to build an initial prototype system from a chosen video dataset and then evaluate the prototype and based on its performance. This approach would incrementally improve the prototype towards the final system. The methodology is best suited for this scenario because, even though the aims and objectives of this project are laid out clearly, there is an uncertainty as to whether the results can be achieved by the experiments.

The initial prototype was aimed to recognise simple actions and interactions, like 'kick' and 'throw', representing human activities on a chosen small set of videos. The initial prototype used only few body parts, such as hands and legs, participating in the interactions. Further prototypes investigated the use of different types of spatial relationships between the body parts and other objects in representing events. Further refinements were included by adding or removing body parts as required by the scenario. The spatio-temporal relationships between objects were represented as interaction graphs using the software developed by Sridhar (Sridhar, Cohn, & Hogg, 2010). The initial prototype was evaluated for the task of detecting human activities and was incrementally improved towards producing the final system that satisfied the minimum requirements and achieved the set objectives.

1.5. Framework Overview

The activity recognition framework is designed based on the following methods.

1.6.1. Collection of video data: A set of videos which serves as input data to the activity recognition framework is chosen. These videos ideally contain an interaction event like 'kick', 'throw, 'catch' or 'bounce'. Use the manually labelled bounding boxes around participant objects in each video, like a person or a ball, as object tracks.

1.6.2. Body Part Detection: Use a pose estimation technique to extract body parts from a detection window identifying a person. Encode these part locations as refined object tracks.

1.6.3. Qualitative Spatio-Temporal Relationships: Determine a set of qualitative spatial and temporal relationships occurring between the tracks of body parts and other interacting objects during the course of a video. Obtain these pair-wise spatio-temporal relationships as interactions.

1.6.4. Graph Based Relational Representation: A relational description of interactions obtained from above step is represented using interaction graphs.

1.6.5. Feature Representation: In order to represent each video interaction graph as a feature vector, the 'Bag-of-graphlets' method described in chapter 6 is used. The idea is to mine frequently occurring sub-graphs and the feature vector corresponding to a video interaction graph is represented as a histogram of these sub-graphs.

1.6.6. Event Learning: A supervised learning approach using machine learning algorithms is used to learn models to detect the human activities.

1.6. Research Questions

The following research questions were set to be investigated at the start of this project.

Will using spatial relationships between body parts be able to distinguish SOME verbs better than when considering a person as a single entity?

Can the framework capture significant patterns of spatial relationship sequences between body parts and the interacting object?

How would using different granularities of body parts help in recognising human activities?

What type of qualitative spatial relationships can model the events better?

Will the feature representation using body parts be distinctive with respect to an event class?

1.7. Thesis Overview

Chapter 2 is a literature review which discusses the current research trends in the areas of human activity recognition and provides a motivation for the proposed approach put forth by this project. It also reviews the existing approaches from which the framework proposed in this project derives its concepts.

Chapter 3 is an overview of the activity recognition framework laid down by this project. It gives the big picture of the different stages involved in the implementing the proposed approach.

Chapter 4 discusses the body part detection framework used in this project in detail. It includes the background theory and how the detection framework is adapted for the purposes of this project.

Chapter 5 explains how the interactions between the body parts and objects are represented as graphs. The concept of 'episodes' and 'interaction graphs' are discussed in detail in this chapter.

Chapter 6 gives a detailed explanation of the relational learning framework used in the project and also reviews the machine learning techniques used.

Chapter 7 details all the experiments that were carried out for the investigating different research questions and evaluating the proposed approach. The results and conclusions drawn from each experiment is discussed here.

Chapter 8 includes all the findings derived from the experiment section and the implementation of the proposed approach. The scenarios for which the proposed approach may not be desirable are presented. Also the scope of the proposed approach and future work which could extend its scope is discussed in this section.

1.8 Resources

For the purpose of implementing the detection framework, the code available from Yang and Ramanan (2011) was used. These algorithms detected body parts from a static image. A basic version of the two available versions was used to implement the purposes of this the project.

Relational Description of Video Scenes (REDVINE) software which was developed by Sridhar (2010) for implementing relational learning using interaction graphs. Hence this software was used to extract the necessary statial relationships and implement learning and graph mininig techniques. This saved the implementation time although some of the programs were modified as referenced in Appendix C.

Chapter 2

Literature Review

2.1. Introduction

Learning human activities from videos is an area of prime focus in computer vision research. Human activity recognition has a wide variety of applications especially in smart video surveillance systems which can detect anomalous activities like a fight or a robbery from CCTV monitored videos. Intelligent environments are also application of human activity analysis where patients and older people are monitored automatically. Intelligent home environments that respond to human gestures are a giant leap made possible with the development of this technology.

The purpose of this chapter is to provide an understanding of the approaches which are currently used for human activity analysis and the motivation to use the proposed approach featuring body parts. This chapter is divided into three main sections; one reviewing the latest trends in human activity recognition and another section laying out the motivation for proposed approach. The last section encompasses a description of the concepts that are used to design the framework of this project. This includes computer vision methods for estimating human poses, knowledge representation concepts of qualitative spatial and temporal relations, relational representations of events as graphs and machine learning concepts for relational learning and classification.

2.2. Current Research

In a recent survey by Aggarwal and Ryoo (2011), human activities are categorised based on the level of complexity. Gestures can be regarded as atomic events which have simple movements like ‘withdrawing’ a body part or ‘stretching a leg’ involving a single person. Actions form the next level, when more than one simple gesture occurs over a period of time, like ‘waving hands’ (stretching arms) and ‘walking’. Actions also encompass a single person. Interactions are more complex and involve the participation of another object or a

person, like ‘kick’, ‘punch’, ‘throw’ etc. The last category is group activities which occupy the highest complexity level where there are multiple persons and objects interacting in the scene like ‘marching’ or ‘group fighting’ (Aggarwal & Ryoo, 2011).

In many approaches, salient features are identified in a video and used for description of the events taking place in the video. These features can be classified as low level, mid level and high level (Lavee et al., 2009). Low level features are based on pixel level properties, like colour, texture and gradient. The mid level features include those where low level features are abstracted into a set of objects or body parts. This can be a good approach as modelling objects participating in the video gives a better characterisation of the events. High level features combine low level or mid level features into higher level semantics like sequences or relational entities like logic predicates (Krishna S. R. Dubba, Cohn, & Hogg, 2010) or graphs (Sridhar et al., 2010).

There are different state-of-the-art approaches to analyse human activities from videos. These methods include space-time approaches, sequential and exemplar based approaches, state space models, syntactic approaches which use grammars and description based approaches that can represent high level activities (Aggarwal & Ryoo, 2011). The following sections discuss these approaches in detail.

2.2.1. Methods

Space-time approaches are one of the most adopted methods in human activity recognition. In this approach, an input video is represented as a 3-dimensional (3-D XYT) volume in which the dimensions are space (XY) and time (T). Some methods extract a set of features from the 3-D XYT volumes and represent it as a feature vector. Hence a three dimensional space-time volume would correspond to an activity, which may be labelled and compared with the space-time volume of an unseen video to recognise that activity (Aggarwal & Ryoo, 2011).

Campbell & Bobick recognised human actions in ballet movements by modelling the body part movements as curves in phase space, where each axis represents a body part, for instance the joint angles (Campbell & Bobick, 1995). In phase space, the points correspond to a static state of the person while movements make a curve. Given a test video, the system

would verify if the points generated by the video are on the learnt curves that is representative of a movement. This method was applied to recognition of actions that involved a single human like dance movements where significant changes relative to body parts occurred.

Laptev and Lindberg extracted local features from the 3-D XYT volumes as 'sparse spatio-temporal interest points (STIP)' features to represent human activities (Laptev & Lindeberg 2003). The interest points are detected in the spatio-temporal domain using Harris and Forstener point operators, which detects interest points that are distinctive. These spatio-temporal points are scale invariant in the 3-D XYT space and can depict changes in motion patterns like an object changing its direction in a bounce activity. The interest points hence correspond to meaningful events as they characterize non-constant motion. Laptev et al. (2008) extended this work by using spatio-temporal histograms formed by dividing the space-time volume into several grids. This gives an idea of where the event has occurred in the video since one can see into which grid a spatio-temporal interest point might fall.

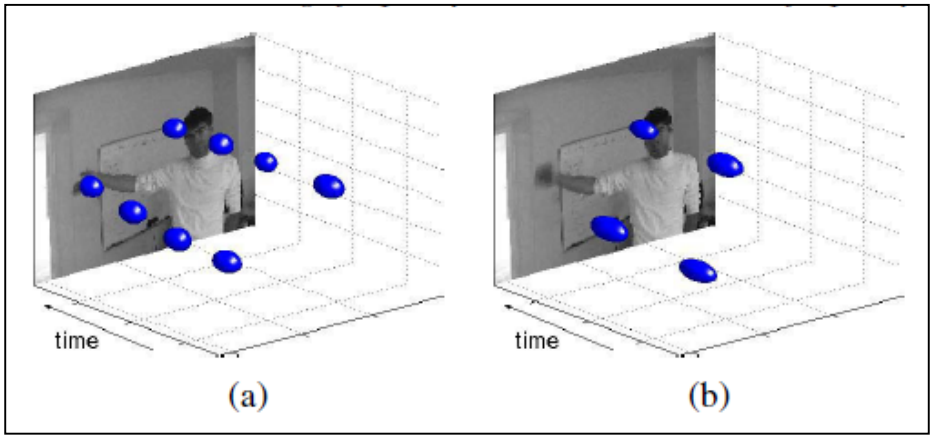


Figure 1: Space- time interest point detection on a hand waving video sequence. (a) STIP features for hand waving at high frequency (b) STIP features for hand waving at low frequency(Laptev & Lindeberg 2003).

Space-time approaches perform better in periodic actions than non-periodic actions and simple human activities and are quite robust against clutter and background noise. However, one disadvantage is that it is not suitable for recognising multiple or more complex activities that are not periodic in nature. It also uses the 'bag- of- words' approach which ignores the spatio-temporal relationship between the features which is not desirable in modelling high level activities (Aggarwal & Ryoo, 2011).

Sequential Approaches: Another direction is to model input videos as a sequence of observations and infer that an activity has occurred in the video if that particular sequence occurs in the video. An example of this method is demonstrated by Veeraraghavan et al. (2006) where activity sequences are modelled by considering the intra- and inter- personal speed changes in performing the same activity. They attain this by modelling a time warping function and showed considerable accuracy in modelling human actions like 'throw', 'kick' and 'pick up'.

State-based approaches represent human actions as a sequence of states. In this approach an activity is modelled using a statistical model which has a certain probability. Hidden Markov Models (HMM) and Dynamic Bayesian Networks (DBN) are used extensively in this approach where human actions have a set of hidden states. The underlying concept is that at each time frame in a video, the person is assumed to be in one state and there is a transition to another state in the following time frame. The states are recorded as observations or features and their transition probabilities are calculated and used for training a model for an activity.

Oliver et al. (2000) introduced an approach using a variant of HMM called coupled HMM (CHMM) to overcome the limitations of HMM to model complex activities involving more than one person or objects. This combines multiple HMMs where each HMM models the motion of a single person or object. This could recognise events involving two persons like 'meeting', 'approaching' etc. Another approach was used by Natarajan and Nevatia (2007) with Coupled Hidden semi Markov models which extended CHMM to characterise the duration of sub-events in each activity.

State space methods can recognise non-periodic activities like interactions between two persons as shown in Oliver et al. (2000) and Natarajan & Nevatia (2007) and in sign language recognition. A probabilistic analysis is made but these methods have been shown to require large training data as the complexity of activities increases.

Damen & Hogg proposed a framework for recognising and linking visually confusing events like dropping a bicycle and picking it up (Damen & Hogg, 2009). They applied Bayesian networks using Markov Chain Monte Carlo (MCMC) to recognise and relate sub events. This is an example of high level activity recognition using a statistical model.

This method is suitable for recognising activities that occur in a sequential order but would not perform well where there are concurrent simultaneous activities taking place.

Syntactic approaches use a grammar or a set of specific rules that can generate strings of a formal language (Aggarwal & Ryoo, 2011). The activities can be represented as a set of rules using a string of atomic actions and these rules are parsed in order to recognise an activity. Context-free grammars and stochastic context free grammars (SCFG) are two common methods implementing this technique. Ivanov and Bobick implemented a syntactic approach to model high level activities involving multiple object interactions using HMM and SCFG (Ivanov & Bobick, 2000).

Joo et al. use attribute grammars for recognising events. Attribute grammars are an extension to SCFG and are able to describe the constraints on features in atomic actions (Joo et al., 2006). Anomalous events are detected when the unseen video input does not abide by the grammar rules or satisfy the given constraints. Events pertaining to parking lots based on the interaction between the cars and humans were detected, like 'car parking', 'drop – off', 'park-out' etc.

These syntactic approaches provide a probabilistic framework that is robust against noise but requires the sub events to be sequential or occur in a temporal order. Hence other high level approaches are required to model activities where multiple sub-events co occur.

Description based approaches: The high level description based approaches model an activity in terms of simpler sub-events and model the spatial, temporal and logical relationships among them. Ryoo and Aggarwal (2006) use a description based approach which can describe composite events based on simple atomic actions. The representation uses CFG and is hierarchical, where poses and gestures are detected from input video data from which actions and interactions are identified. The temporal relationships between different atomic actions are described using Allen's temporal relations (Allen, 1983). Hence concurrent activities can be recognised using this method.

Another approach is based on logic and relational learning which can be said to be more expressive than grammar based models since it encodes complex propositions and functions. In work done by Dubba et al. (2010), Inductive Logic Programming is used, which offers a framework to learn logical relationships from the input. Here complex activities occurring in an aircraft domain like loading and unloading of trolleys, where multiple sub-events occur are considered. An advantage of the relational learning approach is that it can model multiple parallel events.

2.3. Motivation for Proposed Approach

This project is based on the approach proposed by Sridhar et al. (2010) which focuses on using a higher level representation of events by modelling spatio-temporal relationships between the objects to identify the corresponding activities. The spatial and temporal relationships between objects are represented using graphs and a relational learning paradigm is used for activity recognition. The idea of using pose estimation to describe human activities has been explored before in human activity recognition research.

The relational representation of interactions between humans and objects as proposed by Sridhar et al. (2010) was based on capturing patterns of spatio-temporal relationships between these objects. It was also seen from the results that some verbs were confused with other events, for instance, the verb 'approach' got confused with the verb 'catch' since both these verbs would entail similar spatio-temporal relationships (Sridhar, Cohn, & Hogg, 2011). Hence it followed from this observation that a finer representation may be useful for describing some verbs like 'kick' or 'catch' where the interaction is evidently between the specific body parts and the object. This formed the motivation of the project to investigate if it would be useful to represent some interactions at a finer level using body parts. Also, this project seeks to identify significant patterns of interactions between the body parts and other objects, if there are any.

The novelty of this approach itself is the use of a fine-grained representation of interactions between humans and objects using body parts in a relational representation and learning framework introduced by Sridhar et al. (2010).

2.4. Background

2.3.1. Introduction

The approach undertaken by the project derives its core concepts from different existing methods. These are methods to detect and extract body parts, qualitative spatio-temporal relationships, which are crucial to the implementation of ideas laid down by the project and the relational learning framework using graphs defined by Sridhar et al. (2010), which constitutes the backbone of this approach.

The following sections review the above aspects in brief and it is the theoretical framework it builds towards is explained in later chapters.

2.3.2. Pose Estimation

Pose estimation is a process that finds the configuration of human body parts in an image. It has wide applications like in analysis of sport videos, human computer interactions in computer games, gait analysis in physiotherapy and motion tracking in video surveillance. This project uses a pose estimation approach to identify different parts of the body from a given input image frame.

Body parts are detected using the state-of-the-art part-based models presented by Yang and Ramanan (2011). The work of Yang and Ramanan (2011) performs human pose estimation from static images and is based on deformable part based models proposed by (Felzenszwalb, Mcallester, Ramanan, & Irvine, 2010).

In their work, Yang and Ramanan (2011) model co-occurrence relationships between mixtures of parts as well as spatial relationships between locations of parts. This maintains constraints that favour particular combinations of parts which can capture local rigidity of parts, for example two parts of the same limb should have same orientation. This results in a detection of body parts at high speed and greater accuracy and makes it a suitable choice to detect the body parts in each frame of a chosen video. The framework produces a 26 different parts where each part has a part type (Yang & Ramanan, 2011). This helps to identify which part of the body a detected region belongs to and hence abstract away only those parts that are required for the purpose of this project.

