Robust Human Behaviour Anticipation in Multiple Camera Environment

Doctor of Philosophy (Computer Science)

By

Shafina Bibi

92-FBAS/PHDCS/S13

Supervisor

Dr. Nadeem Anjum

Assistant Professor

CUST, Islamabad

Co-Supervisor

Dr. Tehmina Amjad

Assistant Professor

DCS&SE, FBAS, IIUI

Department of Computer Science and Software Engineering

Faculty of Basic and Applied Sciences

International Islamic University, Islamabad

2020

Dissertation

This dissertation is submitted to
Department of Computer Science and Software Engineering,
Faculty of Basic and Applied Sciences,
International Islamic University, Islamabad
As a Partial Fulfilment of the Requirement for the Degree of
Doctor of Philosophy (Computer Science)

Publication Status:

1. Published paper

Bibi, S., Anjum, N., & Sher, M. (2018). Automated multi-feature human interaction recognition in complex environment. *Computers in Industry*, *99*, 282-293.

Declaration

I hereby declare that this thesis **"Robust Human Behaviour Anticipation in Multiple Camera Environment"** neither as a whole nor as a part has been copied out from any source. It is further declared that I have done this research with the accompanied report entirely based on my personal efforts, under the proficient guidance of my supervisors Dr. Nadeem Anjum and Dr. Tehmina Amjad. I also declare that the work presented in this report has not been submitted in support of any other application or degree or qualification in any other University or Institute.

Shafina Bibi

(92-FBAS/PhdCS/S13)

Dedicated to

My Father

The late, Syed Abid Hussain Shah, a strong and gentle soul who taught me a lot of life lessons. He did not only nurture and raise me but also guided me to right actions. He is the person who has made the biggest influence in my life.

My Mother

For her motherly support and care during the moments of despair and discouragement. She has always been a source of motivation and strength for me. She taught me to keep trying without losing trust upon Allah.

Acknowledgement

In the name of Allah, the most Merciful and the most Gracious, the Lord of the worlds, and prayers and peace is upon Muhammad (S.A.W) His servant and messenger.

Alhamdulillah, all praise to Allah for giving me strength and the patience to complete my dissertation finally, after all difficulties and challenges. Words are inadequate to describe my gratitude and appreciation to Him in the whole process of completing this research.

This dissertation would not have been completed without the guidance, help, support and efforts of many people.

I would like to express my deep gratitude to my supervisor Dr. Nadeem Anjum for his constant help and guidance throughout my PhD research work. I would like to extend my gratitude to my Co-supervisor Dr. Tehmina Amjad for her help and advice regarding the process of writing this dissertation. I am thankful to Dr. Muhammad Sher for providing me guidance in initial stages of my thesis.

My deepest gratitude goes to my parents for their support and encouragement. My deceased father, Syed Abid Shah was a source of great inspiration and motivation for me. It was my father's wish to see me with a doctorate degree. Although he could not see this thesis completed, I believe that he is in heaven and will be so proud of me for achieving his dream. Not forgotten, I owe my gratitude to my mother who has been such a source of strength and encouragement throughout my PhD journey. Today whatever I have achieved in my life, all the credit goes to my parents.

I owe special thanks to my siblings, especially my brothers for their financial support and help throughout my education. Also thanks to my sisters-in-law for sincerely taking care of my parents in my absence. I am also thankful to Nageena who is always there for me and guided me in thesis formatting. Each of my family members truly deserves my deepest indebtedness for their continuous supplications for me.

I am very grateful to my teachers Mr. Kamran Hameed, Dr. Faisal Riaz, Dr. Yasir Mehmood and NTC team at MUST for their help and assistance in dataset creation part. They willingly lent me their time, expertise and resources to prepare dataset under multiple camera scenarios. Special appreciations go to Atif Butt and his fellows for recording videos of interaction dataset.

I would like to express my thanks to all my friends, especially Bushra Tufail, for the support, encouragement, patience and for being there for me when I needed. Thanks and appreciations to Asia, Faiza, Noor, Shehla, Dr. Sadia and Dr. Humaira for motivating me through my tough times. I am thankful to all my teachers at DCS & SE, IIUI, especially to Mr. Muhammad Saqlain for giving me his time and valuable feedback.

Last but not the least, no words of appreciation can sum up the gratitude that I owe to Dr. Saeedullah Shah, for his kindness and moral support during my research work. He has always encouraged me to persue academic excellence and perform at my highest level of work. My deepest thanks go to all people who took part in making this thesis real.

May Allah bless them all (Ameen)

Shafina Bibi

92-FBAS/PhdCS/S13

Abstract

Automatic video analysis (AVA) has been an active research area of computer vision and incessant efforts have been made by computer vision researchers to imitate the intellectual real-world video understanding capabilities of human brain onto autonomous vision systems. AVA is increasingly becoming vital in the context of public security, smart homes, facility protection, public transportation, human-computer interaction and automated inspections etc. Despite human activity recognition which relies on full observation of an activity; there are numerous situations like fighting, snatching and stealing in which activities must be anticipated before occurrence. It is nontrivial to achieve robust human behaviour recognition and anticipation in multiple camera views, having partial occlusions and illumination variations. The objective of this thesis is to propose solutions for resolving the aforementioned challenges for the recognition of single person actions, small unit interactions and anticipation of high-level interactions in multiple camera scenarios.

First, a new feature descriptor called Histogram of Oriented Gradient-Median Compound Local Binary Pattern (HOG-MDCLBP) is proposed to recognize single person actions in multiple camera views. In particular, MDCLBP is the proposed descriptor wherein the texture feature extraction is performed using sign difference along with median value difference. MDCLBP eliminates the impacts of illumination variations and partial occlusions. The HOG is an appearance based descriptor that provides illumination invariant representation. Actions are represented by combining histograms of both descriptors. HOG-MDCLBP achieves 96.58% average accuracy on multiple views and un-occluded dataset. On occluded sequences, it achieves 91.58% average accuracy.

Second, to recognize small unit social behaviours/interactions from sequences having illumination variation and occlusions, this thesis proposes to incorporate single person actions detected in the first step and collective poses of persons along with trajectory features. Individual actions are considered as contextual knowledge for the recognition of person-to-person interactions. The solution is based on the assumption that interactions can be correctly identified by analysing actions separately and poses collectively. The proposed method achieves 98.25% accuracy in multiple camera views.

Third, to cope with cluttered background and illumination variations for complex human interaction anticipation, this thesis proposes to use Deep CNN features and temporal features

(CNN-TOFCs). Deep features are extracted from colour images which provide spatial information about ongoing interaction. A new technique is presented to extract temporal information which utilizes the optical flow components (magnitude and orientation). To reduce the cluttered background effects and to enhance motion flow vectors; this thesis proposes to apply a transformation on optical flow magnitude and orientation. Deep CNN features and optical flow features are combined to represent person-to-person interactions from a portion of the input video for anticipation.

The effectiveness of the proposed approaches is validated on challenging multiple camera datasets having partially occluded persons, illumination variations and cluttered background.

Chapter 1

1. Introduction

AVA is an emerging trend in the field of computer vision over the past few years. It refers to the automatic monitoring of normal as well as abnormal behaviours and activities of individuals captured by the cameras to prevent a breach of security and criminal events. It has been an active area of research due to increased demand of automated monitoring in public places such as hospitals, airports, shopping malls, railway stations, sports arena, military monitoring, cinemas, parking lots, etc., and also in smart homes and elder care (see Figure 1.1). With the growing demand for public security, the use of CCTV (closed circuit TV) cameras and IP (Internet Protocol) cameras have become vital to provide maximum surveillance in above mentioned areas. Clearly, video cameras have become an implicit part of our social lives. These cameras are not only used to detect unusual events but also used to prevent the occurrence of these events. A study conducted by Mazerolle et al. [1] shows that people's behaviour has also changed when a public space is monitored by surveillance cameras which result in the reduction of crimes.

Large space public areas required the use of multiple cameras to monitor the entire scene from different viewpoints to analyse activities. CCTV and IP cameras footage monitoring has become ubiquitous which requires a human agent to monitor the screens all the time. Recorded footages are also used later to investigate crimes. Manual analysis of recorded videos is very inconvenient and time consuming. Human resources to monitor the screens are very limited because it is a very tedious job for a human to look at multiple screens constantly, which results in poor monitoring. With the growing number of available technology and easy access to videos, the requirement for their automatic understanding has become vital. It is required to have an intelligent surveillance system that performs intelligent detection of ongoing activities within the area of interest.

Figure 1.1: Various areas with CCTV monitoring; top row (airport, shopping mall, sports arena), bottom row (military monitoring, parking lot and smart home)

The key objective of AVA for activity recognition is to monitor and recognize the activities of people in a specific area to provide video surveillance. The application of video surveillance is not limited to the recognition of unusual activities [2]. These systems are now also implemented in other areas like in offices to monitor employees' activities, traffic monitoring and in playgrounds etc.

Only the complete activity after its occurrence is recognized by these systems. However, researchers are now migrating towards the ***early recognition*** of on-going human activities or behaviours in a video.

The goal of behaviour ***anticipation*** is to infer an activity in its early stages [3]. The anticipation of an activity before it is completely executed is essential in many applications, e.g. smart homes for elder care (falls detection of elderly people), human-robot interaction (to make the robot able to plan ahead to aid human) and many other surveillance applications designed to prevent abnormal events. In this thesis, we have analysed, recognized and anticipated person-to-person interactions under multiple camera environments.

1.1 Human Action, Behaviour Recognition and Anticipation Taxonomy

A moving human can perform different types of activities in a scene. Various taxonomies have been proposed in the literature to define these activities. Human activities can be

divided into four key categories: (1) gestures, (2) actions, (3) behaviours and (4) interactions [4].

1.1.1 Gestures

Gestures are the movements of body parts to express an action or to communicate a message. Main visible body parts include head, hands and legs. The common gestures are stretching arms or raising legs, moving head etc.(see Figure1.2) Previously, many researchers have focused on recognition of gestures [5] [6] [7]. Gesture recognition has been the major research topic for human-robot interaction systems.

Figure1.2: Example of Gestures, DvsGesture dataset [8]

1.1.2 Actions

Gestures are the atomic elements and actions are composed of these elements. The actions can be described as distinct and periodic motion patterns. Single motion patterns include *jumping*; *bending*, *pointing* etc. and periodic motion patterns are *sitting*, *running*, *walking*, *swimming*, and *jogging* [4] (see Figure 1.3). One action can be further divided into sub actions i.e. three sub actions are observed in sitting: *stand*, *bend*, *sit*; *stand up* is subdivided into: *sit*, *bend* and *stand*.

Figure 1.3: Example of Action, KTH [9]

1.1.3 Behaviour

Human behaviour is the way in which a person behaves in response to a particular situation or stimulus[1]. Behaviour is a combination of actions and activities. In a real-world scenario when a person interacts with another person or object, responses are generated due to those interactions. For example; in a street, a person may find his friend, approach him, and talk to him and finally either they walk separately or walk together. A person may carry an object and put that object

[1] http://www.oxforddictionaries.com/definition/english/behaviour

into a vehicle (see Figure 1.4). In this study, ***the terms behaviour and interaction are used interchangeably*** because we are talking about social behaviours i.e. interaction between persons.

Figure1.4 Examples of behaviour (a) sports (b) robbery (c) Catching a bus (d) interaction with another person [4]

Interaction behaviour between two persons can be divided into two categories:

1.1.3.1 Small Unit Interactions

Small unit interactions are the "sub interactions" of an interaction sequence between two persons [10]. High-level actions and interactions are composed of these small unit interactions. These interactions include*: talking, walk separately, walk together and stand together* etc. Recognition of such interactions is mainly required in environments where people interact for a short period of time such as in office lobby and playgrounds.

This research assumes that small unit interactions can be recognized accurately by considering the individual person's actions and trajectory information.

1.1.3.2 High-Level Interactions (Complex activities)

High-level interactions or complex activities between two persons are the longer sequences which include: *handshake, hug, kick, punch, push,* and *point* etc. [11], [12]. Recognition of these high-level interactions is required for high-level surveillance and many other applications.

This thesis aims to anticipate high-level interactions between two persons under multiple camera environments.

1.2 Individual Human Action Recognition

Actions are single or periodic motion patterns performed by humans. Individual human action recognition is achieved by analysing actions of each person exclusively in a video frame. Human action recognition has been an active area of computer vision research due to its concern with numerous key applications like video surveillance, smart homes, digital entertainment or human-computer interaction. The objective of automatic human action recognition is to analyse the actions automatically from unknown videos. Generally, computer system observes an environment using a video camera and the task is to extract useful information and recognize actions on the basis of that information.

Security cameras are installed and connected with a computer system for activity monitoring. Several actions can be observed in a scenario i.e. *walk, jump, clap, sit, stand* and *bend* etc. The main purpose of using security cameras is to monitor the activities in a public place for automatic monitoring. Surveillance applications require monitoring of large public spaces from different viewpoints. The use of multiple security cameras (either CCTV or IP cameras) has become requisite in public places to monitor the entire area.

The task of human action recognition is done by extracting representative features (action representation), learning the classifier using extracted features and recognition of incoming actions on the basis of learned examples.

1.3 Human Interaction Recognition

The analysis and recognition of person-to-person interaction behaviours is a vital task in automatic video analysis. It has attracted an enormous body of research during the last few decades. Due to the increased demand of automatic monitoring in public areas, researchers are now moving from recognizing simple human actions [13]–[16] towards complex human behaviours/interactions [17]–[21].

This thesis also focuses on the recognition of the small unit social interactions i.e. person-to-person interactions in public areas under multiple camera views. *Walk together, stand together, walk separately* and *talking* are examples of small unit interactions. Interaction recognition requires joint observation of persons in a video to observe the relationship between them. Single person action recognition system alone cannot fulfil the requirements of a full surveillance system. In this context, recognizing interactions is a very important task because activities in any public environment are hardly ever performed distinctly [22]. Owing

to above observation, individual actions are considered for the analysis and recognition of person-to-person interactions.

1.4 Human Behaviour/Interaction Anticipation

Unlike traditional interaction recognition techniques, human interaction anticipation aims to recognize an interaction before it is fully executed. AVA for surveillance applications has attracted the considerable attraction of researchers towards the ***recognition and anticipation*** of human interactions.

The field of human behaviour recognition has grown efficaciously in the last 10-15 years. However, many real-world scenarios demand the early recognition of human behaviours to prevent criminal activities such as physical violence, theft, robbery and snatch etc. Consequently, human behaviour anticipation has become an active research area in this decade. Specifically, in the past five years, the domain of automatic behaviour anticipation has developed rapidly. Although many researchers have focused on behaviour anticipation [3], [23]–[25], it is still rather a new area for computer vision researchers.