2.3.3. Qualitative Spatial and Temporal Relations

One of the main aims of this project is to apply different types of qualitative spatial relationships between body parts of a person and an object to describe events. The qualitative spatial relations are used to represent interactions between objects participating in a video. Qualitative spatio-temporal reasoning techniques facilitate interpretation of low level computations into higher level descriptions of the scene (Cohn, Magee, Galata, Hogg, & Hazarika, 2003).

The three main types of spatial relations are based on (i) topology, where one object is with respect to another in space like touch or inside,(ii) orientation (left of or above) and (iii) distance (near or far). Region Connection RCC-8 and RCC-5 (Cohn & Renz, 2007) represent jointly exhaustive and pair-wise disjoint relations characterizing topological relations between a pair of regions in space, refer Figure 2. RCC-5 partition does not take into consideration the difference between the regions that touches the boundary of the two regions as described in RCC-8. Qualitative spatial relationships can either be manually given, e.g. RCC-5 or learned as in Galata et al. (2002).

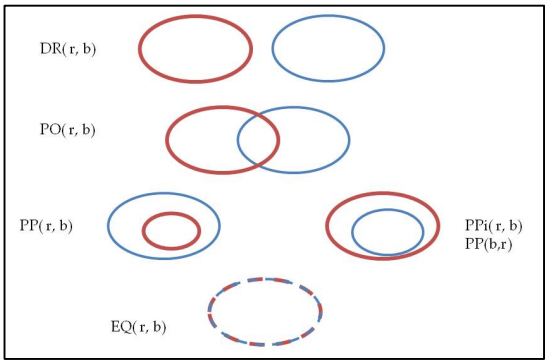


Figure 2: The five spatial relations between two regions r, shown in red & b, shown in blue given by the Rcc-5 calculus

The temporal relationships are modelled using Allen's temporal relations. Allen's temporal interval calculus (Allen, 1983)lays out 13 pair-wise disjoint relations that exist between two time intervals. Pair-wise disjoint relations have no two objects in a domain related by more than one relation, refer Figure 3.

Seven basic predicates Allen defined are 'before', 'meets', 'overlaps', 'during', 'starts', 'finishes', 'equals' along with their inverses of the first six predicates (Figure 3). The predicates 'before' and 'meets' represent relationships that follow a sequence where one happens after another and the rest show simultaneous relationships where one occurs at the same time as the other (there is partial or total overlap of time intervals).

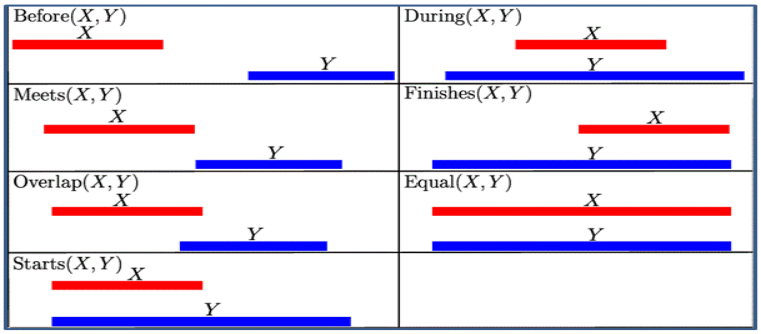


Figure 3: The seven basic predicates of Allen's temporal algebra where X and Y represent two time intervals(Sridhar, 2010).

2.3.4. Relational Learning Approach

This project adopts a graph-based relational learning framework proposed by Sridhar et al. (2010). The main idea is to represent interactions or events as graphs that characterise spatio-temporal relationships between interacting objects. Graphs are a relational description of interactions between objects in this approach. Another method to represent these spatio-temporal relationships is using logical predicates (Dubba et al., 2010).

Under the relational learning approach using graphs, an event is defined by a set of interaction graphs. Interaction graphs are three layered structures where first layer represents the object tracks across the video frames, second layer represents the qualitative spatial relationship existing between the object track pair and the third layer represents temporal relations. Chapter 5 gives a detailed description of this relational representation.

The underlying idea is that events having similar spatio-temporal relationships tend to belong to the same event class. In order to learn and classify two events the respective event graphs are encoded as feature vectors using a 'bag-of-graphlets' approach and the resulting features are given to a machine learning algorithm to learn an event model. The features are represented as histogram of 'graphlets'. 'Graphlets' are the most frequently occurring sub-graphs mined across all interaction graphs corresponding to the videos of a particular event

class. These ‘graphlets’ capture patterns in the interaction graphs and are hence used as features describing a particular event class. These features would be similar for a given event class and distinctive for different event classes. The feature representation is discussed in detail in Chapter 6.

When a test video is given, the task of event detection seen as a problem of finding the most probable covering of an interaction graph corresponding to that video with the interaction sub-graphs using a learned event model (Sridhar, 2010). Chapter 6 describes how an event detection task is translated into finding the most probable covering of an interaction graph.

This project is based on a supervised learning setting where the event labels are given to the classifier during training and a machine learning algorithm, k-Nearest Neighbours which works under supervised setting is applied.

2.5. Conclusion

This chapter gave an overview of the state of the art approaches for human activity recognition. The different approaches are mentioned with their possible advantages and limitations. Then a motivation of undertaking the approach outlined in this project using body parts for representing human interactions is discussed. In the later sections the background concepts that is used for the formulation of the proposed framework is briefly described. This included the pose estimation method used in this project proposed by Yang and Ramanan (2011), Qualitative spatial relationships and Allen’s temporal relations which describe the interactions and a graph-based relational representation and learning framework proposed by Sridhar et al. (2010). The theoretical details of these concepts are discussed in the later chapters.

Chapter 3

Proposed Approach

3.1. Introduction

The proposed approach laid down by this project is designed to use body parts to model interactions. It builds upon an existing work by Sridhar et al. (2010) and uses a state-of-the-art pose estimation technique for the task of identifying human body parts. This chapter gives a high level overview of the activity recognition framework implemented by this project. This framework is referred to as the 'proposed approach' throughout the project.

3.2. Design

A set of events that are representative of human interactions or actions are chosen to be recognised, for example, 'kick', 'throw, 'catch', 'fall', 'bounce'. Videos which contain these events form the input dataset. The detections of participant objects are obtained using hand annotated manual tracks for each video. The proposed approach has three main stages which are follows:

Body Part Detection: Body parts of the person within detection window are obtained using a pose estimation framework proposed by Yang and Ramanan (2011). The detection part provides different body parts of a person and any of these parts can be considered as interacting objects.

Representation of Interactions: Qualitative spatio-temporal relationships are used to characterise interactions representing an event. These spatio-temporal relationships between objects and body parts' tracks from each frame in the video are obtained and represented using interaction graphs for each video.

Event Learning and Recognition: A relational learning framework proposed by Sridhar et al. (2010) is used where feature vectors corresponding to the interaction graphs are obtained using 'Bag-of-Graphlets' approach. Grpahlets have frequently occurring patterns in the interaction graphs pertaining to an event. The feature vectors are given to a classifier to learn

an event model and perform the task of detection and classification of a given video into one of the target events.

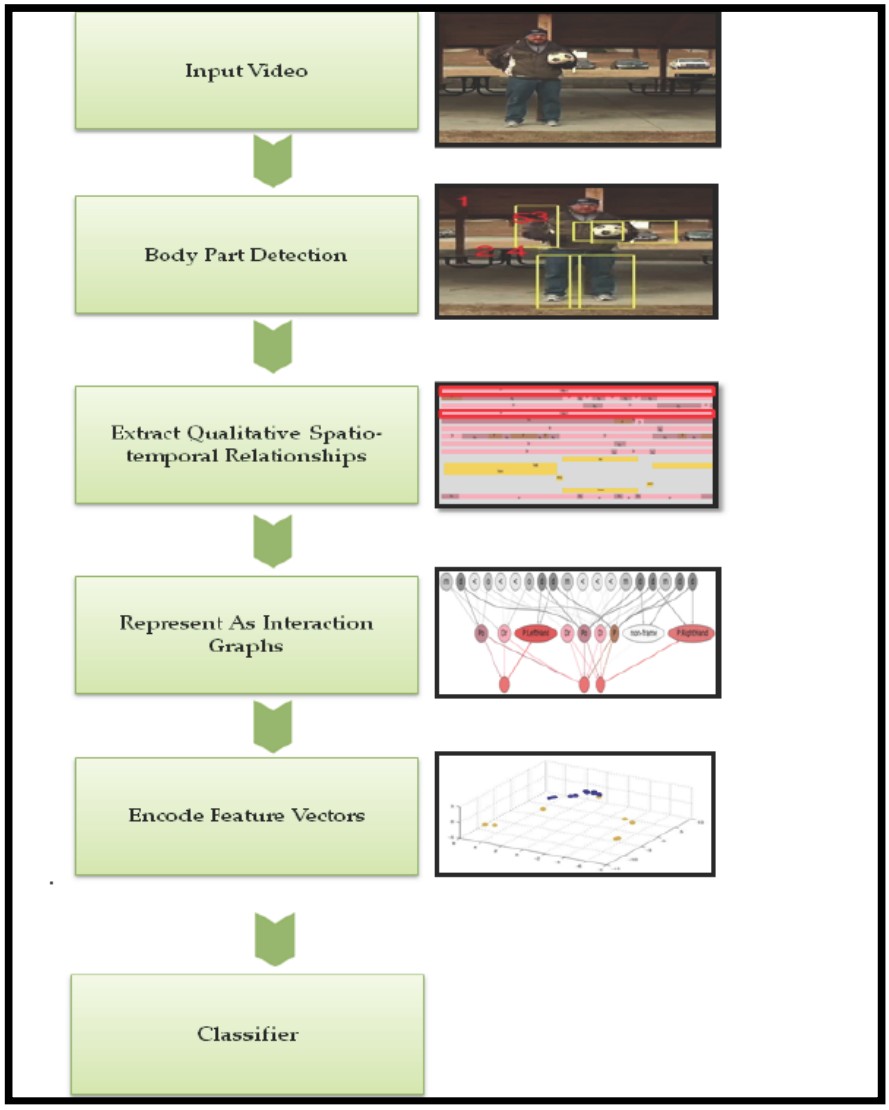


Figure 4: A high level Overview of the Activity Recognition system

3.3. Conclusion

This chapter provides an abstract view of the proposed approach put forth by this project. The three main stages of activity recognition system design, namely Body part detection, Representation of interactions and Event learning and recognition, are defined in brief. Later

Chapter 4

Body Part Detection

4.1. Introduction

The first phase of the proposed approach is to identify body parts of a person from each video frame. The purpose of this chapter is to explain how this task is achieved by the proposed approach laid down by this project. The theoretical framework of the detection method outlined by Yang and Ramanan (2011) and its suitability in the proposed approach design is discussed in the first section. The second section describes how this pose-estimation technique is adapted in the proposed approach.

4.2. Background

The pose estimation task recovers the pose of an articulated object like a human figure, which consists of joints and rigid parts. Pose estimation models are commonly based on pictorial structure representation that encodes objects as a collection of rigid parts that can be connected in different spatial relationships, refer Figure 5. The pictorial structures may be unary templates or pair-wise spring models. The unary templates decompose the object model into local parts templates and spring models impose geometric constraints on part pairs (Yang & Ramanan, 2011).

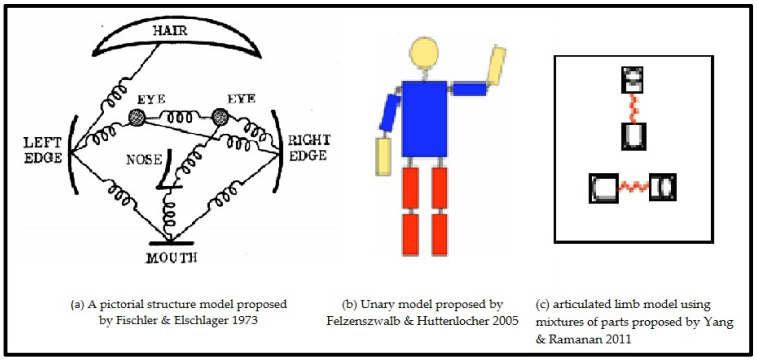


Figure 5:(a) A pictorial structure model proposed by Fischelr and Elschlager (1973) (b) Unary model proposed by Felzenszwalb and Huttenlocher (2005) (c) Articulated limb model using mixtures of parts proposed by Yang and Ramanan (2011)

The work by Yang and Ramanan is a novel representation of body parts which uses the above pictorial structure model that represent spatial relationships between locations of parts and co-occurrence relationships between mixtures of parts (Yang & Ramanan, 2011). Here a 'part type' variable or mixture component is assigned to each part which gives the model flexibility to select between several appearance models. The part type or mixtures maintain constraints that favour particular combinations of parts which can capture local rigidity of parts, for example two parts of the same limb should have the same orientation.

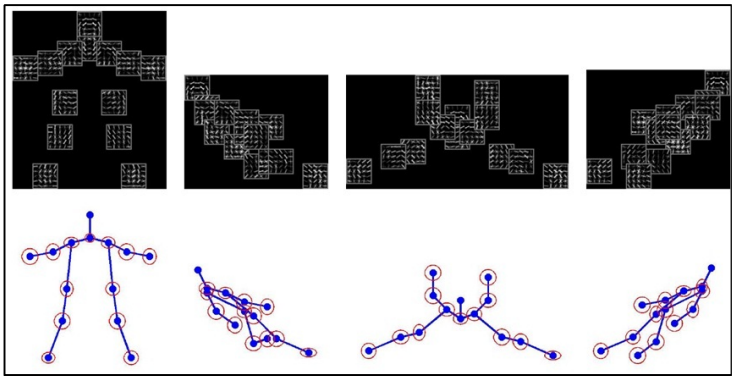


Figure 6: Different body part configurations given by the pictorial structure model used by Yang & Ramanan (2011). A scoring function gives the most likely configuration.

If there are K parts and M mixture components or part types then there are K^M unique pictorial structures (Figure 6). All the pictorial structures are not equally likely and the scoring function which decides the priority for each structure is given by the co-occurrence model. The model used is tree structured and optimization is a attained using dynamic programming (Yang & Ramanan 2011). This results in a detection of body parts at high speed and greater accuracy and makes it a suitable choice to detect the body parts in each frame of a chosen video.

The framework produces 26 different parts where each part is associated with a part type, illustrated in Figure 7. This helps to identify which part of the body a detected region belongs to and hence abstract away only those parts that are required for the purpose of this project. Also, a large number of parts are detected in the frame work which gives a flexibility of modelling granularity of body parts used in the project, e.g. a leg can be represented using its four constituent parts as detected by Yang and Ramanan or as a single part by combining these four parts or modelling only the extremities of the body parts (Yang & Ramanan,2011).

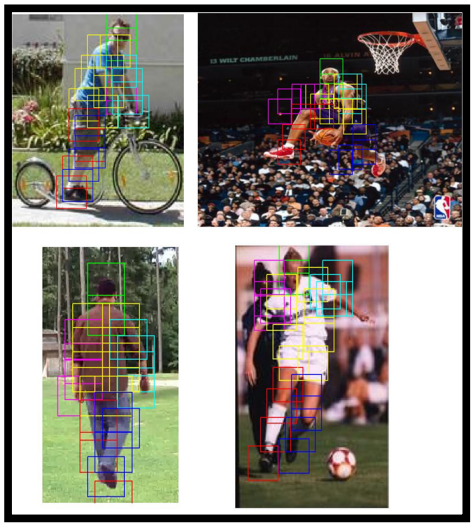


Figure 7: Images with detected body parts returned by the detection framework. The images show different human poses and coherent body part detections.

4.3. Detecting Body Parts

Once the video data set and the corresponding tracks for objects and humans present in each video are obtained, the body part detection method is used to extract body parts of the person interacting with the object. An input video is given to the detection system as a set of images which are cropped to the region of the bounding box surrounding a person in each frame. These are the input image frames fed to the body part detection system.

The detection framework proposed by Yang and Ramanan (2011), as described in detail in the above section, detects 26 different parts of the body identified by its part type on each input image frame. The body parts, Head, Left Shoulder, Torso, Left Hand, Left Leg, Right shoulder, Torso, Right Hand and Right Leg are the main part types of the detected body parts.