The anticipation of ongoing human behaviours is crucial in many domains such as video surveillance (to detect violence activities), robotics (human-robot interaction), intelligent tracking systems (location prediction) and nursing homes (fall detection) etc. furthermore, the vehicle monitoring systems can predict the accidents by using prediction models and can help to prevent accidents. The behaviour anticipation applications can also be used to provide high-level surveillance in crowded areas by monitoring interactions and alerting abnormal movements by making early decisions. In public areas, the goal of the anticipation applications is to generate the alarm in advance of any unusual activity. This is a challenging problem since a human can easily guess the forthcoming event by looking at the event pattern. How can a machine solve this problem? It is tricky to teach a machine using vision algorithms to generate alarm before the event occurs.

Advances in the field of AVA have made this imperative application become real: anticipation/prediction of activities or complex interactions from partially observed videos. Several factors can affect the performance of behaviour classification and anticipation applications: Partial occlusions, illumination variations and viewpoint changes [26].

An example of full and half observations is presented in Figure 1.5 which shows the difference between recognition of complete activity and early recognition of ongoing activity.

This thesis focuses on anticipation of high-level interactions which includes: *bend, faint, handshake, hug, kick, punch and push*.

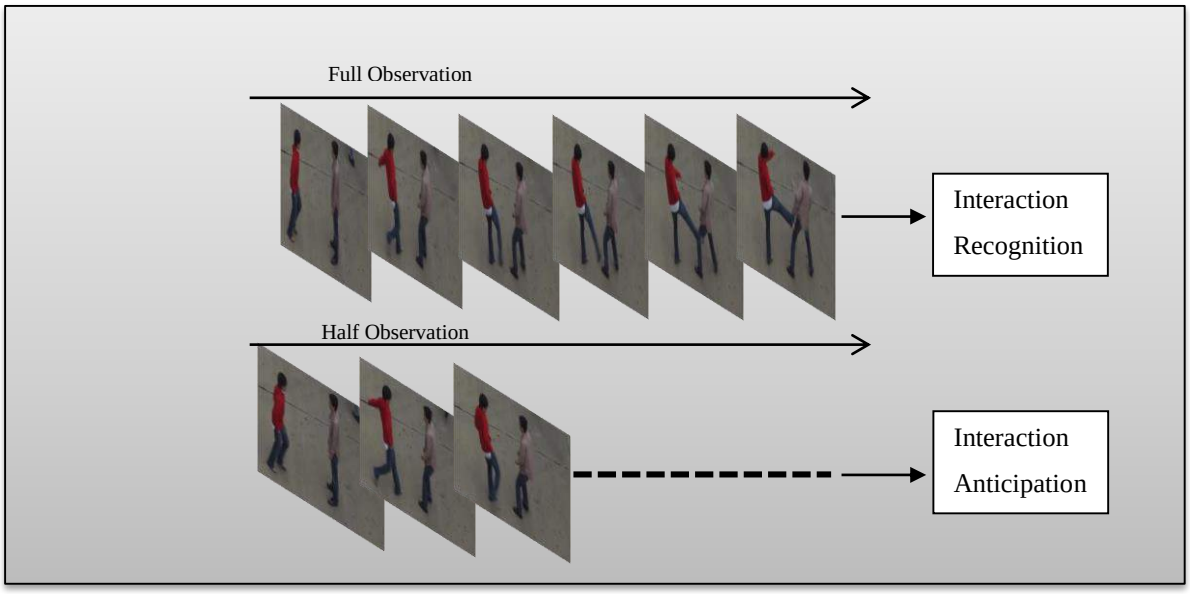

Figure 1.5: Frames from UT-interaction [27] dataset showing full observations and half observations

1.5 Objectives

The primary objective of this research is twofold: a) *to recognize small unit interactions between two persons* and b) *to anticipate high-level interactions between two persons; in multiple camera scenarios*. Specifically, the research focuses on following objectives under multiple camera environments:

1. Individual Human Action Recognition

To develop a method for the recognition of individual human actions from videos captured with multiple cameras in public environments.

2. Human Interaction Recognition

To develop a method for recognition of small unit social (person-to-person) interactions in public environments covered with multiple cameras.

3. Human Interaction Anticipation

To anticipate high-level person-to-person interactions (complex activities) in outdoor public environments monitored with multiple cameras

Next section discusses the primary challenges to accomplish the above-mentioned objectives, which serve as the motivation of this research.

1.6 Challenges and Motivation

Single camera systems are inadequate to fully observe the scene from different views hence lacking information in all aspects. The motivation behind using multiple cameras is to monitor the scene from different views for accurate analysis of human behaviours. Multiple cameras can be installed with different degrees of overlapping (See Figure 1.6) [28]. In this thesis, the experiments are performed on multiple camera views using the layout as in Figure 1.6 (b) (all cameras partially overlapping).

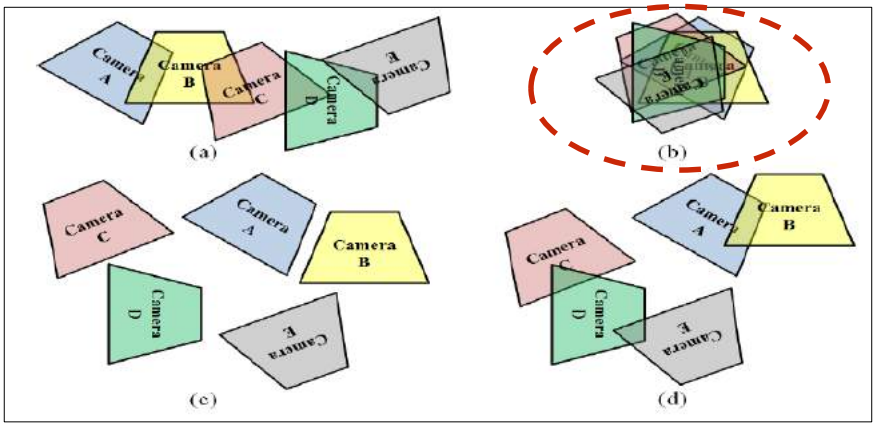

Figure 1.6: Degrees of overlap between multiple cameras: (a) partially overlapping adjacent cameras, (b) Full overlapping, (c) non-overlapping (d) different types of overlapping between cameras. [28]

1.6.1 Analysis of individual actions and small unit interactions under multiple camera views

Multiple camera systems are generally deployed in public places to provide large scale visual coverage for automatic monitoring. Despite large scale coverage, another goal of multiple camera scenarios is to cover the critical areas (entrance of a building, crowded areas, smart homes etc.) from different views. In multi-camera scenarios, all cameras share information to decide which activity is performed in the region of interest. Individual person action recognition and interaction recognition under multiple camera scenarios are mainly hindered by following two issues:

Partial Occlusions

Persons moving across camera views can be occluded by other persons or objects, resulting in difficulty to analyse individual actions and behaviours (Figure 1.7). Analysis of human actions comprises feature extraction and representation; occluded persons make the analysis

task very challenging. Feature representation should be strong enough to recognize the individual actions even if the part of the person is occluded [29].

Figure1.7: Example of partial occlusions due to objects and other persons [15], [30]

Illumination variations

In multi-camera scenarios, cameras are deployed in different positions and persons moving across cameras can experience illumination variations (Figure 1.8). This is the challenging issue that increases uncertainties in the behaviour analysis task.

Figure 1.8: Example of illumination variations. Images are taken with 3 cameras installed in different locations [30].

1.6.2 Anticipation of High-level Human Interactions under Multiple Camera views

The anticipation of human interactions plays a vital role in automatic monitoring systems. The goal of behaviour anticipation systems is to predict the activity before its execution is completed. This research has focused on the anticipation of interactions between two persons in multiple camera views. A concentration of work on the problem of human interaction anticipation has been seen in recent years [3], [31], [32]. However, the area of human behaviour/interaction anticipation ***under multiple camera views*** is still unexplored. Apart from above mentioned challenges, following is the challenge to accomplish behaviour anticipation task:

Analysis of Partial Observations for Behaviour/Interaction Anticipation

As compared to human interaction recognition, interaction anticipation requires to infer the interactions on the basis of partial observations. It is a challenging task to infer an activity at its early stages. The initial pattern of some interactions is much similar i.e *handshake and hug*, *punch and push*, *faint and bend down*. It requires visual observations to be strong enough to discriminate the aforementioned similar patterns.

1.7 Problem Statement

In this research, the issue of recognition and anticipation of person-to-person interactions in multiple camera view environments is addressed. There are various factors that make accurate interaction recognition and anticipation task challenging in outdoor environments.

1. Partial occlusions and illumination variations in videos affect the accuracy of human action and interaction recognition [9-11].
    a) Visual recognition systems rely on visual information of persons. Accuracy of these systems is decreased due to insufficient information if the person to be monitored is partially occluded by some other object or persons.
    b) Illumination variations change the image intensity values.
2. Interaction anticipation is performed on partial observations. The task of anticipation becomes more challenging in the presence of cluttered background, partial occlusions and variations in illumination.
    a) Existing methods for interaction anticipation [3], [23], [33] are not handling the aforementioned problems in multiple camera view outdoor environments.

1.8 Contributions

The contributions of this thesis to individual human action recognition, small unit human interaction recognition and anticipation of high-level interactions under multiple camera scenarios are as follows:

1. A novel texture based feature MDCLBP is proposed. MDCLBP is a novel approach that extracts texture by using median value and sign difference. MDCLBP is invariant to partial occlusions and small illumination variations. The hybrid approach HOG-MDCLBP is introduced to describe individual actions. Both HOG and MDCLBP are appearance based features. HOG is independent of illumination variations. Occlusion

and illumination invariant description of actions is achieved by combining HOG with MDCLBP.

2. Individual human actions and collective poses along with other trajectory features are used to represent small unit interactions between persons in a frame.
3. Deep CNN features and hand-crafted (temporal) features: the Convolutional Neural Network-Transformed Optical Flow Components (CNN-TOFCs) are used to anticipate ongoing human interactions. Optical flow is used to extract temporal features. A transformation is applied on optical flow components (magnitude and orientation) to eliminate the effects of scene variations from input sequences.
4. Compared to human interaction recognition and anticipation in a single camera view, no dataset is available for multi-camera high-level human interactions in outdoor environments. A new dataset has been created to assess the proposed approach on multiple views for the anticipation of high-level interactions.

1.9 DataSets

The efficacy of proposed method for the recognition of individual's actions is validated on two datasets IXMAS (INRIA Xmas Motion Acquisition Sequences) and OIXMAS (Occluded INRIA Xmas Motion Acquisition Sequences) [15] [31]; these are the contemporary benchmark datasets for action recognition under multiple views and partial occlusions. The proposed method for the recognition of small unit human interactions is tested on HALLWAY dataset [30] that is multi-view dataset. A new dataset (MU-Interaction) is also introduced for high-level human interaction anticipation in multiple camera scenarios.

1.10 Classification Method

In this research, SVM is used for the classification of human actions and interactions. SVMs have been the most prominent among other machine learning algorithms for data analysis and recognition tasks. SVMs belong to supervised learning models which use training examples to train and build a function that can classify the input examples and return the class labels as output. It uses kernel trick to perform non-linear classification. The baseline SVM is used for the recognition of individual human actions and interactions due to their successful repute in many computer vision applications [17], [35], [36].

1.11 Thesis Organization

The remaining chapters in this thesis are structured as follows:

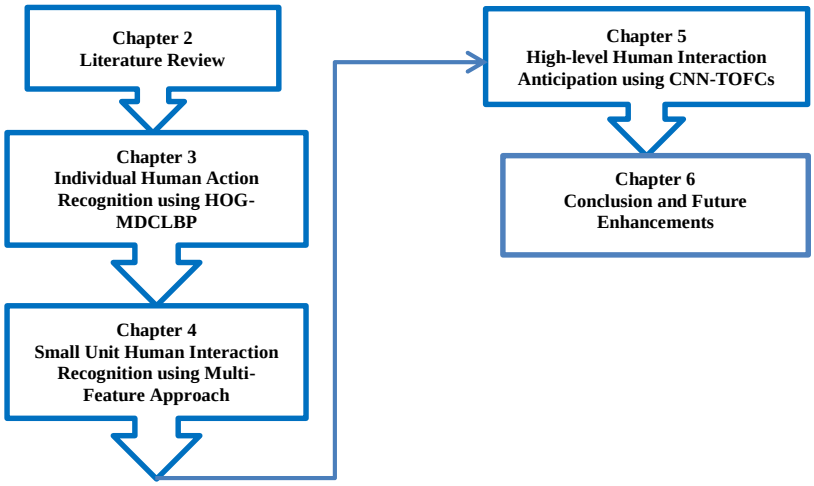

Figure 1.9 Thesis Organization

Chapter 2 provides a summary of previous research on human action recognition, human interaction recognition and human interaction anticipation. Previous work is divided into three categories including both single and multiple camera environments: (1) individual human action recognition, (2) human behaviour/interaction recognition and (3) human behaviour/interaction anticipation. This section provides extensive literature over the last 10 years on representation and classification methods in all categories. Gaps in previous research are identified and the problem is formulated at the end of this chapter.

Chapter 3 presents a new feature descriptor for individual human action recognition in multiple camera views to handle partial occlusions and illumination variations. Mathematical formulation of the proposed descriptor, histogram of oriented gradient- median compound local binary pattern (HOG-MDCLBP) is presented in detail. Robustness of the proposed method against partial occlusions and illumination variations is validated by performing experiments and comparing results with state-of-the-art approaches. The proposed method is also tested on cross camera views for action recognition.

Chapter 4 presents a method for person-to-person small unit interaction recognition in multiple camera environments. In this chapter, the individual action labels detected in Chapter 3 are concatenated with the set of features extracted from trajectories of both persons. Experiments are performed on videos captured with multiple cameras. The efficacy

of proposed interaction recognition method is assessed by performing experiments on multi-view dataset.

Chapter 5 presents a method for anticipation of high-level interactions between two persons in multiple camera scenarios. A new descriptor convolutional neural network- transformed optical flow components (CNN-TOFCs) for human interaction anticipation is proposed in this chapter. A new dataset is also presented in this chapter for the evaluation of proposed anticipation method.

Chapter 6 concludes the thesis with a discussion of contributions that are made for individual human action recognition, human behaviour/interaction recognition and human behaviour/interaction anticipation in public environments under multiple camera views. Further enhancements are also discussed in this chapter.

Chapter 2

2. Literature Review

In this chapter, the detailed study of previous approaches to human behaviour recognition and anticipation is provided. Since a large amount of literature about behaviour recognition and anticipation is available, this chapter will focus on the research that is closely linked to the work presented in this thesis. This chapter provides a review of the existing methods for human behaviour analysis focusing on vision based approaches. Section 2.1 covers individual human action recognition techniques. Human behaviour/interaction recognition techniques are reviewed in section 2.2 and finally, behaviour anticipation techniques are explored in section 2.3. After an extensive review, research gaps are presented in Section 2.4 and problem is formulated in Section 2.5.

Literature review is divided into three categories; (1) Individual Human Action Recognition (2) Human Behaviour Recognition (3) Human Behaviour Anticipation; under single camera and multiple camera environments.

2.1 Individual Human Action Recognition

This section depicts various human action recognition techniques in the literature. Common individual human action recognition methods have been discussed in detail.