From the 26 different parts returned by the detection algorithm, any combination of parts can be used for modelling interactions. For e.g. the body part ‘hand’ can be represented as a combination of its four constituent detected parts into one single part or as four different parts. The project considers the use of modelling different granularities of body parts in the experiments, Figure 8.

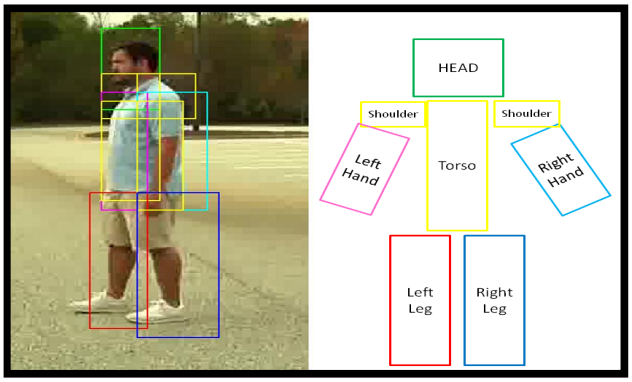


Figure 8: Different body part types identified by the detection framework which can be further used for activity recognition.

The detected body parts are encoded as refined 'object tracks' that contain locations of each body part detected in the step above in each frame of the video. The detection of other interacting object like a ball or a vehicle in each video frame obtained from manual tracks is then appended into the body part track set.

The event ground truth is also defined for each video where the details of the event, for e.g., 'kick' or 'throw' are given. It includes the frame span of the occurring event, giving the idea of when the event happens, the objects involved in the event for e.g. 'Right Leg, Left Leg and Other object' for a 'kick'.

4.4. Conclusion

This chapter discusses the body part detection approach used by this project in detail. First the theoretical framework that the approach builds on is presented. It can be concluded that the body part detection approach proposed by Yang and Ramanan (2011), detects body parts as a set of constituent parts. There are 26 different parts where each part is associated with a part type. This is important for the research undertaken by this project since the freedom to model body parts at any level of abstraction is possible. Also, the speed of detection is high for this approach which makes it an appropriate choice to be used over a sequence of frames. The chapter also reviews in the later section, how this method is implemented for the purpose of this project.

Chapter 5

Representation of Interactions

5.1. Introduction

When the body parts and objects present in a video are extracted, a relational representation is used to describe the interactions among them. Qualitative spatio-temporal relations are used to characterise interactions between objects. This chapter introduces a representation of interactions which are defined by spatio-temporal changes between the object tracks in the form graphs.

The following sections review the types of qualitative spatial and temporal relations used in the proposed approach to model events. It also explains in detail the concept of video-interaction graphs and episodes.

5.2. Qualitative Spatio-Temporal Relationships

Qualitative spatio-temporal relationships provide a high level understanding of a scene since these relationships incorporate spatio-temporal knowledge of the scene. For a particular type of activity, qualitative spatial relations between the interacting objects would be similar.

Five different qualitative spatial relationships and their suitability for modelling human activities are considered in this project, namely, (i) topology, Region Connection Calculus (RCC-5) (ii) direction, DIR4 (iii) relative speed SPD3 (iv) relative size SIZ3 (v) qualitative trajectories QTC6. These are shown in the Figure 9. These relationships were derived from Sridhar et al. (2011).

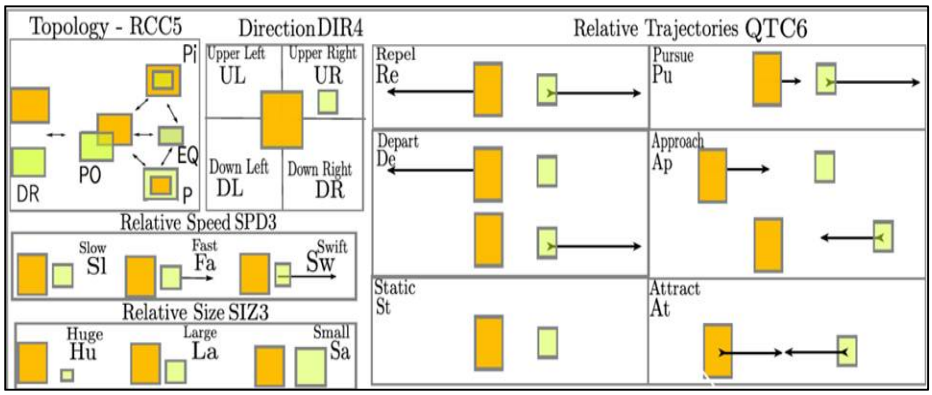


Figure 9: The five different qualitative spatial relationship categories, Topology (RCC-5), Direction (DIR4), Relative speed (SPD3), Relative size (SIZ3) and Relative Trajectories (QTC6) (Sridhar et al., 2011)

Qualitative spatial relationship types chosen for the purposes of event modelling in this project are derived from RCC-5 and QTC-6. The system avoids using spatial relation EQ since it is less likely to occur in video events and combines relations PP and PPi into P. Hence the RCC-5 spatial relationships used are 'Discrete' (DR), 'Partof' (P), 'Partially Overlaps' (PO). QTC-6 spatial relationships represent relative trajectories between the bounding regions given by 'Repel' (Re), 'Depart' (De), 'Static' (St), 'Pursue' (Pu), 'Approach' (Ap) and 'Attract' (At). The follwing Allens temporal intervals are used by ignoring the inverses, {'Before' (<), 'meets' (m), 'overlaps' (o), 'starts' (s), 'during' (d), 'finishes' (f), 'equals' (=)} (Sridhar et al., 2011).

5.3 Relational Representation using Graphs

Qualitative spatio-temporal relationships can be used to distinguish human activities (Cohn et al., 2003). The changes in qualitative spatial relationships between two interacting objects can capture interesting state changes of objects which can represent an event. This project uses the concept of interactions to define these qualitative changes in spatial relationships as explained by Sridhar et al. (2011). Interactions form the essence of modelling events by capturing these qualitative state changes between the participant objects. For instance, for two objects 'Left leg' and 'ball' an interaction could be {DR, PO and DR} and this is characteristic of the event 'kick'.

After obtaining the frame-by-frame tracks of objects and event ground truth for each video, these interactions can be represented in the form of 'interaction graphs'. These interaction graphs portray qualitative spatio-temporal relationships between co-temporally occurring

object tracks(Sridhar et al., 2011). Interaction graphs are a relational description of interactions and provide a mechanism to measure similarities between different interaction graphs and are suitable for applying machine learning techniques (Sridhar, 2010)

An episode is defined by Sridhar et al. (2010) as an interval during which a spatial relationship maximally holds for a certain duration. An episode has the tracks of interacting objects defined by their bounding regions, the spatial relation defining the type of QSR corresponding to the object interaction and the time interval for which the relationship holds. The episodes characterising interactions are extracted from the object tracks data and is shown in

Figure 10.

Figure 10: A set of episodes obtained from a video representing 'kick'. The episodes are shown below with objects LeftLeg Right Leg and the ball. an episode sequence between Left Leg and ball is DR,PO,DR.

An interaction graph can be represented using a set of episodes if there is a sequence of changes in the spatial relationships within the bounding interval of the interaction. The

constituent episodes in an interaction graph are related to each other in the temporal order using Allen's temporal calculi.

An interaction graph is represented by a set of all the episodes between all pairs of objects. The episodes form the layer 1 and layer 2 nodes of an interaction graph where the third layer connects an episode pair by their temporal relationship. A video interaction graph representing all the episodes with pair-wise qualitative spatio-temporal relationships among the interacting objects is obtained. This process is repeated for all the videos.

An interaction graph has three layers. The layer one nodes are mapped to a set of object tracks, layer two nodes to qualitative spatial relations that hold for certain maximal intervals between co-temporally occurring object tracks, and layer three represents qualitative temporal relations between these pairs of intervals. Layer 1 and Layer 2 nodes characterize episodes and Layer 3 nodes connect the episodes using their time intervals described by Allen's temporal relations, see Figure 11.

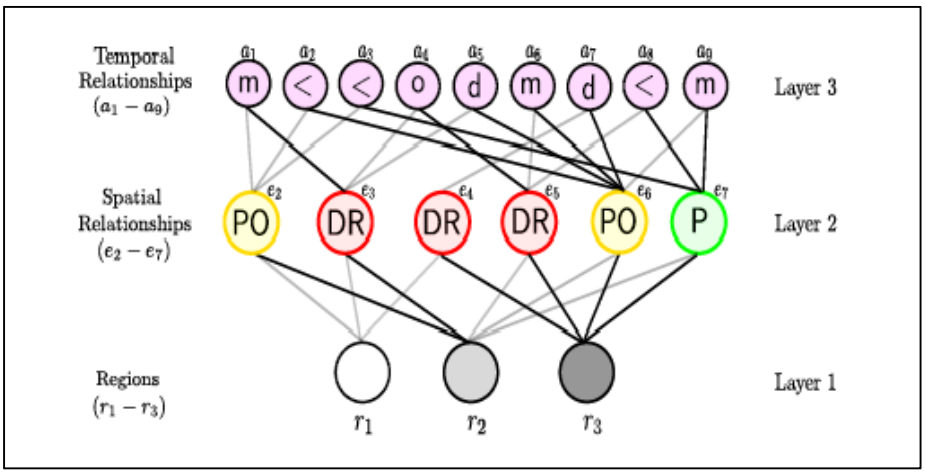


Figure 11: An interaction graph representing the 3 layers of spatio-temporal relationships between objects (Sridhar et al., 2010). Layer 1 represents object tracks corresponding to the object regions r1, r2, r3. Layer 2 represents qualitative spatial relationships between the object track pair. Two episodes are related by temporal relations in Layer 3.

One interesting addition to the designed framework was incorporating the concept of object types into the interaction graph. This is based on the idea that even though the verbs 'kick' and 'throw' would have similar changes in spatial relationships since the ball moves away from the body, there is a distinction in the types of objects that take part in the interaction. For instance, in 'kick' the ball moves away from the leg and while in 'throw' it

moves away from hand. To make this distinction evident in the interaction graphs, the types of objects (body part types) were given as a relationship, see Figure 12.This is done by saying that for the time interval between the start and end frames of a video that gives the entire video length, the type of pair-wise relationship between an object representing left hand, with itself is 'Left Hand'. This gives distinctive video interaction graphs for the two event classes 'kick' and 'throw'.

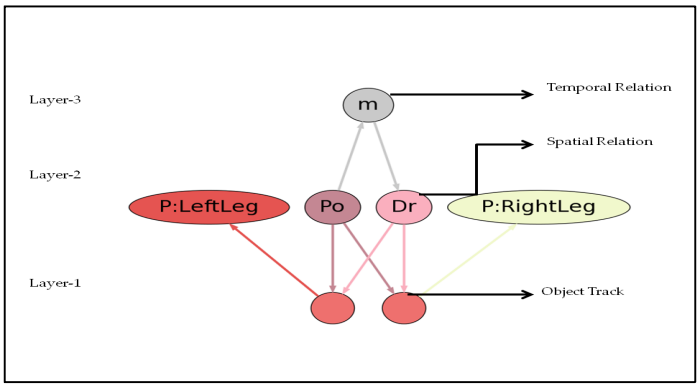


Figure 12: An interaction graph with object types included. It is shown that the object tracks are mapped to the corresponding type of the object.

A video interaction graph or the activity graph would collectively represent all interactions involving all the episodes between all pairs of observed objects during the entire length of a video. Hence a video-interaction graph is a representation of the qualitative spatio-temporal relationships between the body parts and other objects in video.

5.4. Conclusion

This chapter explains how the interactions are represented using qualitative spatio-temporal relationships. The first section gives an outline of different typess of qualitative spatial and temporal relationships used in the proposed approach and later sections explain how these relationships are described using interaction graphs. The concepts of episodes and how they can show characteristic patterns with respect to an event are shown here.

Chapter 6

Event Learning and Recognition

6.1. Introduction

The learning framework in this project uses a supervised setting where the system is provided with the knowledge of the type of the events for a domain. The verb/event labels are given to the learning algorithm to learn a model. Unsupervised methods have minimal knowledge about the type of events that are likely to occur in a video (Sridhar, 2010). The project uses standard supervised setting where the events are manually segmented and labelled.

This chapter discusses in detail how events are recognised using the relational learning approach proposed by Sridhar et al. (2010) in a supervised setting as mentioned above. There are two different phases for event recognition; feature representation and event learning. The following sections discuss each of these phases.

A video interaction graph obtained as explained in chapter 5 represents all the interactions for a given video. An interesting note is that events form only a subset of these interactions seen in the video interaction graph. For instance, the event 'kick' might occur between the objects Right Leg, Left Leg and the ball involving only the spatial relationships DR and PO during the frame interval, say, 231-280. Hence interaction-sub graphs should be extracted from the video interaction graph corresponding to these short time intervals which would be representative of these events, see Figure 13.

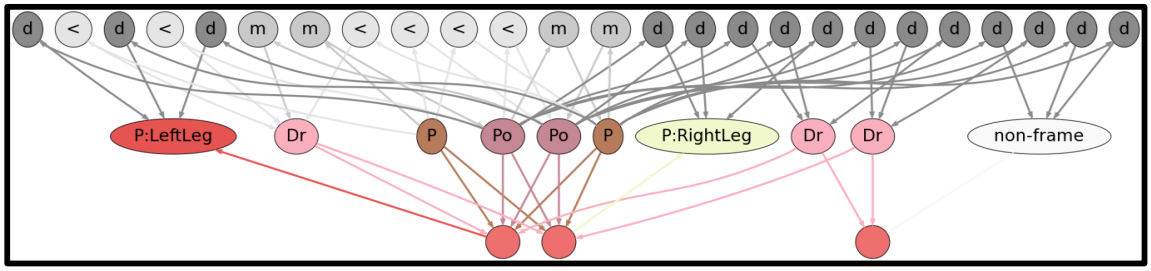


Figure 13: A video interaction graph obtained from implementing the proposed approach. This interaction graph corresponds to the event 'kick'.

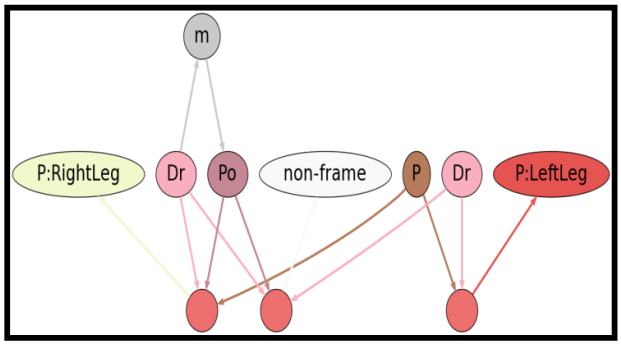
6.2. Feature Representation

A set of video interaction graphs are extracted using the procedure detailed in chapter 5. These video interaction graphs are represented as a set of interaction sub-graphs as mentioned in the above section. The feature representation phase defines a set of features that can characterise these video interaction sub-graphs. The features are a set of patterns or sub-graphs that occurs frequently in a set of video-interaction graphs representing a single event class.

It is seen before that interactions belonging to particular event have similar patterns. If two interactions are similar, the interaction sub-graphs corresponding to those interactions will be isomorphic (Sridhar et al., 2010). Sridhar et al. (2010) mine a set of which frequently occurring sub-graphs called 'graphlets', capturing these patterns from the interaction graphs representing the training samples.

A similarity measure is necessary to identify a set of frequently occurring sub-graphs. A frequent sub-graph is chosen based on a measure called 'support' which represents the numbers of times a sub-graph occurs in a given set of video interaction graphs. If the support measure exceeds a certain threshold the sub-graph is chosen as a frequent sub-graph or 'graphlet'. The 'graphlets' are mined using a greedy depth- first- search.

A 'graphlet' is hence a frequently occurring sub graph belonging to an event class mined across all the training videos representing that event. Since a graphlet identifies a pattern across several interaction graphs that belongs to the same event class, it has high dependencies with respect to a target event class (Sridhar et al., 2010).