2.1.1 Motion and appearance based features

Motion and appearance based action representation approaches are commonly used in vision based human action recognition methods. Action representation as feature vector is a vital step for the recognition of actions. ood features are likely to be robust against different variations i.e. clothing, illumination, scene etc. Silhouette-based, motion-based, body part-based, frame-based and interest points based feature representation approaches are very common in literature.

In earlier works, the motion information was extracted from input video frames for action representation. Silhouette-based motion description approaches are very common, where the human silhouette is evaluated to extract activity features. These approaches are generally based on the process of background subtraction. Motion energy image (MEI), Motion history image (MHI), Localized motion energy image (LMEI) and Non-parametric weighted feature

extraction (NWFE) are some key silhouette-based feature representations methods. In [7] an appearance based approach for human action recognition is presented to decrease the view point dependency. Actions are represented as temporal templates i.e. MEI and MHI. 7 Hu moments are computed given MEI and MHI of an image. Each of the testing images is recognized by finding the Mahalaonobis distance between the mean and variance of saved image moments and input moment. Power of recognition method is increased by using two cameras and incidental motions are handled by using only background subtracted images instead of using temporal templates. Experiments are performed on aerobic data with single and multiple views. This approach is applicable to a single person and fails when there is another person in the field of view.

The spatial distribution of motion energy is captured along with MEI and named as LMEI [37]. A score based fusion method is presented to handle the random orientation of the person with reference to cameras. Classification is performed on extracted LMEI features by using linear discriminant analysis (LDA) classifier. 90% accuracy is achieved when fusion is performed on LMEI based classifiers. The problem with this technique is that by increasing the number of classes, the LMEI tends to be similar for different actions. Moreover, they have ignored the camera network effects which play a vital role in real time processing.

Wo and Shao [38] extended MHI by adding two holistic descriptors: gait energy image (GEI) and inversed recording (INV) to recompense for the loss of information in MHI. They represented human actions using silhouettes and constructing correlogram matrix from a sequence of poses and named as bag of correlated poses (BoCP) that is an extension of the bag of visual words (BoVW) model. The extended MHI is presented by adding INV and GEI information. BoCP is fused with extended MHI to further improve the results. Classification is performed with SVM using Gaussian kernel. Experiments are performed on Weizmann and IXMAS datasets and used only the single view for training and testing. Maximum accuracy of 90.3% with IXMAS dataset in camera 2 and 97% accuracy is achieved for Weizmann dataset.

Hu et al.[39] recognized indoor human actions captured from depth sensors. 3D MHIs is extracted from depth data and combined with color information and scene semantics for the classification of actions. The 3D depth MHIs contain both forward and backward depth MHIs which help to encode increase and decrease in depth information. Experiments are performed on self extracted depth data in an office environment containing six different actions. 93.33%

accuracy is achieved on actions by considering scene semantics and 85% accuracy is achieved exclusive of scene semantics.

Han et al. [40] propsed a localized temporal representation for the recognition of human actions. MHI and MEI templates are extracted form input video sequences. Binarized statistical image features (BSIF) are used to obtain local information from MHI and MEI templates. A linear SVM classifier is used to classify localized temporal representation. Experiments are performed on KTH dataset with 90% accuracy.

Singh et al. [41] computed directional vectors (DVs) by representing human silhouette boundary with chain codes. An adaptive algorithm for the separation of background from foreground is applied and an edge detector is applied to extract silhouettes from foreground image. To reduce the size of feature vector, silhouettes are represented with chain codes which are named as DVs. Normalization is applied on DVs to achieve the scale invariance. The angular distance among two DVs is computed to analyse the extracted vectors. The image frames having similar activities are clustered on the basis of angular distance between DVs. UoS-HID and CMU-Mobo datasets are used for experiments. 85% to 95% accuracy is achieved without temporal smoothing and accuracy is increased to 100% when temporal smoothing is applied. Though this technique can handle the changes in scale, background, clothing and view angle, it is not compatible with persons having different body shapes.

Lin et al. [42] presented a nonparametric weighted feature extraction technique for human activity recognition. The distance signal feature and the width feature are extracted from human's silhouette to build NWFE features. Both features are combined and principle component analysis (PCA) is used for dimensionality reduction. A codebook is then generated from extracted features using k-means clustering algorithm. Extracted features are mapped to symbols and represented with histogram vectors. Experiments are performed on Weizmann dataset with 100% accuracy. The proposed feature extraction technique becomes challenging if the person is occluded.

Vishwakarma and Kapoor [43] first extracted key frames of human silhouettes and extracted features from these frames. To produce the descriptor, the key poses of silhouettes are represented by cells and grids. Parameters of cells and grids (count of white pixels) are computed to model the feature vectors. A hybrid classifier "SVM-NN" is constructed to improve the recognition accuracy. The main drawback of the silhouette-based approaches is

that that the extracted silhouettes are not robust to occlusions. These methods perform well in controlled environments.

Besides the silhouette-based approaches, many motion-based techniques compute motion descriptors directly from consecutive images. The Popular ones include Motion Binary Pattern (MBP), Volume Local Binary Pattern (VLBP), Optical Flow and Histogram of weighted Optical flow.

Histogram of oriented optical flow (HOOF) feature is proposed by Chaudhary et.al. [44] to make the action representation independent of a person's scale and moving direction. HOOF features are modelled with a non-linear dynamical system (NLDSs). The similarity between two NLDSs is measured by using Binet Cauchy Kernels. Experiments are performed on Weizmann gait dataset with 94.4% accuracy. The problem with this method is that it cannot handle multiple disconnected motions in a scene.

Histogram of weighted optical flow (HOWOF) is presented by Mukherjee et al. [45] to derive the motion and pose information from sequence of frames. The vocabulary of poses is built and key poses are selected by applying centrality measure algorithm. The discriminatory codebook is built with finally selected best poses and classification is performed with SVM model.

MBP is a motion descriptor presented by Baumann et al. [46] for multi-view action representation. The MBP is inspired by the VLBP [47] to describe motion information in spatio-temporal space. MBPs are computed from three consecutive images to describe grey-level changes in an image. Motion pattern in an image is represented with histogram and classification is performed with Random Forest classifier. Experiments are performed on KTH, Weizmann and IXMAX datasets. MBPs are evaluated on multi-view dataset but not tested for occluded scenarios.

Kihl et al. [48] presented a motion descriptor by performing a coding step on optical flow field. They performed half-wave rectification on the vector field of optical flow and the proposed descriptor is named as series of polynomial approximation of flow (SoPAF). Vector of locally aggregated tensors (VLAT) indexing method is used to obtain signatures from the descriptors and linear SVM is used for classification. Experiments are executed on UCF11 and HOLLYWOOD datasets which contain realistic videos of action classes taken from YouTube and video clips. These are single view dataset and accuracy on both datasets is improved when compared with previous methods.

Cai et al. [49] proposed a two stream architecture for improvement in the accuracy of hockey actions recognition. Part affinity fields model is used to derive significant hints from the palyer. Temporal features extracted using optical flow and fused with pose information. The fused information is then passed to fully connected layers for the recognition of hockey actions. Experiments are performed on HARPET dataset with 85% accuracy.

Optical flow based features are very sensitive to background noise and movement change. The variations of optical flow can be used to overcome these limitations.

2.1.2 Part-based approaches

Part-based approaches consider human body parts motion instead of analysing the motion of whole body for action classification. After pose estimation from the human body, features are extracted by using different feature representation techniques [50] [51].

Ben Arie et al. [52] represented the activity with human poses along with velocity vectors of major parts of the body, such as hands, legs and torso. A database is constructed and poses and velocity vectors of each body part are stored in multidimensional hash tables. Voting for each body part is performed and combined the votes of all body parts in a frame. Final activity class is decided on the basis of the highest score from all body parts. This method performs well with smooth motion and un-occluded scenarios.

To overcome the challenges of low resolution videos, Danafar and Gheissari [53] captured local and global motion by extracting histograms of optical flow. The global motion descriptor is described by the histograms of horizontal and vertical motion. Local motion is extracted by dividing the human body into three parts and histogram of optical flow is computed from each body part. A SVM model is trained on horizontal and vertical optical flow field for action recognition. Experiments are carried out on KTH dataset with 85% accuracy on un-occluded scenarios.

Another part-based motion descriptor for human activity recognition is presented by Tran et al. [51]. They represented motion of each body part using polar space. Both local and global representations are combined by transforming the global coordinate space into local coordinate space with centre of torso being the origin. The motion of each body part is represented by a 2D histogram and combination of body parts histograms are used to represent each action. Each histogram bin represented number of time the individual body part is observed at a location. They proposed a recognition algorithm which utilized sparse

representation for classification. Experiments are performed on KTH and UCF sports datasets with 97.83% and 91.62% accuracy respectively. This technique is not tested on occluded and multi-view scenarios.

Ke et al. [50] extracted human silhouette and reconstructed 3D human poses from 2D body parts blobs. Hands (right and left), head and feet (right and left) are tracked in each frame on the basis of colour, shape and texture information. After constructing 3D projections 15 features are computed by applying geometric relational features (GRF) conversion. The GRF included nine distance related features and six angle related features. Continuous actions are recognized by concatenating different trained cyclic HMMs. They trained different Cyclic HMMs for different types of actions and proposed a graphical model for concatenation of all Cyclic HMMs. Experiments are performed on self-recorded dataset and IXMAS dataset.

The part-based feature representation techniques perform well if the respective body parts are not occluded.

2.1.3 Frame-based approaches

Another common approach is the frame-based feature representation, where temporal relationship between frames is not considered. HOGs and space-time interest points (STIPs) are most widely used frame-based feature descriptors. Frame-based feature extraction is a straightforward task and it is useful for the images which do not contain rapid local changes. Since these features do not include motion information, researchers have used a combination of frame-based and temporal features to encode motion information.

The HOG was originally proposed by Dalal and Triggs [54] for human detection. Further, HOG along with other descriptors is used for human action recognition. Weinland et al. [34] proposed the extension of the HOG descriptor and represented a sequence of images with 3DHOG to achieve robustness to occlusions and multiple viewpoints. They computed 3DHOG at densely distributed locations within the selected region. Local classification is performed on densely computed descriptors and followed by global classification to provide robustness to viewpoints and occlusions. Global classification is performed by combining the results of local classifiers. Experiments are performed on Weizmann, KTH, UCF, IXMAS and OIXMAS datasets. This technique is tested on both multi-view and occluded scenarios and improved the performance on occluded scenarios.

Cao et al. [55] combined HOG and HOWOF for action representation. They proposed a codebook reduction method by selecting key poses from visual words. Feature descriptors are created by concatenating histogram of HOG and HOWOF which are extracted from each video frame. HOG provides local structure information and HOWOF provides local motion information. Dimensionality reduction is performed using locally linear embedding (LLE) algorithm and PageRank algorithm is applied to select the key poses. Experiments are performed on UT-Tower and KTH datasets. Since the descriptors are extracted from the whole image, computation becomes complex in high quality videos.

Mosabbeb et al. [56] extracted histogram of gradient (HoG) and histogram of optical flow (HoF) features for human action recognition in multiple distributed camera network. They proposed consensus based multi-view distributed network classification framework that depends upon matrix completion method. They have extracted dense features under each camera view by computing HoG and HoF on three different scales. Each camera contributes to the final decision by locally processing the input video sequence and deciding the activity label. Experiments are performed on IXMAS and MuHAVi datasets.

A low dimensional representation of multi-view data is proposed by Murtaza et al. [57]. They represented actions with HOG of MHI images. They presented a silhouette-based approach for action recognition from multi-camera videos by first extracting MHIs and then computing HOGs of all MHI images. Classification is performed by using a NN classifier and experiments are performed on MuHAVi dataset.

Saho et al. [58] proposed to fuse optical flow and HOG based features for action recognition. To distinguish the acitivies that vary in speed, bag of histogram of optical flow (BoHOF) is proposed. HOG features are extracted from color images and PCA along with pearson correlation is applied to reduce HOG features. Befor computing BoHOF, humans are segmented from background and BoHOF are obtained at segmented regions boundaries. Both features are normalized and then fused together to attain the advantages of both features. Experiments are performed on KTH dataset with 96% accuracy.

2.1.4 Space-time interest points

Spatio-temporal local features are extracted by detecting the STIPs using any corner detection algorithm i.e. [59]. These STIPs either can be directly used as BoW representation [14] or can be described by extracting other features around those STIPs.

Mattivi and Shao [60] represented local information by extracting cuboids from STIPs. They described each cuboid by applying Local Binary Pattern (LBP) on three orthogonal planes which is called LBP-TOP descriptor. K-means clustering is applied to build a visual vocabulary. Classification is done using SVM on KTH dataset and achieved 88.19% accuracy.

A multi-camera view action recognition method is proposed by Lin et al. [59]. They first extracted STIPs by using Harris 3D detector and then HOG/HOF descriptors are extracted from detected interest points. Bag of words (BoW) model is constructed with extracted space time (local) features and under each view. Secondly, global motion context feature is extracted which includes histogram of silhouettes and optical flow inside human bounding box. Global features are also used to create BoW model. Both local and global features are then concatenated to form a hybrid BoW model. A pair of dictionaries is learnt and from two action videos taken from two different views. Features between two pairs are linked using the learnt dictionaries to provide view invariant space. This representation improves the performance of cross view recognition because the model learnt in source view can be used to recognize the action in target view. Experiments are carried out on IXMAS dataset with average accuracy of 98.6%.

Multiple kernel learning based fusion framework is proposed by Gu et al. [61] for real time action recognition. They used Gaussian mixture model to extract the moving person and detected space time interest points, HoG and HoF from detected person. A BoW model is built from extracted features. Multiple kernels are considered that correspond to data sources from multiple views and combined using convex function. Experiments are performed on IXMAS dataset with 95% accuracy.

Table 2.1: Characteristics of various Human Action Recognition techniques

Ref.	Features	Classification	Comments
[7]	MEI, MHI	Mahalanobis distance	• Real time and multi-view. • Cluttered background can affect the system accuracy. • Cannot handle self-obstructed movements.
[41]	DVs extracted from chain codes	Angular distance	• CRRs range from 85% to 99%. • Not compatible with the dataset having persons with significantly different body shapes.

			• Not tested for occluded persons
[37]	LMEI	LDA	• Combined information from multiple views • As the number of action classes increases, the localized motion history image turns out to be more similar.
[39]	3D MHIs and scene semantics	SVM	• Videos captured with multiple depth cameras in indoor environments. • Uses scene sematics as context • Use of semantic information increased recognition accuracy
[40]	MHI and MEI	SVM	• Single view • Experiments are performed on scenarios where only one person is performing actions. • The proposed advantages of MHI and MEI are not tested on challenging datasets.
[42]	Width and distance signal of body contour	NWFE	• Feature extraction is challenging if the silhouettes are not properly extracted due to occlusions.
[38]	Extended MHI	Gaussian kernel SVM	• Multi-view • Applicable to single person and non-cluttered background
[43]	Number of White pixels in key frames	Hybrid classification SVM-NN	• Significantly improved results as compared to other silhouette-based approaches [62]–[65]. • It is vitally important that only one person is in the input video. • Less effective if the person is occluded.
[46]	MBP, VLBP	Random Forest	• Multi-view • Not tested on public datasets with moving background. • The method to choose optimal cell size is not defined.
[48]	SoPAF	SVM	• More realistic dataset is used as compared to previous action recognition approaches. • Maximum achieved accuracy is 86.0% that can be further improved.
[44]	HOOF	Extension of Binet-Cauchy	• Background subtraction and human detection is not needed for feature extraction.

		kernel to NLDS	• Unable to handle multiple disconnected motions in a scene.
[45]	HOWOF	SVM	• Key poses are used for classification • Need to limit optical flow vectors for noisy observations
[49]	Pose and optical flow	Fully connected layers	• Part-based method • Tested on un-occluded single person dataset.
[50]	GRF descriptor: distance features and angle features	Cyclic HMMs	• Tested on single actions and continuous actions. • Mainly depends upon the accurate extraction of body parts
[51]	Polar histogram of human body part motion	Sparse MDI	• Both local and global information is incorporated for robust representation. • Not tested on multi-view dataset. • There is a need to encode temporal information to discriminate certain actions.
[53]	Histogram and vertical Optical flow	SVM	• Action recognition from low quality videos. • Tested on different scale and illumination conditions. • Periodic actions are recognized efficiently. • Not tested for occluded scenarios.
[52]	Pose and velocity vectors from major body parts	Maximum Likelihood Estimation	• Performs well with smooth motion. • Difficult to track occluded body parts.
[66]	Informative gaits extracted with LSTM	Softmax Classifier	• Global context memory cell is introduced in LSTM to improve the attention ability of LSTM.
[34]	3D HOG	SVM	• Tested on multi-view and occluded dataset. • Temporal information is also embedded along with HOG features. • Maximum accuracy with occluded scenarios is 76.7% . • Need to improve accuracy.
[55]	HOG, HOWF	SVM	• Slow computation in high quality videos
[57]	HOG over MHIs	NN	• Multi-view • The actions can be described in a more distinctive way

			by selecting key MHIs.
[56]	HoG, HoF	Matrix Completion	• Multi-view • Not tested for occluded scenarios
[58]	HOG, BoHOF	SVM	• Single view • Tested on single view non-noisy sequences.
[59]	Local STIPs, Global Optical flow	Hybrid BoW	• Multi-view • View-independent dictionary is provided
[67]	HoG, HoF and MBH	BoW model	• Multi-view action recognition • Multiple kernel learning method outperform other fusion techniques. • Tested in controlled multi-view environment without occlusions
[60]	LBP-TOP	BoW , SVM	• Maximum accuracy is 90.72%. • The accuracy can be improved by incorporating alternative texture feature extraction methods.
[68]	Motion context descriptor	RLS-TWSVM	• TLS-TWSVM resolved the outliers and heteroscedastic noise problems • Applicable to only one person and un-occluded data
[69]	Shape and motion features from silhouette and joints	MKL based SVM	• Single view • Accurate joint positions cannot be extracted if the person is occluded.
[70]	Shape, Optical flow	HMM	• Multi-view • Applicable to single person scenario • Experiments are performed in controlled environment.
[71]	Depth key points of human body joints	HMM	• Single view • Feature extraction is performed in controlled environment.
[72]	Dense Trajectories	R-NKTM	• Multi-view • R-NKTM is scaled to incorporate new actions • Only the fixed canonical frontal view is used as target view

Table 2.2 Description of datasets used for individual human action recognition

Dataset	Description
Aerobic [7]	Set of 18 aerobic exercises performed by a single person recorded from seven views.
Ballet Movement	
CMU-mobo [73]	This data set contains only walking action performed by 25 persons on a treadmill in a 3D room.
Hollywood2 [74]	Set of 12 human actions in 69 clips. 150 samples of each action are available in this dataset. The data set is extracted from video clips, recorded with different zoom settings, spatial scales and compression artifects.
IXMAS [15]	Multi-view dataset rerorded with 5 cameras and comprises of 13 action classes: walk, sit down, wave, kick etc. 11 actors performed each action 3 times.
KTH [9]	Single camera dataset comprising 6 different human actions: walking, running, handwaving, handclapping, boxing and jogging recorded in four distinct situations.
MuHAVi [75]	Multiple camera dataset recorded with 8 cameras located on four sides of a platform. Illumination variation due to varying lighting conditions due to night street lights.
OIXMAS [34]	Multi-view dataset containing same actions and actors as in IXMAS datset. The actors are occluded by adding artificial occlusions in video frames.
UoS-HID [76]	A large dataset of gait that is designed in indoor and outdoor environments. The scenes are captured with two cameras and 100 actors.

UCF 11 [77]	Sigle camera dataset containing 11 actions like jumping, juggline, swinging, diving,tennis swing etc. This dataset contains challenging videos having camera motion, illumination variation, cluttered background, scale and viewpoint variations.
UCF sports [78]	Single camera UCF Sports dataset comprises of a collection of actions from different sports typically featured on TV channels such as the BBC and ESPN. It comprises of 10 sports actions.
UT-Tower [79]	Single camera dataset captured from the top of University of Texas comprising 108 videos of low resolution. This dataset contains 9 different types of actions performed by 6 actors each 12 times.
Weizmenn	Single camera dataset comprises of 90 low resolution sequences, 10 actions are performed by 9 different persons.

It is concluded from the above literature that existing individual human action recognition techniques [67] work in controlled environments with only one person present in the video and static background. The illumination conditions are also balance in controlled environments. In multiple view approaches [34], [56], [57], [59], experiments are performed in indoor environments with single person actions. In literature, silhouette-based motion features are discussed [7], [38], [41]–[43], [80] in which extraction of shape and silhouette is done by segmentation. The accuracy of these methods depends upon exact segmentation which is not possible in occluded public environments. Part-based approaches [50]–[53] are based on the information of human body parts, which is very challenging to extract in public environments. Combination of frame-based and temporal features based approaches [34], [55]–[57] are discussed in which [34] is tested on artificially occluded persons but still, there is room for improving the accuracy. These approaches are not evaluated in real-world public environments. Thus there is a need to propose a method to recognize individual human

actions in multiple camera public environments where illumination, view variations and occlusions are major issues.

2.2 Human Behaviour/Interaction Recognition

In previous work, the term behaviour is interpreted as person-to-person interaction [81], multiple human/ group interaction [82] and human-to-object interaction [83]. In this thesis, the term *Behaviour Recognition* is used to indicate: person-to-person interactions and multiple person interactions. Many approaches have been proposed for human behaviour recognition for surveillance purpose. Previously vision based human behaviour recognition algorithms are applied on sports videos, smart homes, public place such as shopping malls, railway stations and airports etc.

2.2.1 Trajectory-based features

Most of the existing work on human behaviour recognition is focused on persons tracking and trajectory evaluation.

Oliver et al. [81] tracked the person's positions in subsequent frames by using Kalman filter and measured position, velocity and orientation for each person. The degree of alignment of two persons and the magnitude of their velocities are used as a feature vector. CHMMs are used to formulate the interactions between two persons.

A bottom-up approach for abnormal human behaviour detection is presented by Antonakaki [84] under multiple camera environment. Two types of features including trajectory features (speed, algebraic mean, mean optical flow and max standard deviation) and motion-based features are extracted for abnormal behaviour recognition. The short-term behaviour abnormality is recognized using SVM and the trajectories are classified using CHMMs. The final decision is made on the basis of the results of both classifiers i.e. the behaviour labelling is reflected by short-term behaviour and trajectory information. Experiments are performed in lab environment where a single person performs different activities inform of cameras. This method is tested only on single person behaviour and interactions between persons are not considered. Performance of the proposed method is decreased when tested on CAVIAR dataset because this dataset is captured with single camera which does not encode enough information for abnormality detection.

Geometrical and motion visual features are extracted in [85] to analyse single person behaviour, person-to-person and multiple person interactions at railway station. Trajectories

are extracted by using a S-T MRF model and state transition information is created from geometrical and motion (location, pose and attributes) features. Experiments are performed on videos captured in railway station. The problem with this approach is that parameters for feature extraction are set beforehand that is not applicable to real scenarios. Furthermore, contextual information is necessary to differentiate between normal and abnormal behaviours.

Suk et al. [10] analysed interactions between two persons (person-to-person) on the basis of walking trajectories of persons. A feature vector comprised initial positions, moving directions and distance between both persons. They divided an interaction into small units called sub-interactions. The Dynamic Probabilistic Models (DPMs) are utilized for modelling of sub-interactions and a complete interaction is represented with the network of DPMs (NDPMs). Experiments are performed on four different datasets. The problem with this method is that it can only recognize the predefined interactions and NDPMs require prior knowledge about the structure of interaction.

Calderara et al. [86] detected anomalies in people's trajectories under single and multiple camera views by representing trajectories in space as a sequence of transitions among nodes on graph. On a graph, the shared distinct trajectories represent simply a small subspace of all possible trajectories. This small subspace is categorized by dominant connected components of graph. The graph is then projected on low frequency eigenvectors and anomaly detection is performed using divergence measure which is defined by canonical angle among subspaces. Experiments are done on trajectory data captured with two cameras in university campus.

Chen and Aghajan [82] presented a method to fuse information from multiple cameras for social behaviour analysis in work environment. They focused on localizing human, estimation of head pose and interaction detection. In this method, only the head is tracked instead of full body tracking. Feature vector includes relative distance between people and their relative head angle. The classification of social interactions is performed using SVM classifier under all camera views separately. Fusion is done at interaction decision level by considering the estimations of all cameras and their relevant confidence score. Experiments are executed on multi-view dataset recorded in an office environment with 0.79 precision. The fusion method has achieved comparable performance though it requires very little information to be shared among all camera views. The role of an individual person is not considered while analysing the interactions. Accuracy can be improved by considering an individual person's actions.

Trajectory data is also used for detecting group interactions by Chen and Cavallaro [87]. They extracted the trajectories from sampled frames and trajectory at a specific frame is denoted by the availability of trajectory, its position and velocity. Each object in the scene is modelled as a moving agent and group interactions as collective interests between objects. Temporal association problem is solved by tracking the group interactions over consecutive video frames. The mutual influence between interacting objects is modelled by defining motion direction aware interest map. Experiments are performed on JAIST dataset and APIDIS basketball dataset and group interactions are detected with 80% accuracy in both datasets.

Lin et al. [35] proposed network transmission based algorithm for detection of abnormal human activities group behaviours. They divided the scene into patches and movement of an individual from one patch to any other patch is modelled as package transmission process in network. Abnormal trajectories are detected by calculating network transmission energies consumed to transmit a package.

Recently, Ouyed and Said Allili [88] represented interaction with features extracted from the motion of human body joints. To get the trajectories of all joints, each joint is tracked over video frames. For each trajectory, a group of features is defined and the interaction is represented by concatenating the group features. Group feature weighting is incorporated in kernel logistic regression for interaction classification. Experiments are performed on UT-interaction dataset with 95% average accuracy. Accuracy is reduced when applied on UT-interaction Set II that has slightly moving background. This method may fail in occluded environments with an increased number of persons.

Shape context and trajectories are also used for the recognition of collective activities in a frame by considering pose and individual actions of each person in a single camera view [89]–[91]. Poses of individual persons and atomic actions [91] are incorporated for interaction recognition and further this information is used for group activities recognition.

The trajectory-based features for human behaviour recognition are very common. The accuracy of behaviour recognition methods depends upon the information extracted from trajectory data. Individual person’s activities must be monitored to deal with person-to-person or group interactions.

2.2.2 Body motion and part-based motion features

Full body motion and part-based motion descriptors are extracted around spatio-temporal interest points and used for behaviour recognition.

Kong and Fu [92] recognized close interactions by extracting motion descriptors from interactive regions. Spatio-temporal interest points are detected and motion around interest points is described by utilizing Gradient descriptors. Human bounding box is split into non-overlapping spatio-temporal patches and histogram of video words is used as patch descriptor. A patch-aware latent SVM is proposed to formulize the interactions between close persons. Experiments are performed on BIT interaction and UT-interaction datasets which are single view datasets with an accuracy of 85.38% and 93.33% respectively. The proposed method is also tested to recognize individual actions. Problem with this method is that the proposed features are not able to discriminate the occluded interactions and the visually similar interactions are also misclassified.

Ji et al. [20] represented the interaction with local and global characteristics. Local characteristics are represented with improved BoW descriptor of STIPs and global characteristics with HOG descriptor. A frame-by-frame NN classifier is applied on both descriptors separately and voting histograms are obtained. The final recognition result is attained by applying weighted fusion on voting histograms of both descriptors. Experiments are performed on UT-interaction dataset with average accuracy of 83.3%. This method is simple to recognize interactions but it is not tested on multiple views with varying illuminations and occlusions.

Ahmed and Yousaf [93] proposed to recognize human interaction in challenging partially occluded and noisy environments. MHIs are extracted from input sequences and then HOG and histogram of oriented energy (HOE) features are extracted from MHI templates. Codebook is then constructed from extracted features and linear boosting SVM is applied for classification. Experiments are performed on UT-Interaction and YouTube datasets with 93% and 91.6% accuracy respectively.

Motiian et al. [21] proposed a real time system for the analysis of human behaviour. Motion (HOOF and motion histogram of each individual), proximity (by computing distance between people trajectories) and audio features (mel frequency cepstral coefficients) are combined to form person-to-person interaction trajectories. The temporal sequence is modelled by using a kernel-state space (KSS) model and pairwise kernels with special symmetry are designed.

They validated their approach on four publically available datasets: UT-Interaction dataset, TV Human Interaction dataset, BIT-Interaction dataset, SBU Kinect-Interaction dataset, and two self-created datasets: HAUS-PI (single camera) and MVHAUS-PI (multi-view) datasets. This method is tested only on two person's scenarios and the multiple view dataset is captured in a controlled indoor environment.

Murthy et al. [17] investigated the effect of fusing (early fusion) human body parts based representation with local information (Harris 3D points) and also with densely sampled trajectories for human behaviour recognition. BoW and SVM based approach is used for classification of fused features. Separate classifiers are also learnt on each type of descriptor and late fusion is performed on classifier scores. Experiments are performed on UCF50 and HMDB51 datasets and the best results are attained upon fusing trajectory representation with part-based representation.

Later Huynh-The et al. [36] recognized person-to-person interactions by extracting features from human pose estimation. Spatio-temporal relation features are extracted from the articulated pose coordinates, which consists of intra and inter-person features extracted from distance and angle of joints. A codebook is constructed from the joint coordinates of human body and the correlation between codewords is described using topic modelling. Multi class SVM is used to classify the interactive activity. Experiments are performed on BIT-Interaction and UT-Interaction datasets. They studied the effect of different features on recognition accuracy and reported that merging joint distance and angle features acquire the best accuracy of 91% when compared with other features.

Problem with the part-based features is that enough information cannot be extracted for behaviour recognition if the persons are occluded with other persons or objects.