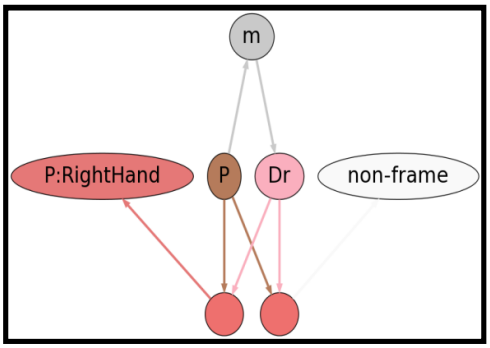


Figure 14: A 'graphlet' representing 'kick' event. Figure 15: (i) A 'graphlet' representing 'throw' event.

In order to represent each interaction sub-graph corresponding to a video as feature vector, the first step is to generate a graphlet vocabulary for the 'Bag-of-graphlets' representation. For this task, all the 'graphlets' for an event are mined as mentioned above. Once the graphlet vocabulary is defined, each sub-interaction graph belonging to a video is represented as a histogram of 'graphlets'. The histograms count the number of times a graphlet taken from a 'graphlet' vocabulary occurs in the interaction-sub graph for a video. Graph isomorphism is used to measure the occurrence of 'graphlets' in the interaction sub-graphs. The Histogram-of-graphlet encoding can be seen from the following figure.

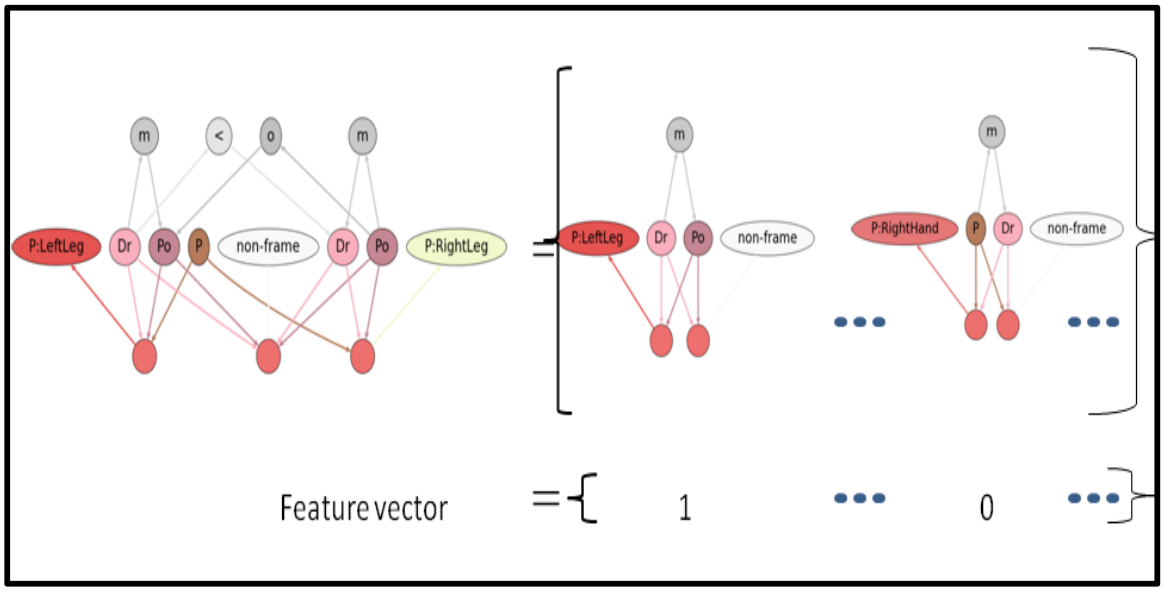


Figure 16: The bag-of-graphlets approach. An interaction sub-graph is encoded as a histogram of graphlets, which counts the number of times a graphlet occurs in the interaction graph

6.3. Event Learning

In this phase, a set of feature vectors obtained from the training videos are used for learning an event model which classifies and detects an event class in an unseen video. The event learning has two stages, training and testing.

6.3.1 Training

The ground truth corresponding to the events consist of smaller time intervals pointing out the start frame and end frame in the video and are labelled with the events or verb names. Hence interaction sub-graphs that correspond to the time interval given in the ground truth

are used for training the event model in a supervised setting. Also, many other interaction sub-graphs from the same training video might temporally overlap with the ground truth time interval. These interaction graphs are also considered to be representative of the events.

If an interaction sub-graph has a time interval corresponding to that of the ground-truth interval, then that sub-graph is labelled as the event class given in the ground-truth. Also if there are other sub-interaction graphs which overlap with the ground truth time interval by a certain threshold, then those are also labelled with corresponding event class. For instance, if the ground truth interval for 'kick' corresponds to frames 231-280 then the interaction sub-graphs with intervals, say, 231-280, 220-260 or 240-290 will be labelled as 'kick. Hence one video interaction graph might have more than one interaction sub-graphs that characterise an event.

Other interaction sub-graphs that have a large gap with the ground truth interval are considered as background class. All the sub-interaction graphs mined from the larger interaction graph for an entire video are mapped to the respective class as explained above. Now a video is represented by a set of interaction sub-graphs which correspond to the event class or background class.

These representative sub-interaction graphs are encoded as feature vectors using a histogram of 'graphlets' as explained in the above section. All the features used to encode the feature vector may not be discriminative, i.e., it may not uniquely define an event class characteristic. Hence a method known as Minimum Redundancy Maximum Relevancy (MRMR) is used to select a subset of features that has maximum dependencies with respect to a target event class (Sridhar et al., 2010).

6.3.2 Testing

The aim of this phase is to classify an unseen video which is assumed to contain the target event class into the respective class based on a learned event model. Also, it should detect the event in the video, i.e. predict when and where it occurs in the test video.

It was seen above that an event class, for e.g, 'kick' represents spatio-temporally similar (DR, PO, DR) ways of performing some task. It is appropriate to say that an event class is a probability distribution over a certain set of similar interactions or interaction graphs

describing the interactions. Here, standard problem in activity analysis is mapped to a formulation in graph based relational learning(Sridhar et al., 2011).

When a new test video is provided to the event model, the problem of event detection in the video can be translated into a problem of finding the most probable covering of the video interaction graph corresponding to the given event with the interaction sub-graphs mined from the new video using a learned event model.

Given a test video, the goal is to find the most probable covering of its video interaction graph with the interaction sub-graphs using a learned event model. To achieve this, the interaction sub graphs are mined first from the test video and represented as a feature vectors using the histogram of 'graphlets'. The 'graphlet' vocabulary defined from the training examples is used to encode the test video interaction sub-graphs. Then each interaction sub-graph represented as a feature vector is classified using the learned model and assigned a likelihood of representing the target event. The most probable video sub-interaction graph has the highest likelihood and is detected as an event. Each detected sub-interaction graph is now labelled with its predicted class and is mapped back into the corresponding video as event detections. If there are more than one detections for a given video, then a threshold of 50% is set for the likelihood so that only those sub-graphs which fall above the threshold will be regarded as event detections.

In the supervised setting used for the task of event learning in this project, two machine-learning algorithms were used to learn an event model to predict the event class for an unseen test video. K-nearest neighbours is explained in the following section.

6.4. Machine Learning Algorithms

In a supervised learning setting where each training example is labelled with its corresponding class. The learning system should now associate the test item to a given label which is learned from a trained model. The prediction accuracy can be evaluated against a ground truth which is already known. There are different machine learning algorithms to perform the task of classification and two of those methods are detailed here which is used for the purposes of the project.

6.4.1. K-Nearest-Neighbour

K-nearest-neighbour (KNN) is a instance based learning algorithm (Witten, Frank, & Hall, 2011) which assumes that features can be represented as points in n-dimensional space and it will use k nearest points to predict the class of the test point. The nearest neighbours are defined in terms of Euclidean distance. The Euclidean distance between the test item and every training item is calculated and k-nearest neighbours or k training examples with the shortest distance from the test data are selected. The majority class associated with the k neighbours is taken as the predicted class for the test data. This method is easy to implement and straight forward but the value of k should be chosen carefully so as to avoid under-fitting of over-fitting.

6.4.2. Cross Validation

When the amount of testing and training data is limited, cross validation method (Witten et al., 2011)can be used to divide the available data set into testing and training sets. This method reserves a certain amount of data for testing and the rest is used for training. For example, in N-Fold cross validation, with N = 10, the data is randomly divided into 10 parts in which each part is chosen as a test set in turn and the rest of the partition is used for training. The cross validation process will be repeated 10 times and the results are averaged.For the purpose of this project, five-fold cross validation is used. The input videos are partitioned into five parts referred as folds. Each fold is used as test data for detecting events while the other four folds are used for training with respect to the target event class.

6.5. Conclusion

This chapter explained the relational learning framework in detail including how interactions graphs were represented as features vectors and how event learning was implemented. First the concept that events occur as subsets in video interaction graphs is introduced. Then patterns characterizing an event were captured using graphlets. These graphlets were used for encoding video interaction graphs as feature vectors using a Bag-of graphlets approach. Then event learning is explained with training and testing phases mentioned in detail.

Chapter 7

EXPERIMENTS & EVALUATION

7.1. Introduction

The purpose of this chapter is to evaluate the 'proposed approach' outlined in chapter 3 using different sets of experiments. The following sections discuss the problem investigated by an experiment, how it was implemented, present the results and conclusions derived from it. The experiments provide a means to investigate the initially set research questions. The first section characterises the video-data set used for activity recognition. The following sections defines different experiments undertake at different stages of development of the proposed approach. As an evaluation strategy, a baseline method is also implemented to provide a comparison against the proposed approach.

7.2. Video-dataset Procurement

One of the main purposes of this project is to investigate the use of modelling interactions between the body parts of a person and other objects to recognise an activity. Hence in order to investigate the validity of the hypothesis presented by the 'proposed approach', a simple setting considering only two verbs was chosen. Also, a set of verbs where these interactions are evident were considered. Two verbs which qualify to capture this characteristic selected for the purposes of this project are (i) 'kick' where the interaction is between the leg of a person and an object and (ii) 'throw' where the interaction is between the hand of a person and an object.

Obtaining suitable video data for analysis of these event classes is crucial to perform the designed set of experiments. It is important to ensure that the videos clearly depict the interaction between the body parts and the object. A set of videos that represent 'kick', where a person kicks a ball or a vehicle and a set of videos that contain the verb 'throw', where a person throws a ball or a bag were taken to form the input video dataset in this project.

For the efficiency of computing tasks, the lengths of the videos selected were short, where most of the videos included only 500-600 frames. Also, to maintain the simplicity of the recognition task, single person scenarios involving an interaction between a single human and an object were preferred. The videos were taken from a single view point and had relatively less clutter or background noise.

Also, it was of prime concern to choose videos which represented the verb in a similar pattern which would in turn give similar interaction graphs for an event. For example, the video instances which portrayed the verb 'throw', as an object being thrown from a car and the person throwing the object was not shown, were discarded. Also videos that involved the interaction between two persons, like one person kicking another were discarded.

All the videos selected had evident distinctive interactions between the body parts and an object. Hence only a small subset of video instances could be pooled from the Mind's Eye video dataset. Nine videos that represent 'kick' and 11 videos represent 'throw' formed the final dataset.

The selected videos are decomposed into set of constituent frames in order to process the image data frame by frame. The already available manual tracks for the selected videos were obtained and ground truth was gathered for each video. The tracks encode the location of the bounding boxes for each object, i.e. for a person and other objects present in the video. It also contains the event ground truth stating when and where the events 'kick' or 'throw' occurs in the video.

Figure 17: Videos representing 'kick' and 'throw' events used in the proposed approach

7.2. Experiment 1: Body Part Detection

The purpose of this initial experiment was to see how body parts were detected over the image frames. The very first requirement of this approach was to identify how well the body parts could be abstracted using the chosen approach by Yang and Ramanan (2011).

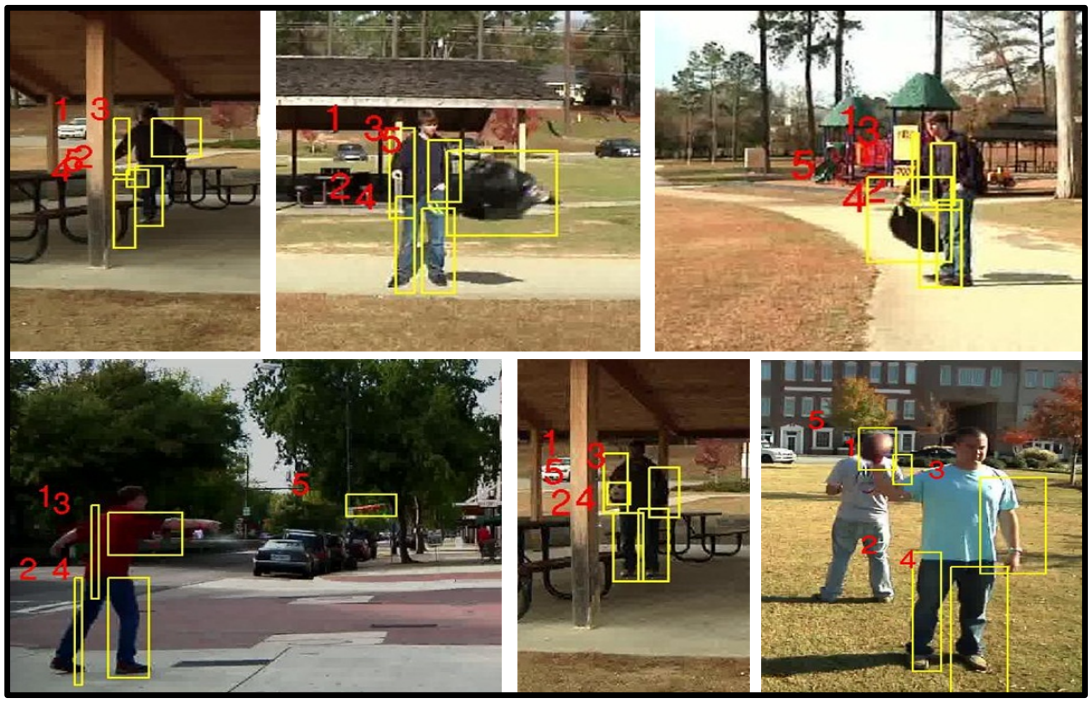


Figure 18: Body Part detections returned by the detection framework in different scenarios used in the proposed approach

It can be seen that the body parts were detected fairly well across all the videos. The average detection time taken to detect videos per image frame was 6-12 seconds. The next decision to be taken was on the types of body parts to be used for characterising the two target events, 'kick' and 'throw'. Initially, all the types of body parts were chosen as objects taking part in interactions. But, the pair-wise episodes became very large when such a choice was made. Hence only those parts that take part in the interactions were abstracted from the detection framework. In this project, where 'kick' and 'throw' are modelled, attention was given to the limbs of a person, refer Figure 19. All the other parts were not found to be relevant to model these verbs by observing the video instances, for instance, the body parts head or shoulders were not as important as the leg in the interaction 'kick'.

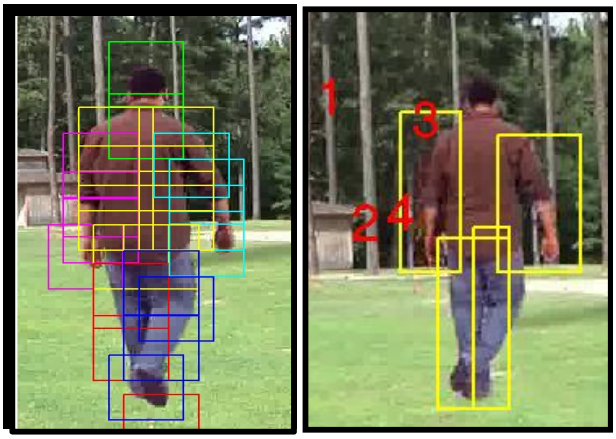


Figure 19: (i) Body parts originally returned by the detection frame work. (ii) Body parts abstracted away from the detected parts used in the proposed approach

7.3. Experiment 2: Identifying intra-class similarity in interactions

The purpose of this experiment is to investigate if there are significant patterns in the qualitative spatio-temporal relationships between the objects participating in a video. This is a crucial experiment that validates the basic hypothesis of this project, that using body parts can better describe some interactions. Hence these interactions should have a similar pattern across all the videos pertaining to an event.

One way to identify these patterns was to observe the episodes obtained from different videos. These episodes characterised a spatial relationships that existed maximally between pair of object tracks as explained in Chapter 5**.** An interaction had a sequence of episodes that followed similar patterns. This was investigated on the episodes to analyse the feasibility of applying relational learning to these interactions since the proposed recognition system is based on identifying and learning these observed patterns.