2.2.3 Audio-visual features

Some researchers used audio features along with visual inputs to overcome the limitations of descriptors under multiple camera environments. A combination of audio and visual features is used by Brdiczka et al. [94] for human behaviour detection in smart homes. Taj and Cavallaro [19] estimated object movement in scenes that are not covered by camera field of view by taking input from microphones. Trajectories are estimated by using audio and visual features. Following features are extracted to recognize the interaction between two people: relative direction, relative distance and its derivative and magnitude of velocity of each person.

Though audio visual features proved to be the promising descriptors for interaction recognition, these features are not suitable in noisy environments.

2.2.4 Contextual features

Contextual information is modelled by many researchers for human behaviour recognition [22], [89], [95]. Contextual information of nearby persons is analysed to recognize collective activities, like standing in queue, talking etc.

Choi et al. [89] proposed a new spatio-temporal local (STL) descriptor to capture the spatio-temporal dissemination of nearby persons . The STL descriptor is centred on focal person and histogram of nearby persons, their pose and movement is computed. Since there are many persons in a scene, a collection of STL descriptors is gathered in each frame. Collective activities are classified by using SVM classifier.

Lan et al. [22] also extracted contextual information and proposed action context (AC) descriptor which is the concatenation of focal person action descriptor and nearby person action descriptor. Contextual group activities are recognized by exploring the contextual information in terms of latent variables. The proposed latent structure is capable to jointly model group activities and individual person actions. Experiments are performed on collective activity dataset and nursing home dataset.

Zhu et al. [95] exploited contextual information for detection of abnormal activities. The activities inside a spatio-temporal threshold are grouped together and considered as associated with each other. Related activities are jointly modelled by extracting the motion and context features. Following descriptors are extracted: intra activity motion and context feature, inter activity context feature, spatial context and temporal context. SVM classifier is trained on motion and context descriptors for activity recognition and anomaly detection. Results published on above mentioned approaches demonstrated that combining features with contextual information can provide a significant improvement in collective behaviour recognition accuracy.

Table 2.3 Characteristics of human behaviour recognition techniques

Ref	Features	Classification	Comment
[81]	Derivative of relative distances, magnitude of velocities	CHMM	• Small unit interactions are recognized by dividing interactions into sub-interaction patterns. • HMMs and CHMMs are compared and CHMMs found to be more superior than HMMs.
[82]	Distance between people and their relative head poses	SVM	• Multi-view social interactions are recognized in office environment. • Individual person's role is not considered.
[19]	Relative distance, relative direction, and velocity	CHMM-MAP	• Small unit interactions are recognized from people's trajectories • Audio and visual features are used for trajectory estimation. • Tested on sports dataset
[20]	BoW descriptor of STIPs and HOG	NN classifier	• Efficient in recognition interactions • Not tested on multiple view having occlusions and varying illumination
[21]	Motion, proximity and audio	KSS model	• Tested on single and multi-view datasets • Not tested for occlusions • Multi-view dataset is captured in indoor environment.
[85]	Pose, positioning and multiple object interaction feature	STI	• Single pedestrian and interaction between pedestrians is analysed • Parameters for feature extraction are set beforehand that is not applicable to real scenarios.
[10]	Distance, angle and motion direction	Network of Dynamic probabilistic models	• Five simple interactions are recognized by dividing interactions into sub-interactions. • Need more training to evaluate in real scenarios • NDPM involves a prior knowledge about the structure of interactions
[86]	Trajectories are	Laplacian	• Single view and multi-view

	represented with graphs	filtering of Graph	• Detects anomalies in people's trajectories
[35]	DT energy values	SVM	• Single view • Performs online detections. • Cannot differentiate normal activity pattern once the abnormal activity is detected.
[92]	Gradient features	Latent SVM	• Single view • Unable to handle occlusions • Visually similar interactions are misclassified.
[36]	Distance and angle extracted from joints	SVM	• Single view • Unable to extract joint features when the person is fully occluded.
[94]	Sound, posture, speed and distance	SVM	• Single view. • Difficulty in identifying focus of attention for each person.
[89]	STL descriptor	Markov Chain Model	• Collective activity in single view. • Robust to view point, illumination variations and cluttered background
[95]	Motion and context features	SFG+Context model	• Spatio-temporal relationships within and across the activities are captured successfully. • Unable to recognize multiple activities in a frame
[93]	HOG, HOE	SVM with linear boosting	• Background noise is successfully reduced using MHI tempaletes • Punch and push actions are highly misclassified due to inter class resemblance.
[91]	Randomized Spatio-temporal volume	SVM	• Collective activities are recognized in single view • HOG is used for individual action recognition • Occlusion handling is not performed explicitly

Table 2.4 Description of datasets used in human behaviour recognition

Dataset	Description
APIDIS basketball [96]	Basketball dataset recorded with 7 cameras around and top of a basketball court.
BIT interaction [97]	Single camera view dataset comprises of 50 videos of each interaction class. Total eight interaction classes are recorded with large variations of background, viewpoints, illumination conditions and scale and appearance.
CAVIAR [98]	Single camera view dataset recorded in a public place with 6 different scenarios: window shopping, meeting with others, fighting, entering shops, exiting shops etc.
Collective activity [89]	Single camera view dataset comprises 44 short videos of 5 distinct collective activities: queueing, talking, walking, crossing, waiting recorded to analyse collective behaviour of persons.
HMDB51 [99]	Single camera view dataset created by collecting movies clips from Youtube and Google videos. This dataset comprises 51 different action types with each types containing 101 clips.
JAIST [100]	Multiple camera view dataset captured with 8 cameras in a lab environment. 8 distinct types of actions are recorded by single person. It also contains actions performed by multiple persons in a group.
MVHAUS-PI (multi-view) [21]	Multiple camera view dataset captured in an indoor environment. It comprises the

	sequences of 16 interaction classes between two persons
Multiple Camera Offcie Environment	
Nursing Home Dataset [22]	Single camera view dataset recorded with a fish eye camera in a nursing home dining room. It include actions like walking, sitting, standing and bending.
SBU Kinect-Interaction dataset [101]	Single camera view dataset comprising 8 interaction classes recorded by 7 actors. Depth images and coordinates of joints are also provided in this dataset.
TV human interaction dataset [102]	Single camera view dataset collected from 20 distinct tv shows and comprises 4 types of interactions.
UT-Interaction [27]	Single camera view dataset comprising videos of 6 interaction classes containing cluttered background and multiple persons.

Previous work on interaction recognition revealed that the trajectory-based features have proved to achieve performance gains but these methods have ignored the role of individual person during interaction recognition. Individual person action recognition is significant for recognition of small unit sub interactions. Individual actions, poses and interactions are exploited for collective activity recognition in [22], [89], [91]. These approaches seem to be efficient for collective activity recognition but still, these techniques are not tested on multiple views and occlusions. Literature revealed that not much work is done on the recognition of person-to-person small unit interactions. Previously, experiments have been performed on single view datasets and less attention is given to sub interactions in multiple camera scenarios.

2.3 Human Behaviour/Interaction Anticipation

Human behaviour anticipation has gain importance in the last few years due to the growing demand for automated surveillance systems in smart homes, public areas and human-computer interaction etc. The field of human behaviour recognition is now gradually moving

towards the anticipation of single person and multiple person behaviours [3]. This section presents mainly used feature representation and classification methods for the anticipation of human activities and interactions.

2.3.1 Handcrafted Features

STIPs, HOGs, HOFs are some common handcrafted features used for human behaviour anticipation.

Ryoo [3] extracted 3D spatio-temporal local features and represented with visual words. Activity prediction problem is formulated probabilistically and activity is represented with an integral histogram of spatio-temporal features. Two extensions of BoVW model are proposed: Integral BoVW and Dynamic BoVW for dynamically encoding the ongoing human activities. Experiments are performed on UT-Interaction dataset and 70% accuracy is achieved with Dynamic BoVW and 65.0% with Integral BoVW.

Sun et al. [24] used body parts movements to represent ongoing human activities. They extracted dense STIPs as low level features and scale adaptive mean shift method is used to locate sparse grouplets. A recurrent self-organizing map trajectory (RSOM) is proposed where STIPs are mapped on RSOM network. Human activities are represented by using extracted RSOM trajectory. Prediction is performed by combining DTW distance and edit distance i.e. DTW-E to measure the difference between RSOM trajectories. Experiments are performed on Rochester dataset, UT-Interaction dataset and DARPA dataset. This method achieved highest accuracy of 100% on UT-Interaction dataset. However, the accuracy is decreased when tested on a cluttered background.

Wang et al [31] proposed human activity prediction method. They firstly divide activity video into short segments and each segment is represented with HOG and HOF which are extracted around local spatio-temporal interest points. These features are then represented in BoW model. To compare the segments of different lengths, a temporally weighted generalized time wrapping (TGTW) algorithm is proposed to perform time series alignment of activity segments. After obtaining the alignment similarities, k-nearest neighbour (KNN) is used to predict activity class. Experiments are performed on UT-Interaction dataset, DARPA-Y1 dataset and UCF support dataset.

Barnachon [32] introduced histogram based representation of 3D motion capture data for ongoing action recognition. An extension of classical to integral histograms is presented to

control the lack of temporal information. The comparison of histograms is performed by using Bhattacharya distance. This method is tested on datasets where a single person is performing different actions in front of the camera.

Existing approaches on interaction anticipation mainly used spatio-temporal features and motion features. These handcrafted features lose global structure in the data hence these methods alone are not able to capture significant motion information for interaction anticipation [103].

2.3.2 Deep Feature Representation

STIPs, HOGs, Optical flow, Trajectory features etc. are the handcrafted feature representations, which have their own limitations. Deep learning approaches are now commonly employed in different classification and prediction tasks [104], [105]. Combination of deep and handcrafted features is used by Majtner et al. [106] for skin lesion classification and Wu et al. [104] for person re-identification.

Chen et al. [107] used unsupervised feature learning approach and proposed space-time deep belief network (DBN) for single person action recognition. It builds invariant features from spatio-temporal data by convolving restricted boltzmann machines (RBMs) together with spatial and temporal pooling layer.

Choi et al. [25] used multiple RBMs for unsupervised feature extraction to predict human behaviours in smart homes. They have proposed two prediction algorithms, DBN-R and DBN-ANN, and compared the results with DBN-SVM. Experiments are performed on MIT home dataset for prediction of activities in smart homes.

Ke et al. [23] presented an human interaction prediction method by considering temporal information. Optical flow is extracted from the input video frames. They presented low level optical flow coding images to the Deep Convolutional Network for deep temporal feature extraction. The deep features extracted from each frame are concatenated using temporal convolution. A Softmax activation function is applied to classify interactions from partially observed videos. Experiments are performed on UT-interaction and TV Human Interaction datasets with average an accuracy of 88.3% and 69% respectively.

Dutta and Zielinska [108] presented a probabilistic method by considering object affordance for human-object interaction prediction. Features are extracted at three levels: low level

(HOG, dense trajectory), mid level (onset, actionlet and poselet) and high level (CNN). The extracted features are then passed to probabilistic model for prediction of actions. Experiments are performed on WUT-ZTMiR and CAD-60 datasets.

Deep learning features proved to be more reliable in behaviour prediction tasks. Deep features are extracted on multiple layers, hence able to extract more dense information from input images.

Table 2.5: Characteristics of Human Behaviour Anticipation methods

Ref.	Features	Classification	Comments
[3]	Integral histogram of space time features	Dynamic bag of words	• Single view • Handcrafted features • Tested in controlled environment
[24]	Sparse grouplets to represent movement of body parts	DTW-E	• Single view with one person • High accuracy when compared with simple BoW method • Handcrafted features • Unable to detect interest points if body parts are occluded
[31]	Space time features	KNN	• Single view • TGTW is proposed for time series alignment • Tested on single person activities and interactions • Handcrafted features
[32]	MoCap data	DTW	• Single view • Actions are represented with histogram of motion capture data • Applicable on single person actions
[109]	Trajectory features	Hidden variable MDP	• The interplay between features and environment is focused
[110]	ADV	Various classifiers i.e. SOM, SSOM, SGAS	• Need to generalize the model

[25]	Automatic feature extraction using DBN-R	DBN-R	• Behaviour prediction in smart homes • Prediction is based on learning human intentions • It should be extended to use more realistic data
[111]	Local spatio-temporal features	CSR	• Performance is improved as compared to [3]. • Handcrafted features
[112]	Features automatically learning with deep learning model	Deep Network	• Model contains feature layer and action response layer • Designed for human players • Can be extended beyond two players
[23]	Used pre-trained CNN to extract features from Flow coding images	Softmax	• Deep temporal architecture is presented by learning features from flow coding images • Original RGB images can provide a useful information for interaction observation
[108]	HOG, dense trajectory, onset, actionlet, poselet, CNN features	Probabilistic model	• Human-object interaction is predicted considering object affordance • RGB depth data set is used for experiments • Distance and angular preferences are used to find object affordance • Feature extraction is relatively complex.

Literature review of previous approaches for interaction anticipation reveals that there is very little work done on interaction anticipation. Space-time features and temporal features are commonly used handcrafted features for anticipation. In recent years, deep learning modules are also used for feature extraction and also for the anticipation. UT-Interaction is the popular single view dataset is being used for interaction recognition and anticipation. Multiple

camera views can be used to improve the performance of interaction anticipation. To the best of our knowledge, interaction anticipation in multiple camera scenarios is not explored in previous studies.

Table 2.6 Description of dataset used in human actions/interaction prediction

Dataset	Description
DARPA Y1 [113]	Single camera view dataset comprising videos of 7 interaction classes with variation in actor size, illumination and background.
MIT home dataset [114]	This dataset is recorded from two appartments for two weeks. Total 164 sensors were placed in both appartments. These sensors were placed on different objects like refrigrators, drawers, switches etc.
UCF sports dataset [115]	This dataset contains set of sports actions collected from various sports shows.

2.4 Research Limitations

In this chapter, the existing techniques for *individual human action recognition*, *human behaviour/interaction recognition* and *human interaction anticipation* are discussed. Existing methods have shown promising results in various activity recognition and anticipation tasks. However, there are several open issues and problems that need to be solved.

Existing methods for the recognition of single person actions work in controlled environments with only one person present in the video and static background [34], [46], [55]–[57], [59]. Some methods are robust to handle partial occlusions and some are robust in handling illumination variations. Hence those systems are not handling partial occlusions and illumination variations in multiple camera environments at the same time in real-world videos.

Existing research is mainly focused on collective behaviour of multiple persons under single camera [22], [87], [89], [90], [116]. There is a need to analyse small unit person-to-person interactions in public environments to fulfil the requirements of automatic video analysis applications. Social interactions analysis becomes challenging if the persons are occluded.

The anticipation of high-level person-to-person interactions is necessary for many surveillance applications. Anticipation is concerned with future certainty of ongoing human behaviours. To the best of our knowledge, human behaviour anticipation under multiple camera environments has not addressed yet. Interaction anticipation in multiple camera view outdoor scenarios is a challenging problem due to the presence of illumination variations, partial occlusions and cluttered background.

The focus of this research is to improve recognition and anticipation accuracy in multiple camera view scenarios in indoor and outdoor environments.

2.5 Problem Formulation

Let $C = \{C_1 \dots C_M\}$ be the set of M cameras and $V = \{V^{1,N}\}$ be a video sequence of length N frames. We denote the set of persons as $P = \{P_i^k : i = 1 \dots m\}$ where m is the number of persons detected in a frame k, P_i^k is the i^{th} person or individual in k^{th} frame. Let $T = \{T_t^m : t = 1 \dots L\}$ be the set of the trajectories of m persons with a length of L. Ideally, $N = L$ only if the same person is seen in all frames of the video V. After extracting this information, the objective is to recognize $P_i^k \rightarrow P_j^k$ in V where the symbol $\rightarrow$ represents interaction (person-to-person) between person i and person j. This problem is broken into two steps.

1. Let the function for feature extraction of individual persons be denoted by ψ and defined as $\{O_i^k\}_{i=1}^m = \psi\{p_i^k\}_{i=1}^m$. Here, ψ represents two feature extraction function i.e. HOG and MDCLBP and O is the extracted feature vector. For the recognition of individual actions in a frame, a classification function $g(.)$ is applied on extracted features as $g: \{O_i^k\}_{i=1}^m \rightarrow \{l_i^k\}_{i=1}^m$ which returns the discrete values $\{l_i^k\}_{i=1}^m$, consists of individual action labels predicted by $g(.)$. Fusion is applied on classification results of all cameras to get the final decision.

2. To recognize person-to-person interaction i.e. $P_i^k \rightarrow P_j^k$, predicted action labels and the trajectory features of both persons are considered. If P_j^k be the j^{th} person detected nearest to P_i^k where $1 \leq j \leq m$, we have incorporated $\{l_i^k\}_{i=1}^m$ as contextual information with trajectory features; the final feature vector is represented by FI_{ij}^k.

Let Ω be the feature extraction function for human behaviour anticipation and ROI_{ij}^{k} be the region of interest (ROI) detected by drawing bounding box around both persons. Ω is defined as $FA_{ij}^{k} = \Omega(ROI_{ij}^{k})$. Here FA_{ij}^{k} is feature vector extracted from ROI_{ij}^{k} for anticipation.

2.6 Summary

In this chapter, an overview of previous work in the area of individual human action recognition, human behaviour recognition and human behaviour anticipation is discussed. Extensive literature is available on individual human action recognition from videos having single view and single actor; very little attention is given to public places where occlusions and illuminations are very common. Single person action recognition approaches are mainly validated on simple datasets with acceptable accuracy. The large quantity of dataset having single person actions is also widely available. The individual action recognition problem becomes more complex in outdoor public places. Recognition of small unit interactions between people in public scenarios is very important in automatic human behaviour analysis. Much work has been done on interaction recognition but there is still room for improvement. Unlike individual action recognition and interaction recognition, human interaction anticipation is also essential for surveillance applications.

Chapter 3

Human Action Recognition in Multiple Camera Environments using HOG-MDCLBP as a New Descriptor

3.1 Introduction

Recognising single person actions in multiple camera environments has always been a challenging task due to certain obscurities i.e. pose and illumination variations and occlusions. In particular, the presence of occlusions in a frame increases uncertainties in understanding the actions of a single person. Many action recognition techniques have been proposed, some of which extracted HOGs and combined it with other features to achieve good accuracy. Literature in Chapter 2 revealed that HOGs alone cannot handle the challenges of outdoor multiple camera scenarios.

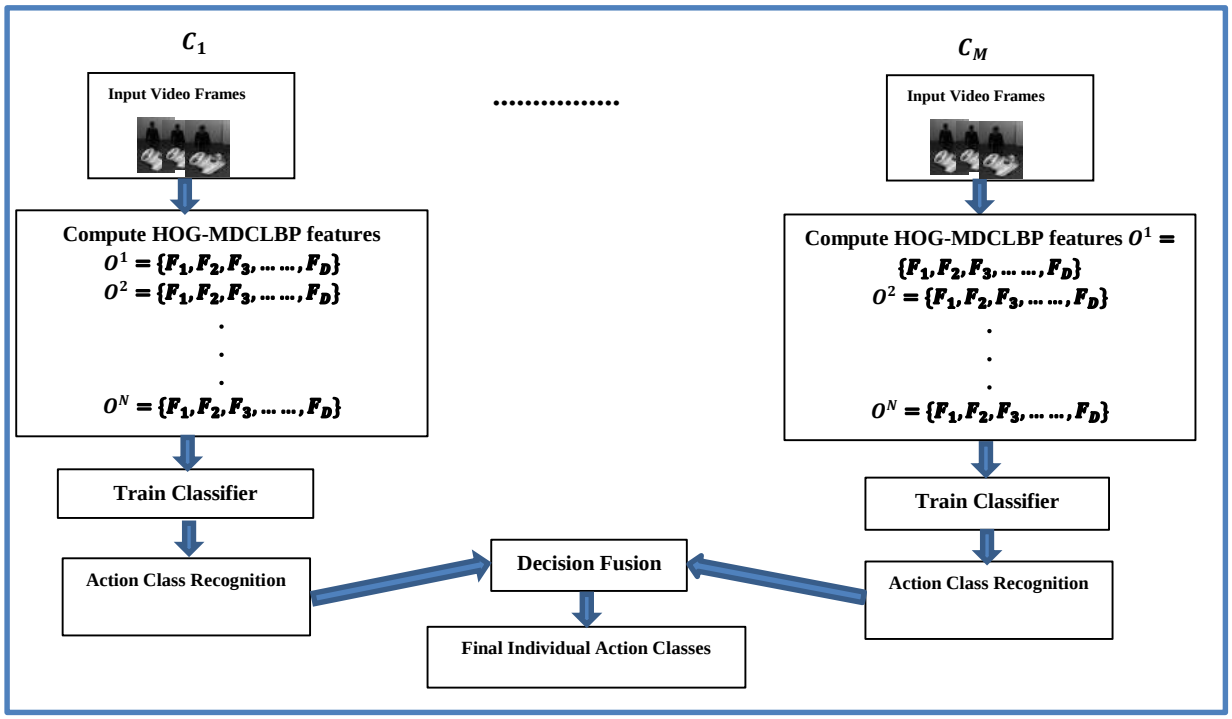

Figure 3.1: A Block diagram illustrating individual person action recognition approach.

In this chapter, a vision based technique has been proposed for human action recognition in multiple camera environments. The process of action recognition is depicted in Figure 3.1.This is a crossbreed approach that takes advantage of two types of appearance features,

HOG and MDCLBP. HOG provides the information about the appearance of gradient distributions in a local patch. In addition, the MDCLBP numerates the arrangements of intensities in a frame to extract local texture features. HOG and MDCLBP features have been synthesized to represent an individual person's actions. Training of the extracted feature is performed using SVM classifier and then experiments are performed on multi-view dataset and compared presented approach with some state-of-the-art approaches.

This chapter is organized as follows: Section 3.2 discusses the motivation of developing HOG-MDCLBP descriptor. In Section 3.3, feature extraction method for individual action recognition is described. Section 3.4 is about experimental setup and results.

3.2 Motivation

This chapter presents the feature extraction method for individual human action recognition. It is proposed to represent an action with the combination of gradient and texture features. MDCLBP is the proposed technique for texture feature extraction that is the variation of compound local binary pattern (CLBP) [117] technique. CLBP uses average value in a small window as a threshold. The average value thresholding discards several important pixels information. Median value on the other hand has been proved to show very good discriminatory properties [118]. This thesis proposed to use median value within 3×3 window as threshold assuming that median value is not as much affected by noise and variations. Local texture information extracted using the median value threshold has eliminated the effects of partial occlusions. HOG on the other hand provides illumination invariant representation. When MDCLBP is combined with HOG, it resolved the issues of partial occlusions and illumination variations occur due to the multiple person interactions and multiple camera views.

3.3 Feature Extraction for Individual Human Action Recognition

This section describes the action representation method for individual human action recognition. Action is represented by combining two types of representations to achieve robust results.

3.3.1 Histogram of Oriented Gradient (HOG)

HOG is an appearance based feature descriptor that extracts the distribution of gradient directions. It was proposed by Dalal and Triggs [54] for human detection. HOG has been shown to be rather efficacious for human detection. Later, HOG along with other descriptors

has been used for action recognition by many researchers [23] – [25]. Given a person's detected window, HOGs are extracted given as:

$$h_i^k = HOG(P_i^k) \tag{3.1}$$

Here $HOG(.)$ is the HOG feature extraction function. After extracting HOG descriptors from a detected person window P_i^k, Histogram of HOG feature descriptors $\widehat{H}_i^k$ is computed. The steps to compute HOG feature descriptors are as follows:

1. Gradient image calculation

Horizontal and vertical gradients of a detected person's window are computed by convolving P_i^k with kernels (Equation 3.2) along horizontal and vertical directions.

$$gx = [-1\ 0\ 1] \text{ and } gy = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \tag{3.2}$$

The horizontal and vertical gradient images are represented by Px and Py .

2. Magnitude and Orientation calculation

The magnitude G and the direction θ of gradient are calculated as:

$$G = \sqrt{gx^2 + gy^2} \tag{3.3}$$

$$\theta = \tan^{-1}(gx/gy) \tag{3.4}$$

3. Orientation Binning

Orientation binning is performed by dividing the input image (person's window) into small cells of size (8×8) and computing a local 1D histogram of oriented gradients of each cell. The 8×8 cell contains 192 pixels i.e. $8 \times 8 \times 3$. Binning is done by using 9 orientation bins spaced over $[0^\circ - 180^\circ]$, ignoring gradient sign. The corresponding bin for each pixel under 8×8 region is decided by looking at orientation θ and vote is selected on the basis of magnitude G.

4. Block Normalization and final descriptor calculation

Block normalization is performed in order to make the descriptor independent of global illumination variations. Normalization is performed over 16×16 region in image to get

36×1 normalized vector. Same process is repeated for entire image region by choosing a block of size 16×16 and normalizing that block repeatedly. HOG feature descriptor is obtained by combining all normalized vectors. Final feature vector is represented with normalized 9-bin histogram as the obtained HOG descriptor size is very large i.e. 1×10296 for the image size of $97 \times 219 \times 3$. Also, the person size in videos captured from different views can also vary, so all descriptors are represented with 9-bin histograms i.e. $\widehat{H}_i^k$. HOG descriptor along with histogram representation is shown in Figure 3.2.

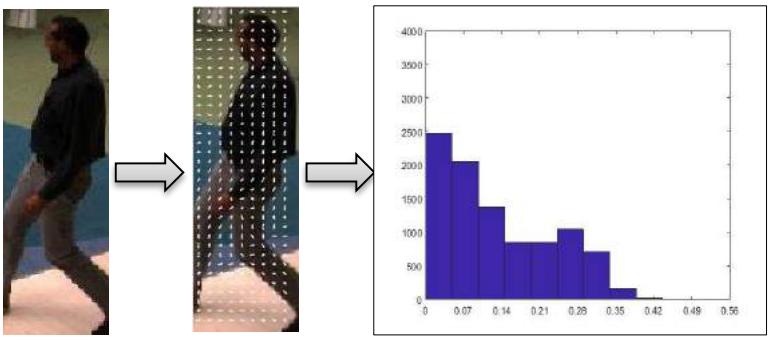

Figure 3.2: Extraction of HOG descriptors. (a) Original Image. (b) HOG descriptor visualization. (c) 9- bin HOG histogram

The effect of illumination variation on HOG features is shown in Figure 3.3. Global illumination change is applied on original image and the histograms of both images are also displayed. The Euclidean distance between original and illumination changed HOG histogram is 0.04 which demonstrates that the similarity between both histograms is high because the distance is closer to zero.

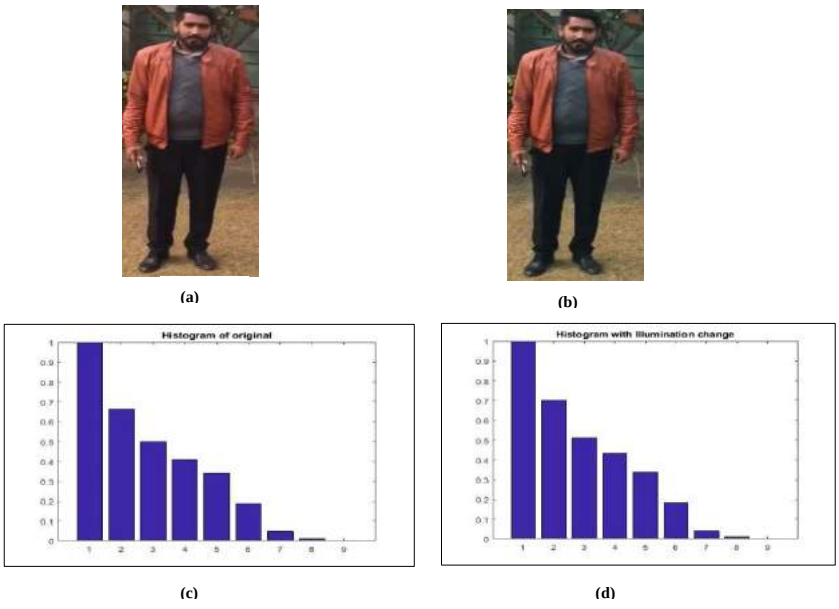

Figure 3.3: Example of illumination variation and its effect on HOG features. (a) Original Image. (b) Image with illumination change. (c) Histogram of HOGs of original image. (d) Histogram of HOGs of image after illumination change. The Euclidean distance between both histograms is 0.04.

3.3.2 Median Compound Local Binary Pattern (MDCLBP)

MDCLBP is a texture operator, it is 16 bit encoding scheme which combines MDCLBP1 and MDCLBP2 for the extraction of texture features. The MDCLBP is defined as:

$$\hat{C}_i^k = \Omega(P_i^k) \tag{3.5}$$

MDCLBP is a variation of CLBP that has originally been proposed in [117] as an extension of LBP [119] for rotationally invariant classification of texture. Unlike LBP, that returns 8-bit output code, MDCLBP operator returns 16-bit code: two bits for each neighbourhood. As in LBP and CLBP, the least significant bit is the sign difference. Moreover, In LBP, the central pixel is used as threshold and difference of neighbouring pixels to the central pixels is calculated to generate 8-bit code. CLBP utilizes magnitude information as well as sign information to generate appropriate binary operator. In the proposed MDCLBP, modification is performed on the second bit that is most significant bit. CLBP utilized average value as a threshold under 3×3 window but we have modified the second bit by selecting median value as threshold instead of using average value. Median value is selected as it is less affected by outer affects. Moreover, the neighbourhood size is chosen 3×3 as in original CLBP operator to reduce the number of features [117].