For the verb 'kick', the ball is first away from the body and then touches and departs from the leg of the person. Hence while using an RCC-5 spatial relation; the possible sequence would be (disconnected DR, partially-overlaps PO and DR) between the objects leg and the ball. This is depicted in the following Figure 20 and Figure 21.

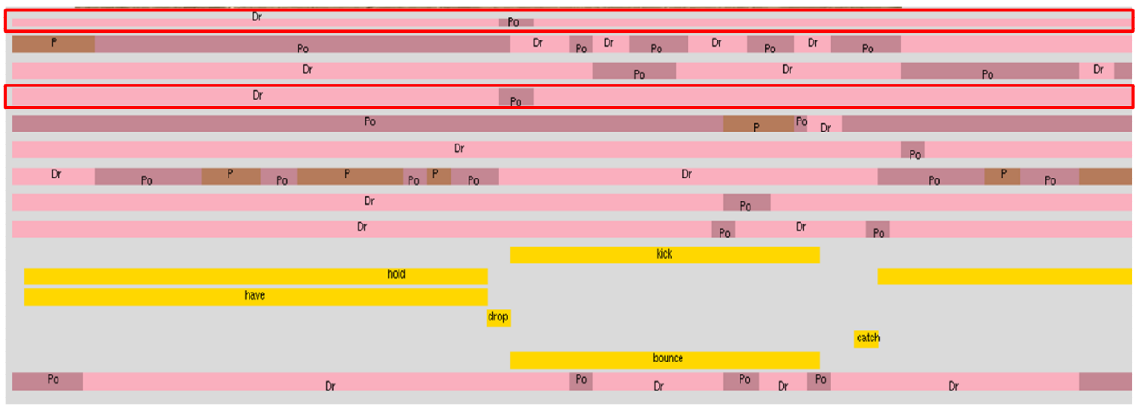


Figure 20: The episodes extracted from a video representing 'kick'. The highlighted episodes are between the legs and the ball. It can be seen that episode sequences in the vicinity of the ground truth interval is DR->PO->DR.

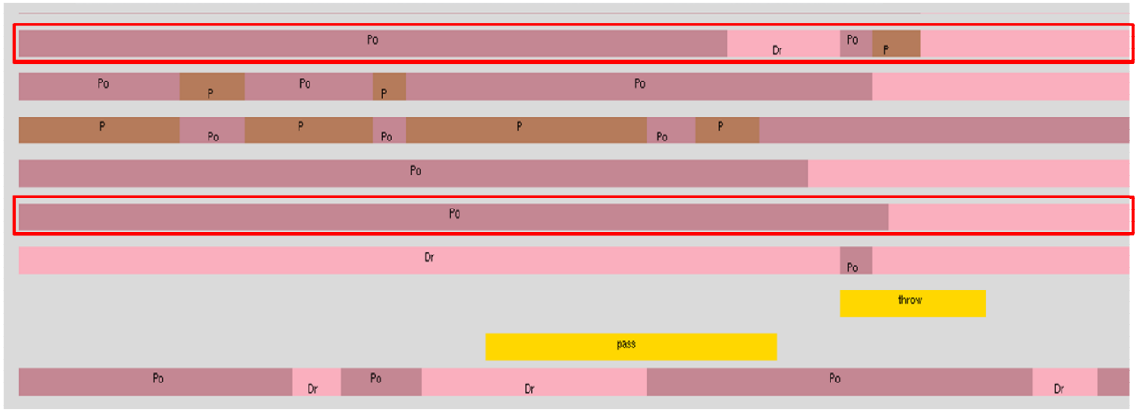


Figure 21: The episodes extracted from a video representing 'throw'. The highlighted episodes are between the hands and the ball. It can be seen that episode sequences in the vicinity of the ground truth interval are P->DR and PO->DR

It was seen that most of the videos representing the event class ‘kick’ showed this pattern.A similar pattern occurs in the verb ‘throw’ where the ball is first in touch with the hand and then departs. {PO, DR} is the RCC-5 relationship sequence between the hand and the ball. The following figure gives a sample episode in a throw video. This pattern was evident in most of the ‘throw’ videos, even though there were instances that did not capture this change well. Hence it was observed that many interactions were similar within an event class and followed a pattern.

7.4. Experiment 3: Identifying Discriminative Inter-Class Patterns In Interactions

There are two purposes for this experiment, one is to investigate if the interactions represented by interaction graphs had discriminative features with respect to an event class and the other is to investigate if the spatio-temporal patterns observed in the above experiment are captured by ‘graphlets’ representing a target class. This is an important concept since if there were distinctive features between interactions representing the events ‘kick’ and ‘throw’, the videos can be classified to the respective event class.

In order to find if there were distinctive interactions between ‘kick’ and ‘throw’, the ‘graphlets’ returned as an output by the activity recognition system were observed. There were two expected observations, (i) The patterns seen visually from the above experiment in the interaction episodes were also captured by a ‘graphlet’, (ii) There were ‘graphlets’ or features that represent ‘kick’ or ‘throw’ exclusively, i.e. a ‘graphlet’ feature present in ‘kick’ interaction graphs was absent in a ‘throw’ interaction graph.

The following Figure 22, is an observed result from the implemented system. It shows a video interaction graph belonging to the class ‘throw’ and two ‘graphlets’ that characterise a pattern occurring in most of the throw videos, {P,DR} or {PO,DR} as seen in the above experiment.

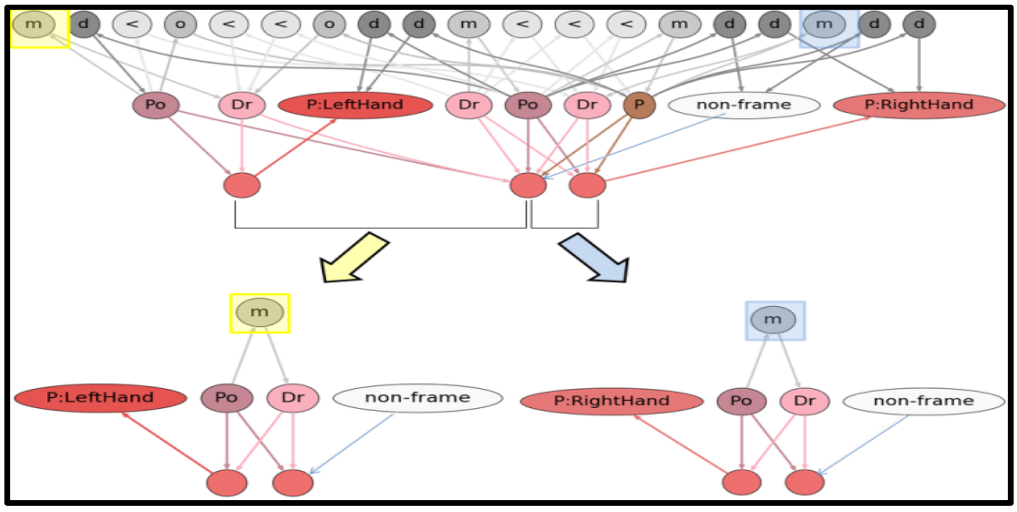


Figure 22: A video interaction sub-graph representing 'throw'. 'graphlets' mined from the videos corresponding to 'throw' are given below. The interaction graph contains the patterns that are characteristic to throw.

Similar observation can be seen in the video interaction sub-graphs representing ‘kick’. The following Figure 23 gives the result given by the implemented system where the graphlets capturing the interaction pattern for kick {DR, PO, DR} between the leg and the ball, as observed in Experiment 2, is shown

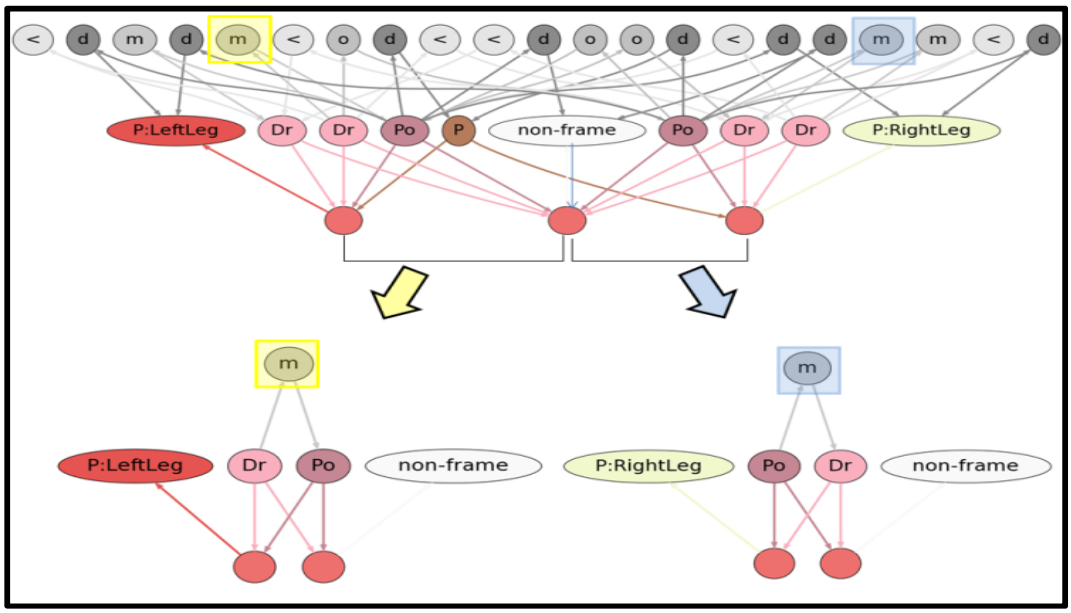


Figure 23: A video interaction sub-graph representing 'kick'. Two 'graphlets' mined from the videos corresponding to 'kick' are given below. The video interaction graphs contain the patterns that are characteristic to kick.

The next task is to find if there were distinctive patterns or ‘graphlets’ representing a target event class. The following subset of ‘graphlets’ are given in the figure which exclusively represented the classes ‘kick’ and ‘throw’. This was given as an output by the implemented system where only those ‘graphlets’ belonging to one class that were present in interaction graphs corresponding to that class but absent in the interaction graphs corresponding to the other class were obtained. See Figure 24, Figure 25.

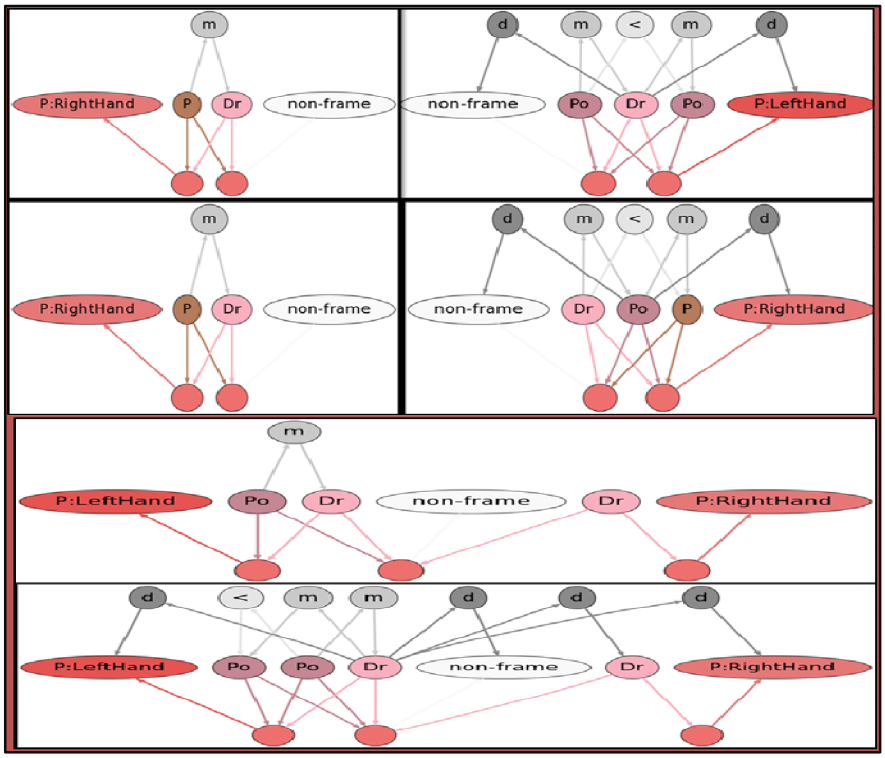


Figure 24: A subset of ' graphlets' capturing distinctive patterns between the leg and the object (non-frame). It can be seen that the pattern {DR, PO, DR } occurs in the graphlets.The 'graphlets' are obtained as a result of the implementation.

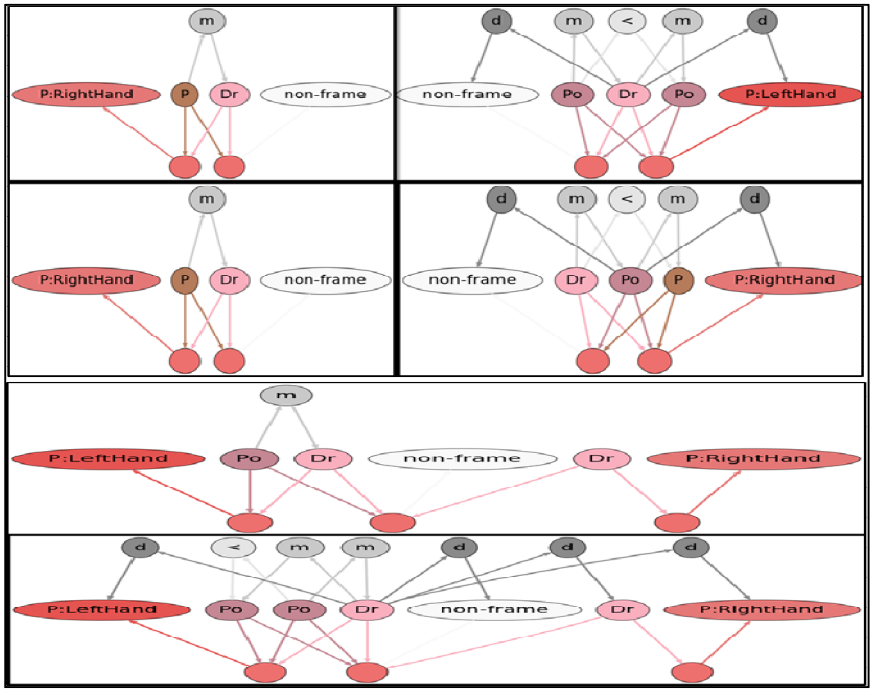


Figure 25: A subset of 'graphlets' corresponding to the verb 'throw are given above. It can be seen that the interactions are between hands and the 'non-frame' or the object. The main pattern is {PO, DR} the object first touches and then departs away from hands.

7.5. Experiment 4: Baseline Vs. Proposed Approach

7.5.1. Baseline

To evaluate the experimental set up used for this project, a baseline system was defined against which the current system performance was compared. The aim of the project is to investigate whether modelling body parts would be useful in recognising human activities and hence it is evaluated against a system that does not use body parts to model activities. Instead it considers humans as a single object interacting with an object.

In this scenario 'kick' and 'throw' are two interactions taking place between a person and an object. The body part detection framework is not implemented for the purposes of baseline system. For obtaining a baseline performance, the same set of videos selected in the designed framework was used. The manual tracks that gave the bounding boxes or the detection windows of persons were used as object tracks for the entire video length.

RCC-5 and QTC-6 spatial relationships were used and interaction graphs were obtained following the procedure discussed in Chapter 6. Video interaction graphs obtained were encoded as feature vectors using the 'Bag-of-graphlets' approach and given as input to a classifier. The results obtained from using the baseline method are discussed in the comparison.

7.5.2. Proposed approach

This experiment forms the core of this concept, which best investigates whether using body parts will improve the modelling of human interactions. For this, the steps are followed as explained in the above sections. For the experiment, 20 videos were used as input video data where 9 instances represented 'kick' and 11 instances represented 'throw'. Body parts were extracted first which contained detections of limbs of a person's body. The chosen qualitative spatial relations are RCC-5 and QTC-6.