To calculate MDCLBP, first of all a 3×3 region is selected from P_i^k and to generate 2-bit code for each neighbouring value sign difference and median value are used. The central pixel value from selected region is compared with neighbouring values (as in LBP). The bit is set to 1 if the sign difference is positive. To compute second bit, the difference of median value with its 8 neighbouring pixels is calculated and the bit is set to 1 if the value of neighbouring pixel is greater than the median value, otherwise 0. Resulting 16 bit code is converted into decimal which results into very high value. For this reason, two sub codes are generated: MDCLBP1 and MDCLBP2. Same process is repetitively applied on entire image by choosing a window of size 3×3 at each iteration, resulting in two MDCLBP images. Mathematically,

Let ∂_1 be the least significant bit and ∂_2 be the most significant bit, computed by applying the MDCLBP operator.

For ∂_1

$$\partial_1(x_c, y_c) = \sum_{c=0}^{7} s(n_b - n_c)2^c \quad (3.6)$$

$$s(X) = \begin{cases} 1 & X \geq 0 \\ 0 & X < 0 \end{cases}, \quad (3.7)$$

Here, n_b is the value of neighbouring pixel and n_c is the intensity value of centre pixel (x_c, y_c) and c is the number of neighbours around (x_c, y_c).

For ∂_2, firstly, the median value is calculated from selected $3x3$ window and denoted as $\tilde{m}$, then ∂_2 is calculate as follows:

$$\partial_2(x_c, y_c) = \sum_{c=0}^{7} s(n_b - \tilde{m})2^c \quad (3.8)$$

Here, $\tilde{m}$ is median value calculated from 3×3 region and selected as threshold. The most significant bits code ∂_2 is generated by comparing threshold value with its associative neighbouring pixel values. The rule is same as used to calculate ∂_1, i.e. If $\tilde{m}$ is less than neighbouring value it attains 1 otherwise 0. This process is applied on entire image and the resulting ∂_1 and ∂_2 values are combined to create 16 bit code. Figure 3.4 shows the illustration of MDCLBP computation process; 142 is the centre pixel value to compute least significant bit and 165 is the median value to compute most significant bit.

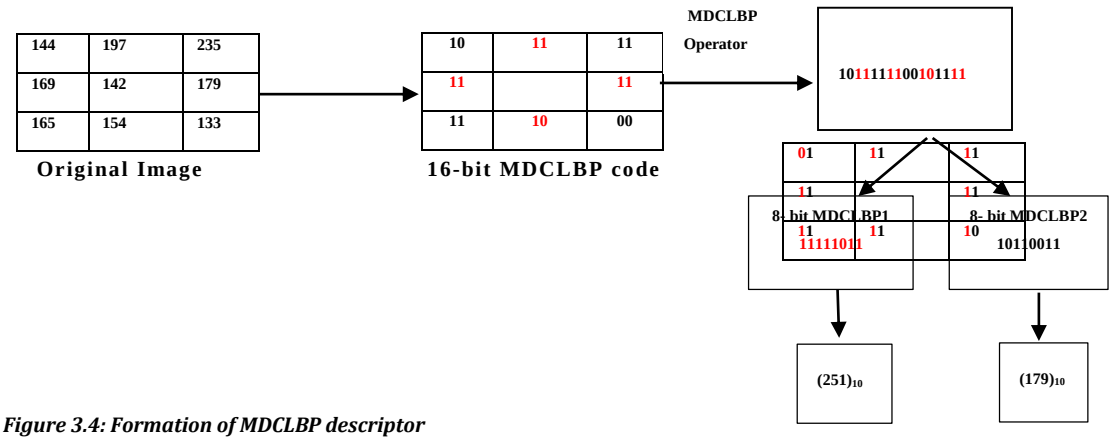

Figure 3.4: Formation of MDCLBP descriptor

A 16-bit MDCLBP code is computed by using Equations (3.6), (3.7) and (3.8); the code generated at middle pixel is discarded to keep the resulting values in the range i.e. 0-255. The 8-bit MDCLBP1 is obtained by combining neighbours at north, east, west and south positions. The other 8-bit MDCLBP2 is obtained by combining the neighbours at diagonal positions of the centre pixel. Each of the MDCLBP codes is represented by a normalized 256-

bin histogram and finally, the histograms are concatenated to get final texture feature descriptor $\hat{C}_i^k$ of size 1x512. The illustration of MDCLBP is depicted in Figure 3.5.

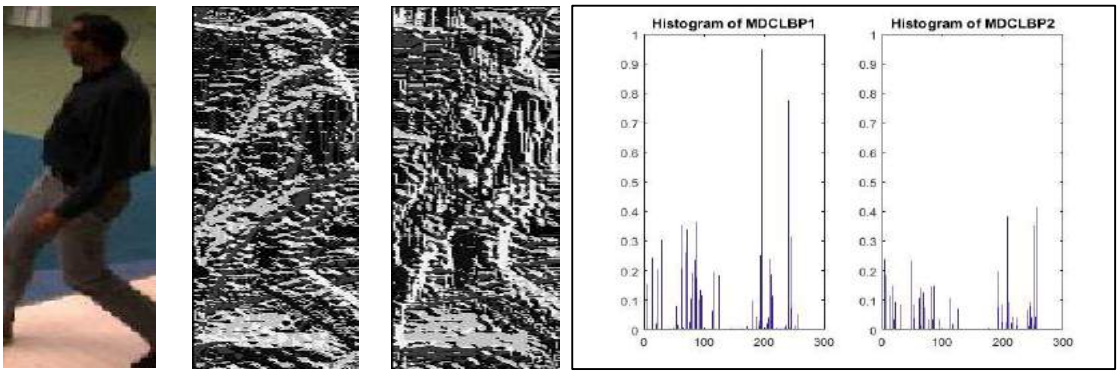

Figure 3.5: Illustration of MDCLBP descriptor. (a) Original Image. (b) MDCLBP1. (c) MDCLBP2. (d) Normalized histograms of MDCLBP1 and MDCLBP2

HOG descriptor $\hat{H}_i^k$(1x9) and MDCLBP descriptor $\hat{C}_i^k$(1x512) are normalized separately and then concatenated using Equation (3.9) for getting a feature descriptor robust to partial occlusions and illumination variations.

$$O_i^k = [\hat{H}_i^k, \hat{C}_i^k] \tag{3.9}$$

Here, O_i^k is the final feature descriptor of size 1×521 for individual action recognition.

The effect of partial occlusion on MDCLBP is presented in Figure 3.6; Euclidean distance between both histograms is 0.2 which shows that MDCLBP histogram is not much affected by partial occlusion.

3.4 Recognition of Individual Human Actions

After feature extraction, supervised learning is performed and the classifier is trained on extracted HOG-MDCLBP features for individual human action recognition. Separate classifiers are trained on each camera view and then the results of all classifiers are fused to get the final action decision.

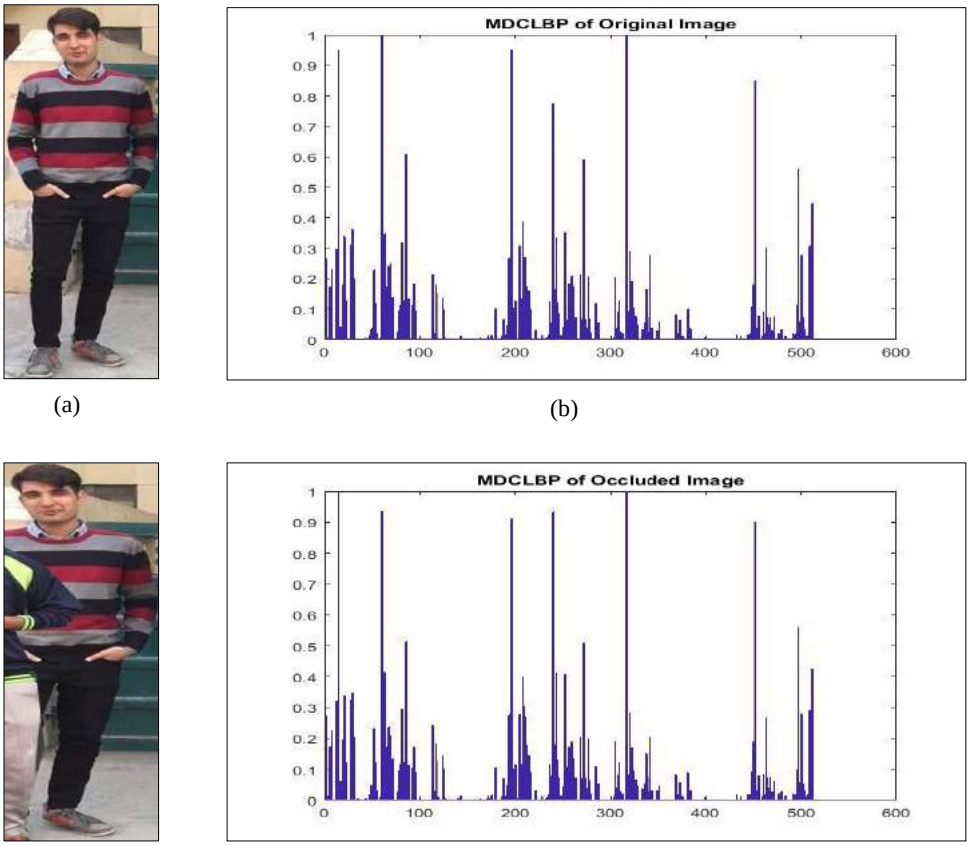

Figure 3.6: Effect of partial occlusion on MDCLBP features. The Euclidean distance between (b) and (d) is 0.2.

3.4.1 Support Vector Machines (SVMs)

SVMs were originally designed for binary classification with maximised margin criterion (MMC)[120]. Nevertheless, multiclass action recognition is required in real-world applications to recognize a variety of actions. The multi-class classification has been performed by using two common methods: one-against-one [121] and one-against-all [122]. It is proved by Zhang et al. [121] that both methods generate almost the same results. The one-against-one method is used in this thesis to perform multi-class classification using radial basis function (RBF) kernel with parameters c=1 and gamma=0.7.

If $\{Y_i^k\}_{i=1}^m$ be the vector containing labels assigned to each instance, the SVM decision function can be written as follows:

$$min\frac{1}{2}\|w\|^2 + ¢\sum_{i=1}^{m}\xi_i \;\; subject\ to\ Y_i\,(\vec{w}.K(o_i) + b) \geq 1 - \xi_i \qquad (3.10)$$

Here, $\left\{(o_i, Y_i)\middle| o_i \in O_i^k, Y_i \in \{1,2, \dots \mathcal{C}\}\right\}_{i=1}^{m}$.

Where ¢ is the regularization parameter, K is the kernel function, $\mathcal{C}$ is the total number of action classes and m is the number of persons in a frame k. A multiclass SVM is trained by using RBF kernel and validation is performed using a 10-fold cross validation scheme. As we need to recognize actions at each frame, each of the individual frames is utilized in learning and testing process.

RBF kernel and its parameters are selected after performing classification with different kernels by varying regularization parameter ¢ and the value of gamma in kernel functions. The best results are achieved using Radia basis function (RBF) with gamma=0.7 and ¢=1.

3.4.2 Decision Fusion

The two commonly known fusion methods to associate the estimations from multiple camera views include early fusion and late fusion, which are also called feature level fusion and decision level fusion respectively. In early fusion method, features are extracted under each camera view and combined with different methods to get the valuable features among all [123]. In late fusion, the classification is performed directly on extracted features under each camera view and the scores of classifiers are fused to get final decision [124]. Late fusion is performed in this research because in early fusion each camera shares huge amount of information, however, only the decisions of classifiers are shared in decision level fusion approach. Different methods of late fusion are available in literature [123], we have chosen majority voting; the action class that receives highest number of scores is considered as final action. The reason to chose majority voting is that in simple majority voting each classifier's output is independent of other classifers. The actors in dataset are free to choose any direction to perform actions, so we are unable to weight a classifier under a specific camera view in advance.

3.5 Experiments

3.5.1 Experimental Setup

In this section, performance of the proposed method for individual human action recognition is evaluated. The effectiveness of the proposed HOG-MDCLBP descriptor is assessed on two datasets and compared with some state-of-the-art approaches.

3.5.2 Datasets

The effectiveness of the proposed approach is evaluated on multi-view datasets for single human action recognition. The experiments are performed on two datasets: IXMAS and OIXMAS, these are the contemporary benchmark datasets in action recognition under multiple views and occlusions. Since this research aims to recognize individual action in each frame, the actions are further subdivided into the sub actions.

IXMAS Dataset

This dataset has been introduced by Weinland et al. [15] for multi-view human action recognition. The IXMAS dataset contains 13 action classes: scratch head, cross arm, check watch, get up, turn around, walk, wave , sit down, punch, point, kick, pick up and throw (overhead) (See Figure 3.7). The actions are performed 3 times by 11 actors. Videos are acquired by placing five cameras on different angles having a frame rate of 23fps and frame size of 390x291.

Figure 3.7: Multi- view IXMAS dataset example images showing five camera views [15]

The orientation and position are freely chosen by the actors. To perform a frame-by-frame evaluation of our proposed method; we have further divided action into segments or sub-actions. For example, the action *sit* is divided into three segments: stand, bend and sit; *get up* is split into three segments: sit, bend and stand; *scratch head* is divided into two segments: stand and scratch; the full action of *point* is composed of three parts: stand, point and stand. The *turnaround* is considered as *walk* in our experiment. In this way, total 11 numbers of action classes are found in IXMAS dataset.

OIXMAS Dataset

This dataset contains 11 actions as in IXMAS dataset, performed by different performers. The performers are occluded by adding artificial occlusions of different shapes, as shown in Figure 3.8. Five cameras are place on different locations for recording. Each action is performed 3 times by all performers.The Position and orientation are freely chosen by the actors and objects are also placed on different locations in each scenario.

Figure 3.8: Multi-view OIXMAS dataset showing artificial occlusions

3.5.3 Evaluation Method

The tests are performed on two datasets captured from multiple camera views. Individual human actions are recognized by using cross validation. In cross validation, k fold cross validation is used i.e. the whole data is randomly split into k equal size subsamples. A single subsample is reserved for validation and remaining $k-1$ subsamples are used for training the classifier. Commonly the value of k is chosen as 5 or 10 bacause these values result in a model with low biasness [125]. We performed 10 fold cross validation to evaluate the performance of proposed method. The value of $k = 10$ is selected empirically because it achieved better performance as compared to $k = 5$.

3.5.4 Evaluation of HOG-MDCLBP on IXMAS dataset

In this section, the proposed HOG-MDCLBP descriptor is evaluated on IXMAS dataset and the comparisons are performed with some state-of-the-art approaches.

Frame size in IXMAS dataset is 390×291, a bounding box of size 64×48 is drawn around the person and features are extracted for individual action recognition. First of all, HOG features are extracted from the person under bounding box of size 64×48. To extract HOG, the cell size of 8×8 is selected, the quality of extracted local information is affected by cell size; if the cell size is larger it loses more information. The input image (person window) is divided into 7 vertical blocks and 5 horizontal blocks, making 35 positions. After performing

block normalization of 16×16 blocks, we get 36×1 size vector from 16×16 block i.e. each block is represented by 36×1 size vector. All vectors extracted from the person window are concatenated to make a feature vector of size 1×1260. Since the size of feature vector is very giant and the persons with different sizes are also observed in real scenarios. HOG descriptor is represented with 9-bin histogram.