The first set of experiments considers four different body part types each of which represent limbs of a person's body. They were labelled as 'Right Hand',' Right Leg', 'Left Hand and Left Leg. This is an apt choice since in the activities 'kick' and 'throw'; the main interaction is between a person's limbs and the object. The object types were included into the episodes as

explained in chapter 5**.** The resulting interaction graphs were encoded as feature vectors as explained above and classified using k-NN classifier.

7.5.3. Comparison

The results obtained from implementing the baseline and proposed approaches are discussed and compared below. The feature vectors characterising the events 'kick' and 'throw' were obtained for the baseline approach and the proposed approach. The figure below shows a plot of feature vectors corresponding to each video sub-graph against the graphlets or features used to encode them.

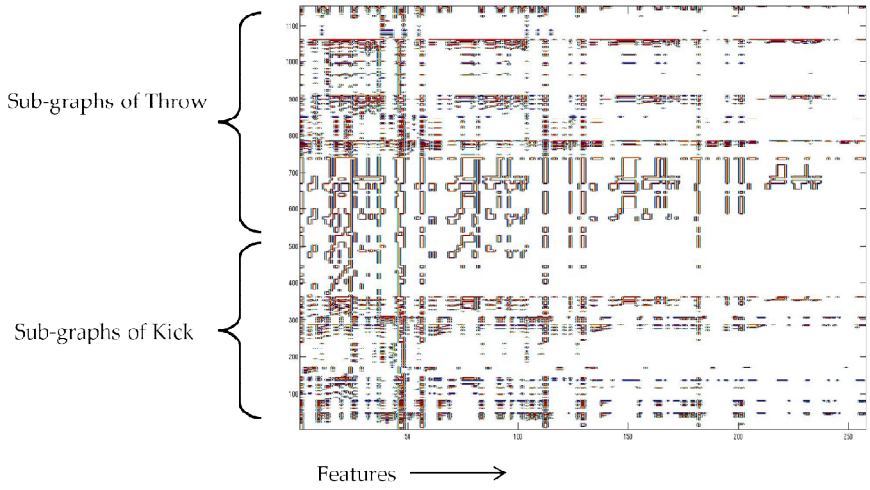


Figure 26: The feature vector from the baseline approach, the sub graphs encoded by features

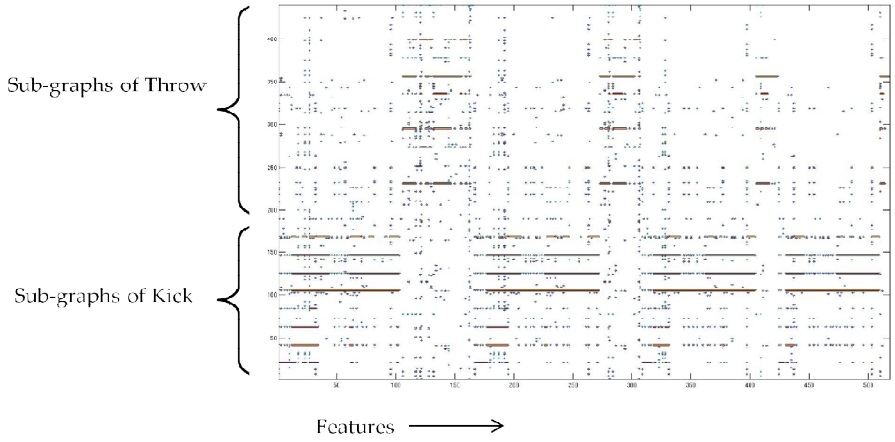


Figure 27: The feature vector from the proposed approach, the sub graphs encoded by features.

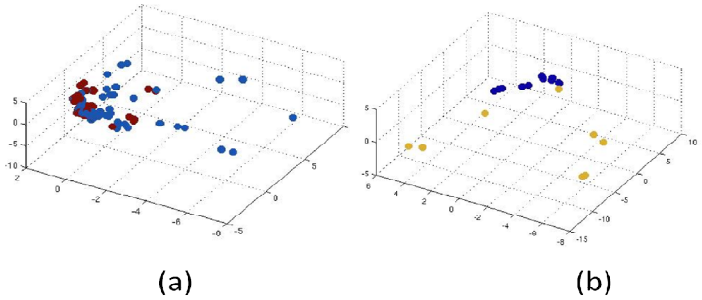


Figure 28: (a) features representation of two event classes in baseline approach (b) feature representation of two-event class in proposed approach.

The Figure 28 illustrates the feature vectors representing the two event classes and it is evident that there is a large overlap between the features from two classes. An inter-class distinction is not evident here. This leads to the conclusion that features used weren't highly discriminative. Ideally a feature must uniquely represent an event class.

The feature vector shows a distinction between the two event classes and hints that the classes can be uniquely characterised. By examining the feature vectors from Figure 27 and Figure 26 it becomes clear that the two classes show distinctive patterns.

7.5.3.1 Results

Results obtained for both the approaches are given as follows:

	F1 Scores	True Positives	False Positives	False Negatives	Precision	Recall
Baseline	0.57	10	13	10	0.4	1
Proposed Approach	0.79	13	0	7	1	0.65

Table 1: Comparison of the proposed approach and baseline using F1 scores along with an in-depth analysis of result

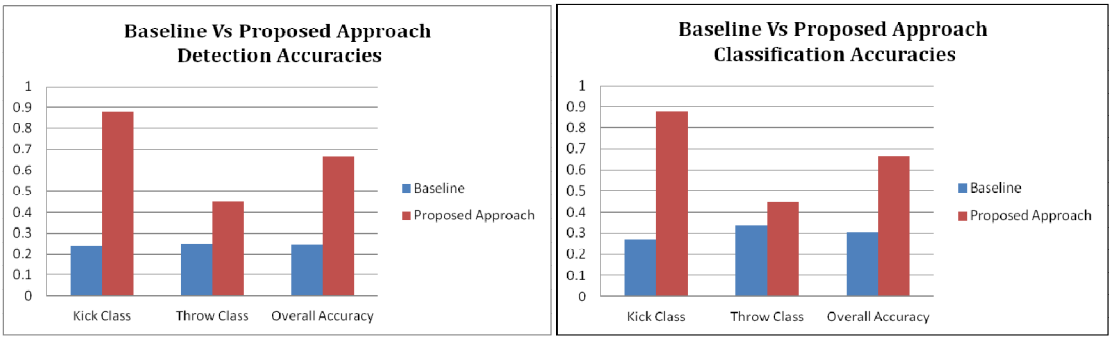


Figure 29: Results Obtained for Experiment 4: Baseline vs. Proposed Approach

7.5.4 Conclusion

From the above results it can be seen that the two event classes have been detected and classified with better accuracy using the proposed approach when compared to the baseline.

In the baseline approach, class 'throw' is shown to have been classified better compared to 'kick' event class. These results may be attributed to the observation that in this baseline approach the interaction is modelled between a person as a single object and an interacting object like a 'ball' or 'vehicle'. Here the distinction between interactions occurring in the two events, 'kick' and 'throw' would not be highly evident, as the ball moves away from the person in both cases. This is evident from the feature vectors as seen in the Figure 26.

The event 'kick' was classified with higher accuracy compared to 'throw' in the proposed approach using body parts. These results could be interpreted as a direct outcome of using body parts to model the interactions, 'kick' and 'throw'. The use of object types also aided the features to be distinctive as it encoded information of spatial relationship changes between the specific body parts involved in the interaction and the object. Patterns were shown in both QTC-6 and RCC-5 spatial relationships. The observations on feature vectors confirm this idea.

Table 1 gives a detailed analysis of the F1-scores obtained in the results shown in Table 1. It can be seen that the proposed approach has obtained a significantly higher F1 score compared to the baseline. It is because the proposed approach gives more true positives and does not give any false positives. This indicates that using body parts has a significant improvement in detecting the events 'kick' and 'throw' over the baseline method. This indicates that while using the proposed method with body parts, the classifier either identified the true class of the given video or it did not classify it as belonging to 'kick' or 'throw'. This can be interpreted as result of the unique patterns obtained using the body parts. The baseline lacked such discriminative features.

Another note is that the detection accuracy of the baseline approach was lesser than its classification accuracy. This indicated that the detections, of where the event occurred were not accurate with respect to the ground truth intervals, even though the predicted class was correct.

7.6. Experiment 5: Using different qualitative spatial relationships

Another set of experiments were formulated to find the types of spatial relationships that are useful for modelling interactions. In verbs like ‘kick’ and ‘throw’ RCC-5 and QTC-6 qualitative relations were chosen first.

Topological spatial relationships, RCC-5 gives the spatial relations between the bounding regions of interacting objects over the video image frames. This can represent spatial relationship pattern in ‘kick’ and ‘throw’ since the ball touches the body parts and then stays disconnected. QTC-6 gives relative trajectories between the interacting objects which can represent if an object moves away from another. This is also a good choice for ‘kick’ and ‘throw’ since the ball moves away from the respective body part taking part in the interaction.

The experiments were run on video dataset with 21 representing ‘kick’ and ‘throw’ events. The first set of experiments used only QTC-6 relations. The second set of experiments used only RCC-5 relations as mentioned in chapter **5**. The observations and results from these two experiments were compared with the proposed approach which uses both RCC-5 and QTC-6.

7.6.1. Comparison

The feature vectors for the above experiments are given as follows:

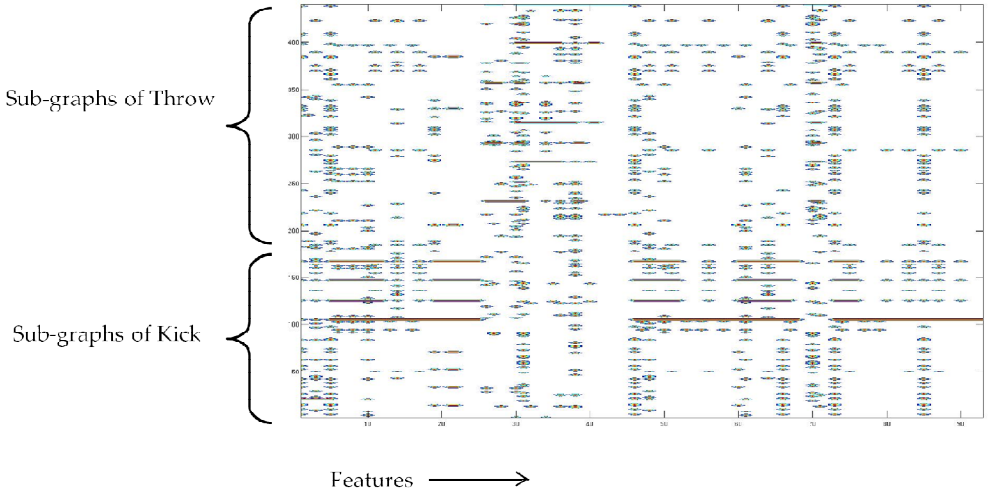


Figure 30: The feature vector from the using Only QTC-6

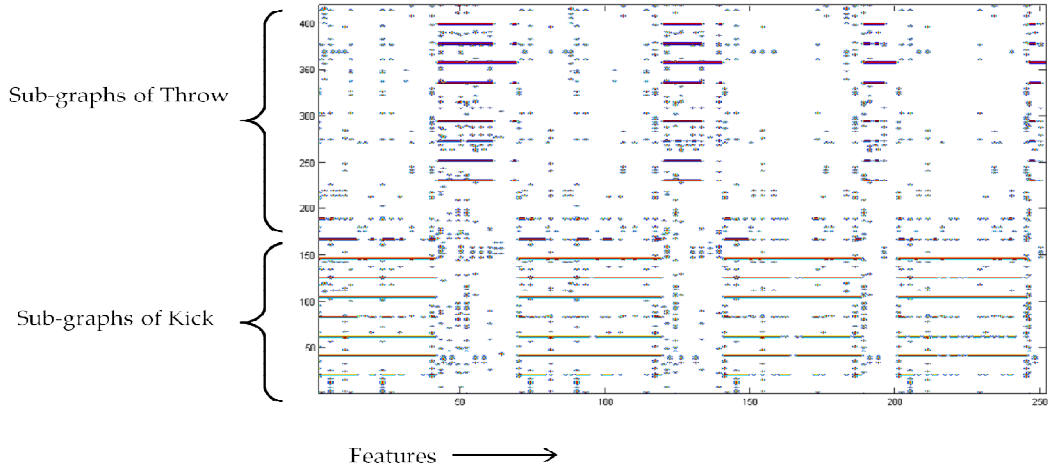


Figure 31: The feature vector from the experiment using Only RCC-5

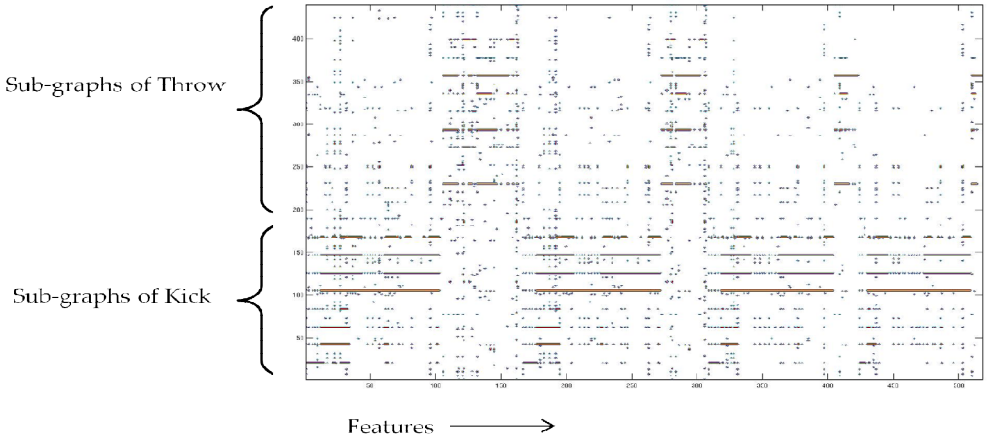


Figure 32: The feature vector from the experiment using RCC-5 and QTC-6

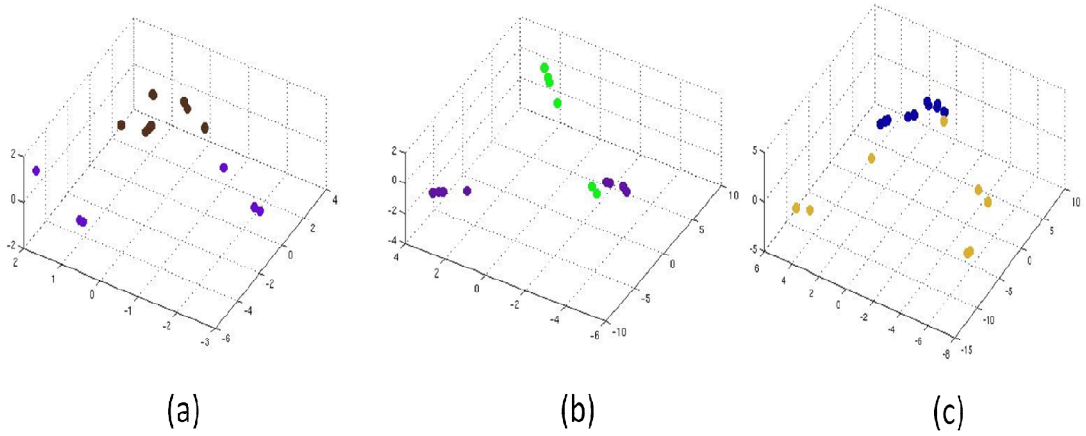


Figure 33: Feature representation for the two event classes,'kick'&'throw'(a) Using QTC-6 (b) Using RCC-5 (c) Using RCC-5 and QTC-6.

From the above figures obtained it can be seen that implementation using RCC-5 spatial relation gives the most distinctive features among the three approaches. When the feature vector plot in Figure 33 and Figure 32 is analysed, an evident distinction of patterns between the two classes can be visualised, with reduced background noise, or insignificant patterns. Implementations using QTC-6 alone shows patterns and incorporates a lot of noise, or patterns that do not distinguish the two classes. The proposed approach used shows a feature vector representation that includes a distinctive pattern, but also characterises noise.

The feature representation for the two event classes are given in the Figure 33. Here it is evident that the observation using RCC-5 alone gives a clear distinction of classes compared to using QTC-6 alone or RCC-5, QTC-6 combination. The features obtained using the other two spatial relationship types are also discriminative.

7.6.1.1. Results

The Detection and classification accuracies obtained when implementing the system using QTC-6, RCC-5 and the proposed approach that uses both QTC-6 and RCC-5 are given below.

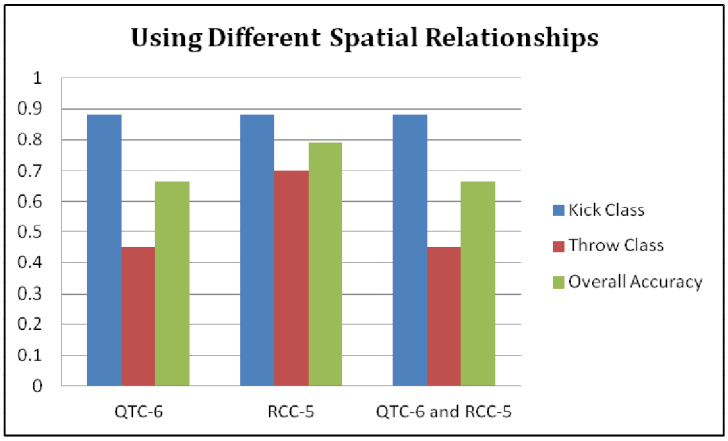


Figure 34: Results obtained for Experiment 5: Using different qualitative spatial relationships

7.6.2. Conclusion

It can be seen that the prediction accuracies are higher when RCC-5 spatial relationship is used when compared to using QTC-6 and the proposed approach that models QTC-6 and

RCC-5. The results obtained when modelling QTC-6 alone was similar to the results obtained by the proposed approach.

It can be interpreted from the features as observed in the above section that RCC-5 has better discriminative patterns when compared to the other two approaches. This follows from the idea that 'kick' and 'throw' can be characterized by the topological changes between the body parts involved and the object. Distinctive sets of 'graphlets' were obtained across the videos that represent an event which was also verified in Experiment 2.

QTC-6 can also be used for representing spatial relationship changes between the interacting object for the given verbs 'kick' and 'throw'. The two events have some distinctive changes in trajectories since the ball moves away from the leg or hand of a person. On the other hand, there are also a lot of observed noises or patterns that are not discriminative. One reason would be that in 'kick', when ball moves away from the leg of a person, it also moves away from the hand of the person. Hence relative trajectories might have similar patterns across the two events and result in non-discriminative patterns.

7.7. Experiment 6: Using finer body part representation

This experiment investigates if modelling finer granularities of body parts would improve performance accuracies or represent the events better. One idea was to model only the extremities of limbs. This was inspired from the idea that for a verb like 'kick', the ball would touch the 'foot' of a person and it would be worthwhile to investigate how the system would perform when only the extremities of body parts are modelled. This is then compared against the proposed approach which uses the entire region of limbs.

7.7.1. Comparison

The following results were obtained for this experiment:

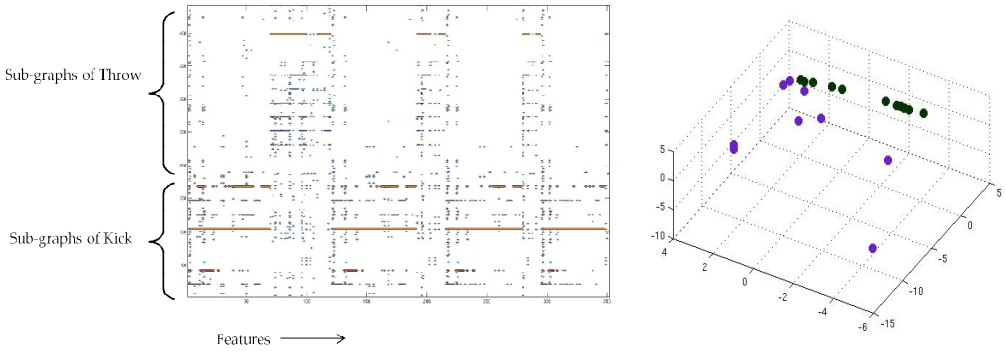


Figure 35: The feature vector plots for modelling body part extremities

When the feature vector graph is inspected, it can be seen that for the event 'kick' , patterns are not very clear, i.e., discriminative patterns are not evident. In the case of 'throw' similar patterns can be seen across all interaction sub-graphs.

7.7.1.1. Results

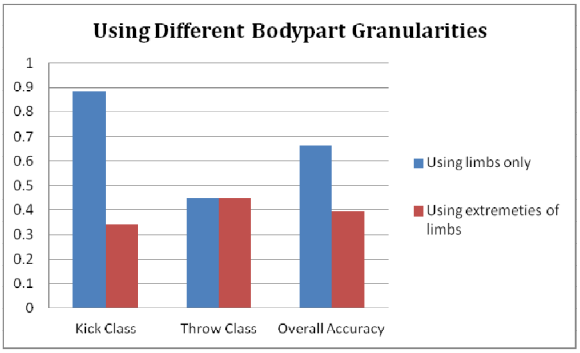


Figure 36: Results obtained for Experiment 6: Using finer body part representation

7.7.2. Conclusions

The results show that prediction accuracies were reduced for the event 'kick' and remained the same for the event 'throw'. This can be due to the observation that while modelling the extremities for 'kick', many of the kick videos had instances of the ball touching the leg of the person and not necessarily the 'foot'. In some videos, the interaction was between the knee part of the leg and the ball. In 'throw', the ball always touches the hand extremities since a person holds the ball first and then throws it. Also, the detections of the extremes

were not accurate in some videos. It is seen from the feature vector plot in Figure 35 that many of the kick videos do not display patterns, or characterising features are absent.

7.8. Experiment 7: Using object types vs. no object types

One experiment was designed to evaluate the system performance when the 'object types' information, explained in the **former sections**, was not provided. The object types give the interaction graphs an additional layer-1 relationship which relates the spatial relationships to its corresponding body parts and objects. The experiment uses the same implementation as the proposed approach, but without encoding object types.

7.8.1. Comparisons

The results obtained are as follows:

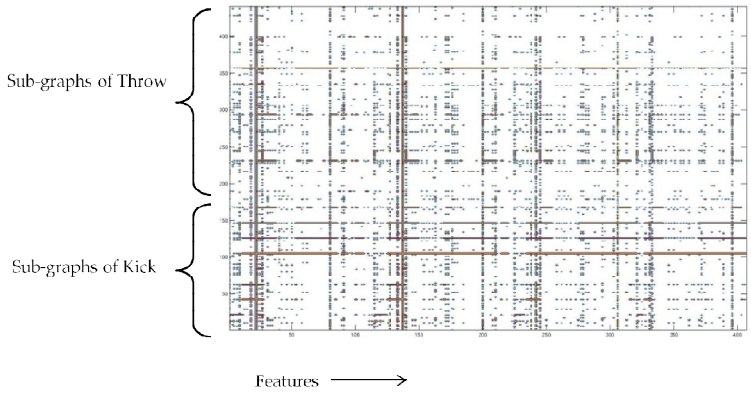


Figure 37: Feature representations when object types are not used

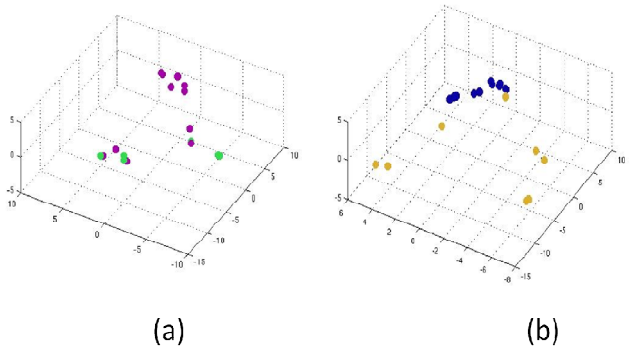


Figure 38: features (a) No object types (B) proposed Approach

The feature vectors in fig x does not encode any distinctive classes as the feature points appear scattered over the feature space. The feature vector graph shows no interesting patterns and contains noise.

7.8.1.1 Results

	F1 Scores	True Positives	False Postives	False negatives	Precision	Recall
Proposed Approach	0.79	13	0	1	1	0.65
Without Object Types	0.19	2	4	14	0.33	0.13

Table 2: Comparison of approaches using object types and no object types

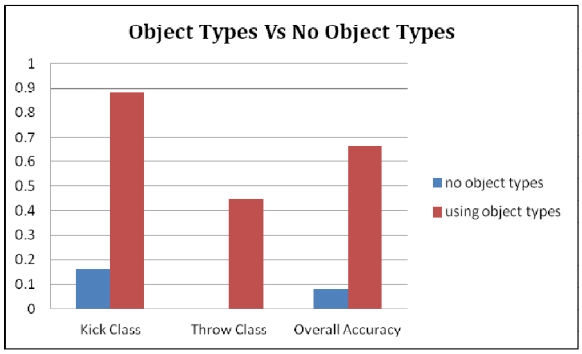


Figure 39: Results obtained for Experiment 7.8:Using object types vs. no object types

7.8.2. Conclusions

The prediction accuracies obtained are very low when compared to the proposed approach. The detection accuracy for the class 'kick' has been reduced drastically while 'throw' is not detected at all. In the Table 2, the F1 scores of the two approaches are given and it can be identified that the number of instances detected by the system that uses body parts without including object type information is very low.

These observations can occur due to the fact that pair-wise relationships between all the body parts and the object have been modelled here, without the knowledge of which objects are interacting. For example, in a video instance representing kick, the legs might move with respect to the hands and these interactions will also be captured. This interaction may occur in a video representing throw.

Also, there are some instances where throw and kick has similar interactions. If the information on what objects are interacting is abstracted away, it will not be possible to characterize discriminative interactions for an event. This is described in the following figure

where the interaction without using object types is {PO, DR}. Two other interactions that characterize a ‘kick’ and ‘throw’ are given with the same spatial relationship changes, {PO, DR}. From this the importance of using object types to distinguish events can be interpreted.

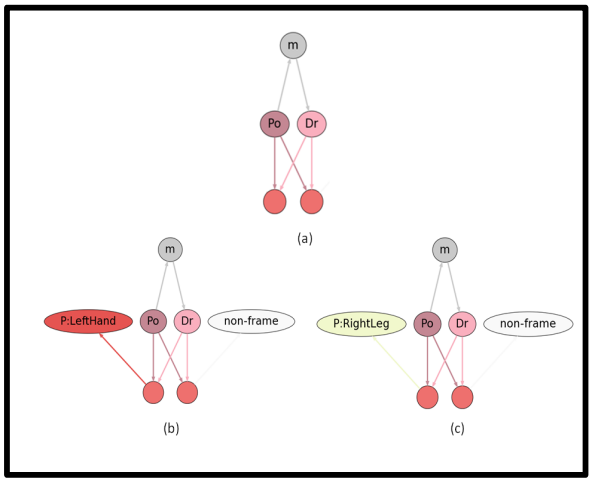


Figure 40: (a) an interaction without an object type (b) Interaction with object type representing 'throw' (ii) An interaction with object type representing 'kick'.

Chapter 8

CONCLUSIONS & FURTHER WORK

8.1. Introduction

This chapter presents the conclusions drawn from the Experiments section in chapter 7 and evaluates if the research questions set at the start of the project has been answered. Also this chapter discusses the problem scenarios the proposed approach might encounter and suggests further work that could be done to improve the proposed approach.

8.2. Findings

This section provides a list of conclusions drawn from the experiments outlined in the previous chapter.

'Experiment 4: Baseline vs. Proposed Approach' showed that the proposed framework which uses body parts to model interactions exceeded the overall performance accuracy of the baseline approach which modelled humans as a single object. This implied that modelling body parts for the events 'kick' and 'throw' gave a better representation of interactions.

'Experiment 5: Using different qualitative spatial relationships' identified which types of qualitative spatial relationships were best suited to describe the events 'kick' and 'throw'. Using 'RCC-5' alone showed the best performance giving an overall accuracy of 79% when compared to QTC-6 which gave 66.5% and'RCC-5 & QTC-6' with 66.5% overall accuracy. Hence 'RCC-5' was shown to best describe the verbs 'kick' and 'throw'.

'Experiment 6: Using Finer Body Part Representation' sought if a finer body parts representation would model the events more precisely. This was evaluated against the proposed approach. It was seen that using the extremities of body parts performed similarly for the verb 'throw' compared to the proposed approach while the verb 'kick' was recognised with reduced accuracy. This was attributed to the video instances describing 'kick' where a 'kick' was not necessarily between the 'foot' and the ball.

It was also investigated if including the 'object types' information into the episodes representing interactions helped to distinguish between the verbs 'kick' and 'throw'. 'Object types' map the object tracks to their corresponding type, for instance, left hand. It was found that not including the type information degraded the performance drastically, with only 19% overall accuracy compared to the proposed approach which includes this information. Hence type information was shown to be an integral addition to interaction graphs for a distinctive representation of interactions.

In the proposed approach, the pair-wise spatial relations between the limbs of the body and an object was modelled. This also includes the spatio-temporal relations between the limbs. These interactions did not have discriminative patterns when the given set of video instances is considered. These could model the noises in the features since these intra-body part interactions may not encode any significant changes representing an interaction between the object and the body part. Hence removing these relations might eliminate noises from the feature set. It can hence be derived that the feasibility of using extremeties depends on the video instances used for training. Also many design decisions are based on the verbs or events that should be recognised.

The proposed approach outlined in this project recognises two interaction verbs or events 'kick' and 'throw'. It should be noted that this approach is feasible for modelling certain verbs or events where there is an evident interaction between the body parts and the objects involved. Some verbs that show this characteristic are 'catch', where the hands and the object is involved, 'pick-up', 'put-down','lift' are some other verbs which would benefit from using this approach. Due to the time limit set for this project, the experiments could not be extended to the modelling of above verbs.

Another set of events that might benefit from using body parts rather than a single person are actions, where there are no external objects involved. For instance, the verb 'fall' there are spatial relationship changes between the body parts and here more body part types could be included and not just the limbs as used in the proposed approach. Some other action verbs that might be modelled using this approach are 'walk', 'run', 'clapping hands', 'a dance move'(e.g. a particular ballet move),'raise', which significant changes that follows a pattern with respect to other body parts.

The proposed approach outperformed the baseline method for the interactions ‘kick’ and ‘throw’. It would be worthwhile investigating which verbs would perform better using the baseline, by considering human as a single object. Verbs like ‘follow’, ‘chase’, ’approach’, where persons can be modelled as a single entity, for instance when a person chases another person, relationships between the bounding boxes of the entire persons are only required and modelling body parts would not be necessary.

Another consideration is to investigate the feasibility of using an approach other than graph-based representations to represent interactions. In a simple scenario where only a single event is recognised, Inductive Logic Programming methods can be used to describe the interactions. This could be implemented and compared with the proposed approach.

8.3. Problem scenarios

The body part detection framework as described in Chapter 4 is automatic, i.e. no supervision is provided for the system to detect the parts. Hence a ground truth is necessary to verify the detection accuracies. Since the ground truth was not already available it was not possible to annotate all the video frames with body part bounding boxes due to the time limit set by the project.

It would give an insight to the question of whether the body part detection accuracy has a direct effect on the overall classification accuracy of the target event. It is intuitive that if for a given set of frames an event occurs and the detection framework fails to capture a key interaction that only took place in one single frame (like a single kick), the framework would fail to record this change. This would have an impact on the classification accuracy.

It is also important to see how the approach would work when two verbs that have similar interactions with respect to body parts have to be recognised. For instance, the verbs ‘throw’ and ‘hit’, or ‘throw’ and ‘drop’ might have similar interactions when the spatial relationships are considered and the body parts involved are mostly the same. Another set would be ‘bounce’ and ‘jump’ where the person moves in a similar way.

It can be noticed that the quality of the findings are largely dependent on the quality of dataset on which the experiments were carried out. The videos representing event classes were chosen such that there was a clear and distinctive sequence of spatio-temporal

relationship changes between objects. The hypothesis introduced in this project was observed to be correct by modelling these events. But, it has to be investigated if the system will be robust against different other scenarios where the same event occurs. Due to the limited time available, the project could not be extended to explore this facet.

The proposed approach recognises two event classes. This can be extended to recognise more event classes, like 'catch', 'bounce', 'drop' etc and find the overall classification performance. It may be follow that including more classes with different set of patterns will help to differentiate the classes because of inter-class distinction in the interactions. Otherwise, if the verbs chosen have inter-class similarities then the classification may not be proper.

8.4. Answering Research Questions

This project aimed to find answers to the following research questions.

Will using spatial relationships between body parts be able to distinguish SOME verbs better than when considering a person as a single entity? The proposed approach used body parts to represent interactions in two event classes 'kick' and 'throw'. It was found that using body parts to model interactions can represent these verbs better when compared to using humans as a single entity.

Can the framework capture significant patterns of spatial relationship sequences between body parts and the interacting object?

For the two recognised event classes 'kick' and 'throw' significant patterns of spatial relationship sequences were observed and encoded into the features as found in the Experiment 2 in chapter 7**.** These distinctive patterns were observed as sequences of episodes and verified that 'graphlets' capture those patterns.

How would using different granularities of body parts help in recognising human activities?

For the two events recognised by the proposed approach modelling granularity using the extremities of the participating body parts did not improve the performance. While one verb gave the same performance as the proposed approach another gave a relatively low performance and this is attributed to the observation that the video instances did not

necessarily represent the interaction as taking part between the extremities of the body parts and the object. Another level of granularity, where is limb was represented by a set of four parts, could not be implemented due to the short time available for the project implementation. This could be investigated further.

What type of qualitative spatial relationships can model the events better?

It is seen that for the events 'kick' and 'throw', RCC-5 can model the interactions in the most distinctive manner compared to the relationships QTC-6 and a combination of both. But, this finding is limited to the target events detected here, 'kick' and 'throw'. When modelling other verbs this observation would not remain the same. This could be investigated as a further work.

Will the feature representation using body parts be distinctive with respect to an event class?

It was seen that the inter-class distinction was evident using the proposed approach in modelling the events 'kick' and 'throw'. This was verified in the experiment x where the feature vectors corresponding to 'kick' and 'throw', obtained using the proposed approach showed patterns that were similar within a class and different between classes. Hence feature representation using body parts were found to be distinctive with respect to the event classes 'kick' and 'throw'.

8.5. Further Work

One way to extend the approach proposed by this project is to make the system autonomous with minimum manual supervision. Instead of using manual tracks a set of automated tracks can be used detect persons in the image frames. The result from this can be given to the body part detection framework used in the proposed approach. Instead of working on a supervised setting, the approach can be implemented in an unsupervised setting where the event labels are not provided for training. In an unsupervised learning setting the event classes are not known in advance and allow the system to learn the classes autonomously.

Another investigation would be to find whether the body part detection accuracy would impact the classification accuracy. For this another detection method which performs better than the current approach, like the work by Park and Ramanan (2011) which is an extension

to the detection framework used in the project giving higher detection accuracies, should be taken and the detection accuracies and overall classification accuracies should be compared.

A direction for future work would be to investigate the applicability of the proposed approach in modelling composite interactions between multiple persons. For example, the verb ‘hand-shake’, ‘exchange’ where two persons interact and their body parts are involved, using the proposed approach might benefit. But, this would be computationally expensive since the body parts have to be detected for two or more persons taking part in the interaction. Also, only relevant body parts with respect to the target event should be modelled.

References

Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis. *ACM Computing Surveys*, *43*(3), 1–43.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. (R. J. Brachman & H. J. Levesque, (Eds.)*Communications of the ACM*, *26*(11), 832–843.

Campbell, L. W., & Bobick, A. F. (1995). Recognition of human body motion using phase space constraints. *Proceedings of IEEE International Conference on Computer Vision*, *pages*(309), 624–630.

Cohn, A. G., Magee, D., Galata, A., Hogg, D., & Hazarika, S. (2003). Towards an Architecture for Cognitive Vision using Qualitative Spatio-Temporal Representations and Abduction. In C. Freksa, C. Habel, & K. F. Wender (Eds.), *Spatial Cognition III* (pp. 232–248). Springer.

Cohn, A. G., & Renz, J. (2007). Qualitative Spatial Reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter (Eds.), *Handbook of Knowledge Representation*. Elsevier.

Damen, D., & Hogg, D. (2009). Recognizing linked events: Searching the space of feasible explanations. *IEEE Conference on Computer Vision and Pattern Recognition (2009)*, 927–934.

Felzenszwalb, P., Mcallester, D., Ramanan, D., & Irvine, U. C. (n.d.). A Discriminatively Trained , Multiscale , Deformable Part Model.

Galata, A., Cohn, A. G., D., M., & Hogg, D. (2002). Modeling Interaction Using Learnt Qualitative Spatio-Temporal Relations and Variable Length Markov Models. In F. van Harmelen (Ed.), *Proc. European Conference on Artificial Intelligence (ECAI'02)* (pp. 741–745).

Ivanov, Y. A., & Bobick, A. F. (2000) Recognition of visual activities and interactions by stochastic parsing. , 22 IEEE Transactions on Pattern Analysis and Machine Intelligence 852–872 (2000). IEEE. doi:10.1109/34.868686

Joo, S., Chellappa, R., & Park, C. (2006). Attribute Grammar-Based Event Recognition and Anomaly Detection ◎.

Krishna S. R. Dubba, Cohn, A. G., & Hogg, D. C. (2010). Event Model Learning from Complex Videos using ILP. *Proc. ECAI* (Vol. 215, pp. 93–98). IOS Press.

Laptev, I., & Lindeberg, T. Space-time interest points (2003), Proceedings Ninth IEEE International Conference on Computer Vision 432–439 vol.1 (2003). IEEE. Lavee, G., Rivlin, E., & Rudzsky, M. (2009). Understanding Video Events: A Survey of Methods

for Automatic Interpretation of Semantic Occurrences in Video. *IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews*, *39*(5), 489–504.

Natarajan, P., & Nevatia, R. (2007). Coupled Hidden Semi Markov Models for Activity Recognition ◎ CHSMM◎: Definition , Learning and Decoding Algorithms.

Oliver, N. M., Rosario, B., & Pentland, A. P. (2000). A Bayesian computer vision system for modeling human interactions . *Pattern Analysis and Machine Intelligence, IEEE Transactions on*.

Ryoo, M. S., & Aggarwal, J. K. (2006). Recognition of Composite Human Activities through Context-Free Grammar Based Representation. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*.

Sridhar, M. (2010). *Unsupervised Learning of Event and Object Classes from Video*, University of Leeds

Sridhar, M., Cohn, A. G., & Hogg, D. C. (2010). Unsupervised Learning of Event Classes from Video. *Proc. AAAI* (pp. 1631–1638). AAAI Press.

Sridhar, M., Cohn, A. G., & Hogg, D. C. (2011). Benchmarking Qualitative Spatial Calculi for Video Activity Analysis. *Proc. IJCAI Workshop Benchmarks and Applications of Spatial Reasoning* (pp. 15–20).

Veeraraghavan, A., Chellappa, R., & Roy-Chowdhury, A. K. (2006). The Function Space of an Activity. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. *Annals of Physics* (Vol. 54, p. 664). Morgan Kaufmann. Retrieved from http://www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0123748569

Yang, Y., & Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts resenting shape. *CVPR*.

P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan Sept 2010. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Pattern Analysis and Machine Intelligence* (PAMI).

D. Park, D. Ramanan. November 2011. N-Best Maximal Decoders for Part Models International Conference on Computer Vision (ICCV) Barcelona, Spain

Appendix A

Personal Reflection

Equipped with his five senses, man explores the universe around him and calls the adventure science – Edwin Powell Hubble.

These words remind me of the wonderful experience I had by indulging myself in a scientific research environment presented by this project. Looking back at the paths I have taken through the course of this project work, I feel extremely delighted to be where I stand now. Working on this project was perhaps more than a part of my academic course; I could do something that I always wanted to, and that was to experience scientific research. I was fortunate to work on a project which amalgamated my areas of interest, computer vision and machine learning. I strongly believe that it necessary to choose a project that one really is interested in, because it is that fervor which keeps you moving, throughout the project.

While doing this project I observed that simple thoughts are developed into ideas and investigated further. It is extraordinary to see that those intuitive thoughts later transform into working systems. This project gave me a platform to think on my own and implement what I feel is right about the concept.

Also, I felt that it is important to think about all the possible directions for a research question and make apt decisions on the routes to travel. Regular supervisor meetings kept me guided and a good communication with my supervisor Prof. Tony Cohn helped me a lot,. Many things were not straight forward at the start, but slowly everything became clearer. I found myself in a ‘fix’ many times, especially when some of the results I got were unexpected. It was my assessor who gave me the insight that process is more important than the result. I was expected to have a thorough understanding on what I was doing and why I was doing it. These are some of the lessons I learned from my project experience.

I faced problems when I had to work with external software and understand the details of its implementation. I underestimated the time it would take for that task and it consumed most of my time, during the later stages of the project work. Also, the 'literature review' section was an entirely new concept, where I had to read and critique on others' work. I spent a lot of time working on that section, since it was crucial to identify what 'not' to include. Therefore I would like to advise future students to start working on the literature review as early as possible.

Writing up the project report was a relishing experience as many concepts I thought I did not know about my work materialised when I started penning it down. The project write up is one of the core parts of a project work as it is a 'window' of the entire project work to the rest of world. Hence the write up demands high attention.

One of the things I would like to advise the future students would be to 'focus on the process- not the result'. If this is followed, I can ensure that the project work would bring an amazing experience to you.

Appendix B

Software Used

For the purpose of body part detection the code available from the authors Yang and Ramanan were used. Any functions that were modified are given below along with the amount of modification made.

demo.m – deva raman's code for detecting body parts from a given image. Slight modification.

The following are the set of functions that were written and implemented in Matlab for the purposes of the project.

BodyPartsLearning.m – main function for applying the framework pipeline.

Showboxes_ellipses.m – function for combining or selecting the body parts returned from the detection framework.

setTracks.m – function that appends the body parts detected from the detection framework as tracks in the format recognized by Redvine and appends the tracks of interacting object into it.

MakeOutputGraph.m – function which encodes all the episodes into a format that can be input to Redvine for graph mining and event recognition.

DisplayParts – Display the body parts detected in each image frame of a video.

Some of the code was used from Redvine to extract edpisodes from tracks obtained by the detection framework and modified to fit the purposes of the approach.

CallDisplayTracksAndEpisodesV2 – code for displaying episodes from tracks. Modified slightly for the project.

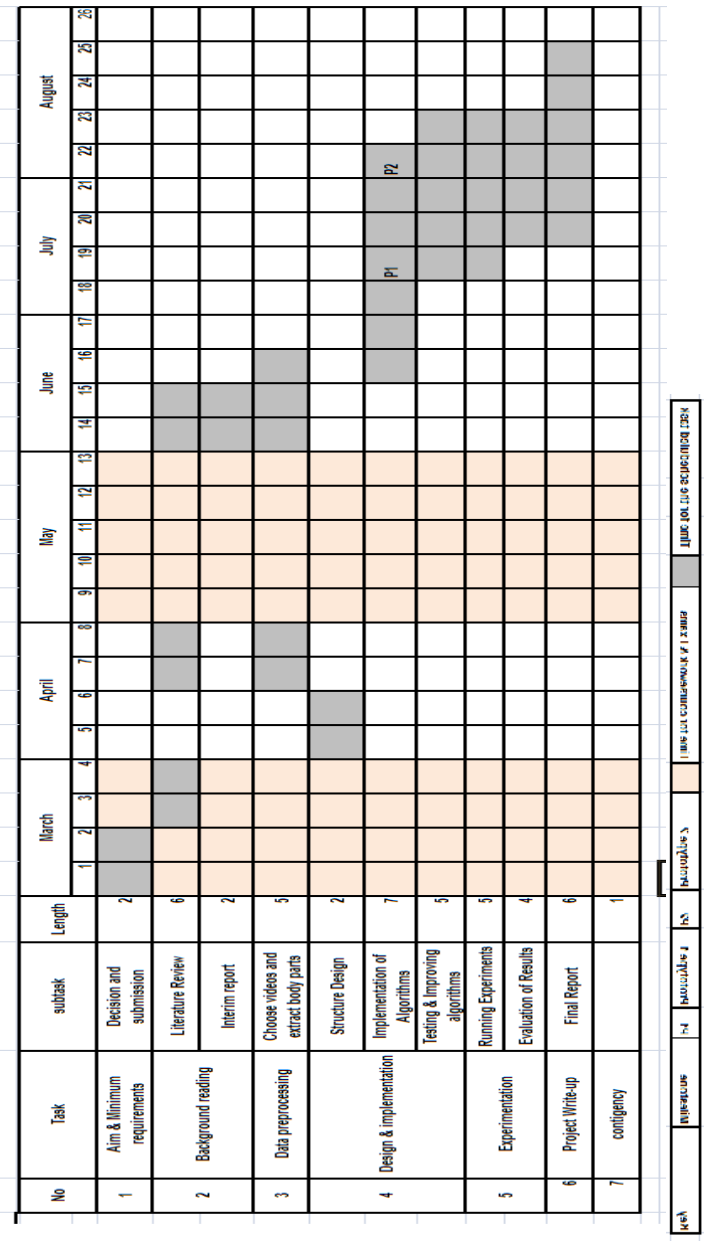
GetGraphFromClipTracks.m – The function extracts episodes from the given set of body part and object tracks. Modified to include object types information.

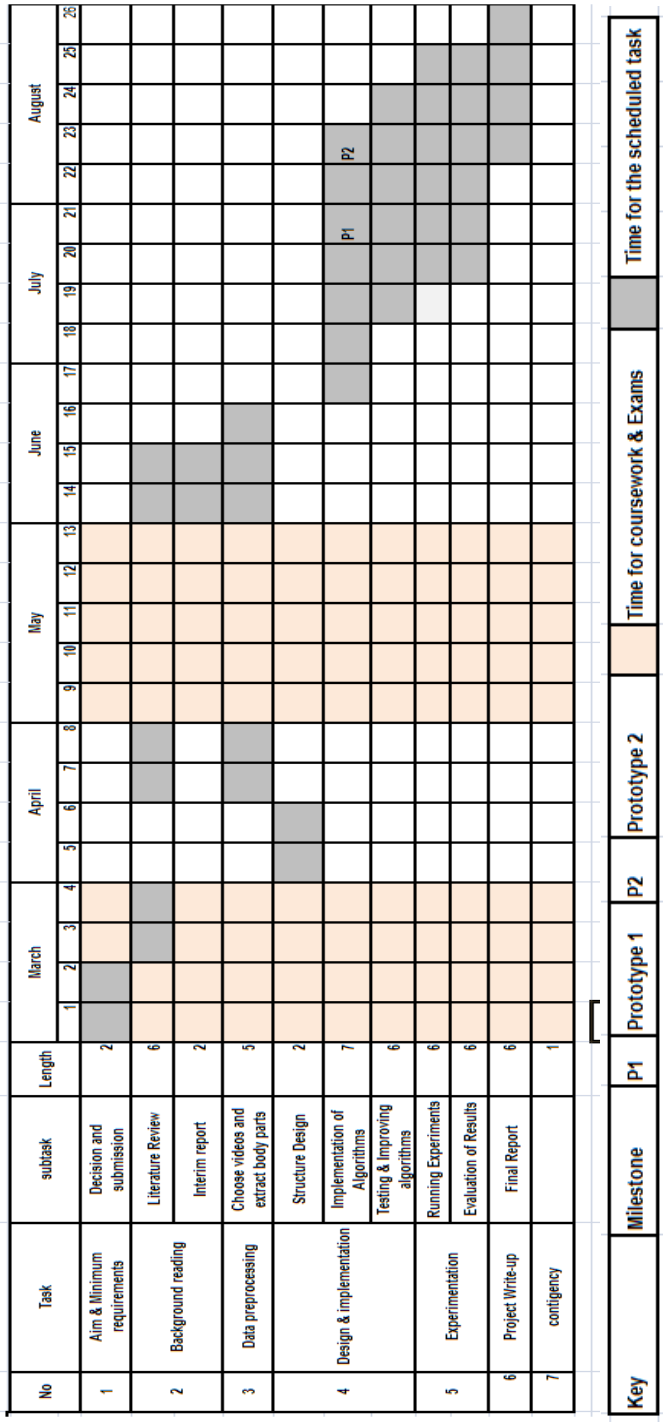
LoadParamQualRel.m – The function was modified to include object type information.

Functions for mining graphlets and event leanring were taken from Redvine without any modifications.

Appendix C

Project Schedule





Revised schedule for the project. The structure used is inspired by the structure used by a previous year project work.