Next, MDCLBP features are calculated from detected person's bounding box. The MDCLBP operator returns 16-bit code; which is further split into 2 codes: MDCLBP1 and MDCLBP2, each of 8 bits. Two texture images are generated by applying MDCLBP operator iteratively on whole image and both images are represented with histograms. Both histograms are then concatenated to get the final texture feature descriptor of size 1×512.

1×9 sized HOG histogram is concatenated with 1×512 sized histogram of MDCLBP, which is named as HOG-MDCLBP feature descriptor of size 1×521 . In multi-frame approach, all frames in a sequence are used for training and decision of a full action is made when the action is completed. Experiments are also performed on full length actions and results are shown in Table 3.1. 96.58% accuracy is achieved; some similar actions are misclassified but the rate of misclassification is lower as compared to correctly classified actions.

Table 3.1: Confusion matrix of action recognition results based on full length actions in IXMAS dataset

	Check watch	Cross arms	Scratch head	Sit down	Get up	Turn around	Walk	Wave	Punch	Kick	Point	Pickup
Check watch	100	0	0	0	0	0	0	0	0	0	0	0
Cross arms	0	100	0	0	0	0	0	0	0	0	0	0
Scratch head	0	0	100	0	0	0	0	0	0	0	0	0
Sit down	0	0	0	94	6	0	0	0	0	0	0	0
Get up	0	0	0	7	93	0	0	0	0	0	0	0
Turn around	0	0	0	0	0	96	4	0	0	0	0	0
Walk	0	0	0	0	0	10	90	0	0	0	0	0
Wave	0	0	0	0	0	0	0	100	0	0	0	0
Punch	0	0	0	0	0	0	0	0	100	0	0	0
Kick	0	0	0	0	0	0	0	0	0	100	0	0
Point	0	0	0	0	0	0	0	0	11	0	92	0
Pickup	0	0	0	0	6	0	0	0	0	0	0	94

The proposed approach is validated by using frame-by-frame and multi-frame approach. Since individual action labels are needed at each frame, frame-by-frame learning is performed by using the features extracted from each frame for learning the classifier. Each action is divided into sub actions to represent that action in a single frame. Classification is performed on all camera views separately and results of all classifier are fused to acquire the final action class decision. Decision level fusion is performed majority voting based method [129] is used to get the final classification decision.

Next experiment is performed by testing the action at each frame separately. Action division is discussed in Section 3.5.2, as a result of subdivisions; a new action class *bend* is added in this scenario. The action *pick up* is further divided into following sequence: *stand*, *bend*, *sit*, *bend* and *stand pick up* action is removed in this experiment. Moreover, *turn around* is also considered as *walking* in this scenario. Table 3.2 shows the average accuracy of 96.81% in frame-by-frame-based recognition method by fusing the results of all cameras. Some misclassifications are detected in very similar actions i.e. *sit down* is misclassified as *bend*, *stand* is misclassified as *cross arms*, *walk* is misclassified as *kick* and *point* is confused with *punch*. This is for the reason that these actions are perceived similar when observed from different positions. The proposed method has correctly identified all remaining actions.

Table 3.2: Confusion matrix of frame-by-frame individual action recognition on IXMAS dataset

	Check watch	Cross arms	Scratch head	Sit down	Stand	Bend	Walk	Wave	Punch	Kick	Point
Check watch	100	0	0	0	0	0	0	0	0	0	0
Cross arms	0	100	0	0	0	0	0	0	0	0	0
Scratch head	0	0	100	0	0	0	0	0	0	0	0
Sit down	0	0	0	94	0	6	0	0	0	0	0
Stand	0	7	0	0	93	0	0	0	0	0	0
Bend	0	0	0	4	0	96	0	0	0	0	0
Walk	0	0	0	0	0	0	90	0	0	10	0
Wave	0	0	0	0	0	0	0	100	0	0	0
Punch	0	0	0	0	0	0	0	0	100	0	0
Kick	0	0	0	0	0	0	0	0	0	100	0
Point	0	0	0	0	0	0	0	0	8	0	92

3.5.5 Evaluation of HOG-MDCLBP on OIXMAS dataset

The proposed method is evaluated on OIXMAS dataset which contains same actions as IXMAS dataset with occlusions. OIXMAS is a challenging dataset as it contains occlusions under multiple camera views. So persons are partially or sometimes completely occluded

under some camera views. Table 3.3 shows the confusion matrix of accuracies with full length action recognition. Few cases of misclassifications are observed among *cross arms* and *check watch*, *pick up* and *get up*, *getup* and *sitdown*, *turn around and walk* and *punch* and *point*. These misclassifications are due to the reason of variety of occlusions in OIXMAS dataset. Confusion matrix shows that the average accuracy of 91.58% is achieved on occluded sequences. Cross camera view action recognition is performed the classification results and compared with state-of-the-art approaches. Tables 3.4 and 3.5 explicate the cross camera view action recognition performance which shows that accuracy is significantly improved with the proposed HOG-MDCLBP method as compared to other approaches.

Table 3.3: Confusion matrix of action recognition results based on full length actions in OIXMAS dataset

	Check watch	Cross arms	Scratch head	Sit down	Stand	Bend	Walk	Wave	Punch	Kick	Point
Check watch	98	2	0	0	0	0	0	0	0	0	0
Cross arms	10	90	0	0	0	0	0	0	0	0	0
Scratch head	0	0	88	0	12	0	0	0	0	0	0
Sit down	0	0	0	93	7	0	0	0	0	0	0
Stand	0	0	0	0	100	0	0	0	0	0	0
Bend	0	0	0	10	0	90	0	0	0	0	0
Walk	0	0	0	0	0	0	96	0	0	4	0
Wave	10	10	0	0	0	0	0	80	0	0	0
Punch	0	0	0	0	0	0	0	0	90	0	10
Kick	0	0	0	0	0	0	0	0	0	100	0

Cross camera view analysis is also performed, in which the experiments are explicitly performed by training the classifier on actions captured with one camera and testing on actions captured with another camera which is not ever seen by the classifier during training. This is the best approach to confirm the robustness of proposed method in handling multiple views and illumination variations. The comparison of cross camera view analysis is performed with two state-of-the-art approaches [127] [128] on IXMAS and OIXMAS datasets. [127] uses sparse code filtering for mining action pattern from multiple camera views. They incoded label information in sparse coding process for dictionary learning. The discriminative sparse codes and classifiers are jointly modelled using collaborative filtering. [128] make use of self-similarity matrices for encoding frame-to-frame respective changes. It is based on tracking joints of human body and then self-similarity matrices are computed from tracked body points. The comparison of average accuracies on IXMAS is shown in Table 3.4 and OIXMAS in Table 3.5 which shows acceptable accuracy in cross camera view

analysis. Diagonal entries demonstrate the results on same views which are considerably better because training and testing is done on identical views. Other entries show the results of cross views which are also acceptable instead of Cam4 which always show low accuracy as compared to other camera views. The reason of accuracy degradation is that Cam4 is fixed on top hence it provides the top view so this view is completely different from other camera views. When similar actions are observed from top, the confusion in recognition is increased. The results of cross camera recognition are averaged over all camera views. Comparison is performed with other approaches show that higher accuracy is achieved when analysis is performed with our proposed approach. This is significant because the proposed approach generalizes to new views not seen by the classifier. Average accuracy is slightly decreased in Cam4 when compared with [127] which is due to top view of camera. Though view-independent action recognition is achieved in [128], the self-similarity matrices computed in this technique are based on low-level features and achieved lower accuracy when compared with proposed approach and [127] especially when the person is viewed from top camera.

Table 3.4: Cross camera view action recognition and comparison of average accuracies on IXMAS dataset

	Testing Data					**Avg. Accuracy**		
Training Data	Cam0	Cam1	Cam2	Cam3	Cam4	**Proposed Method**	Wang et al. [126]	Junejo et al. [127]
Cam0	96.1	89.0	87.2	86.5	67.7	**85.3**	82.9	68.6
Cam1	85.5	97.0	87.9	88.0	57.5	**83.2**	83.2	68.6
Cam2	82.5	89.0	95.5	86.5	69.4	**84.6**	80.5	68.5
Cam3	87.0	89.7	88.4	94.6	82.0	**88.3**	79.4	66.1
Cam4	70.7	67.7	58.5	65.0	92.6	**71.0**	73.8	49.6

Table 3.5: Cross camera view action recognition and comparison of average accuracies on OIXMAS dataset

	Testing Data					**Avg. Accuracy**		
Training Data	Cam0	Cam1	Cam2	Cam3	Cam4	**Proposed Method**	Wang et al. [126]	Junejo et al.[127]
Cam0	98.0	79.7	80.0	80.7	60.6	**79.8**	60.8	58.3
Cam1	75.0	99.0	78.5	80.0	55.0	**77.5**	69.8	63.5
Cam2	78.0	80.5	98.5	76.0	62.7	**79.1**	65.7	53.8
Cam3	75.3	80.7	67.0	99.7	65.6	**77.6**	69.6	48.8
Cam4	50.8	46.0	58.5	53.8	88.7	**59.56**	58.5	45.3

The results of frame-by-frame action recognition are presented in Table 3.6 which shows that up to 92% accuracy can be achieved with the proposed method in a frame-based approach.

The comparison of proposed individual action recognition method with some state-of-the-art approaches is displayed in Table 3.7 which indicates that the proposed method attained performance improvement in occluded scenarios as well as in view point variations.

Table 3.6: Confusion matrix of frame-by-frame individual action recognition on OIXMAS dataset

	Check watch	Cross arms	Scratch head	Sit down	Get up	Turn around	Walk	Wave	Punch	Kick	Point	Pickup
Check watch	95	5	0	0	0	0	0	0	0	0	0	0
Cross arms	12	88	0	0	0	0	0	0	0	0	0	0
Scratch head	0	0	100	0	0	0	0	0	0	0	0	0
Sit down	0	0	0	93	7	0	0	0	0	0	0	0
Get up	0	0	0	2	86	0	0	0	0	0	0	12
Turn around	0	0	0	0	0	90	10	0	0	0	0	0
Walk	0	0	0	0	0	6	94	0	0	0	0	0
Wave	0	5	0	0	0	0	0	85	0	0	10	0
Punch	0	0	0	0	0	0	0	0	90	0	10	0
Kick	0	0	0	0	0	0	0	0	0	100	0	0
Point	0	0	0	0	0	0	0	0	10	0	90	0
Pickup	0	0	0	2	10	0	0	0	0	0	0	88

Table 3.7: Comparison of the proposed individual action recognition method with other state-of-the-art approaches

Ref.	Occlusions	View point variations	Accuracy
Weinland et al. [32]	×	✓	86.3 %
Weinland et al. [32]	✓	✓	76.7%
Baumann et al. [45]	×	✓	80.55%
Wang et al. [126]	×	✓	85.94%
Wang et al. [126]	✓	✓	78.03%
Proposed Method	×	✓	**96.58 %**
Proposed Method	✓	✓	**91.58%**

3.6 Statistical test for measuring significance of results

T-test is conducted to for measuring the significance of results of different methods. The proposed method is compared with five different techniques having occlusion and view point variations. The decision rule is that if $p<=0.05$ then the test is significant i.e there is significant difference in the results of all methods.

Table 3.8: Results of t-test

21.546	t statistics
0.000	P
82.24000	Mean difference

Table 3.8 shows the value of p is 0.000 i.e ($p < 0.05$) which means there is significant difference in the results of all methods. Mean difference is 82.24000 which lies between lower difference (72.9002) and upper difference (91.5798).

A comparison of SVM with other classifiers is also performed and results are presented in Figure 3.9 which explicates that higher accuracy is achieved by SVM on both datasets. SVM is preffered over other classifiers because it provides lower error rate even if the dataset is small.

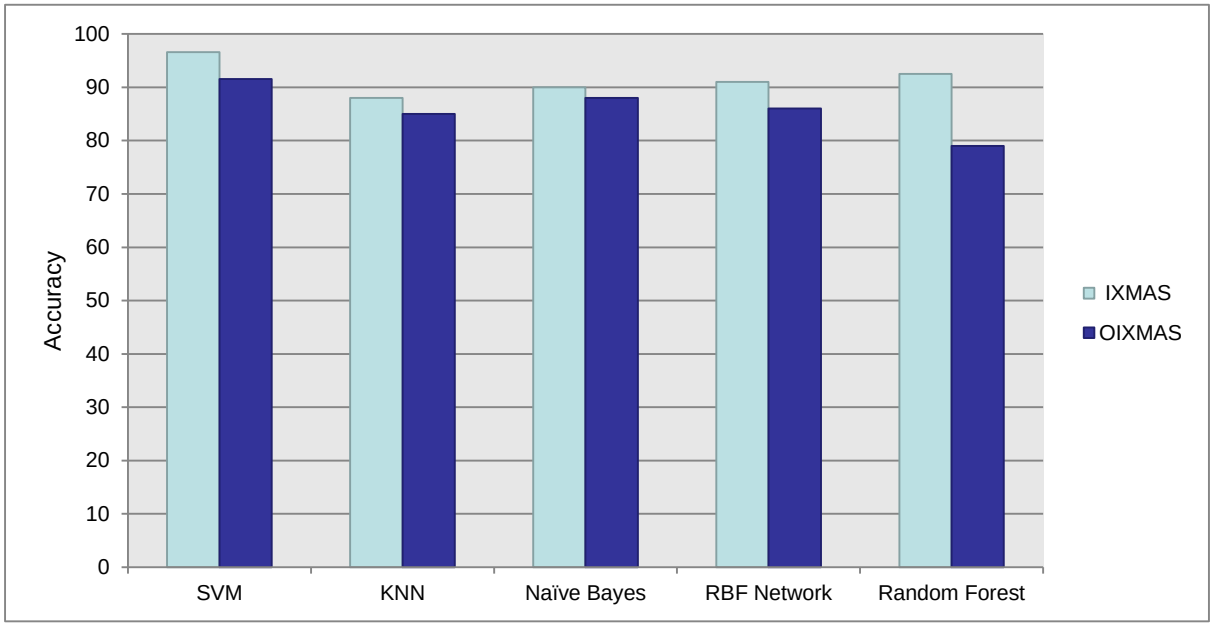

Figure 3.9: Comparison of SVM with other classifiers

A comparison of SVM kernels (linear, RBF, polynomial and sigmoid) is shown in Figure 3.10 which explicates that best results are achieved using RBF kernel. The best results with RBF kernel are achieved by setting gamma=0.7 and ¢ = 1. Second best accuracy is achieved by polynomial kernel. The accuracy dropped to 75% with linear kernel, because the dataset is not linearly separable. The minimum classification error is achieved with RBF kernel in both partially occluded and multiple views.

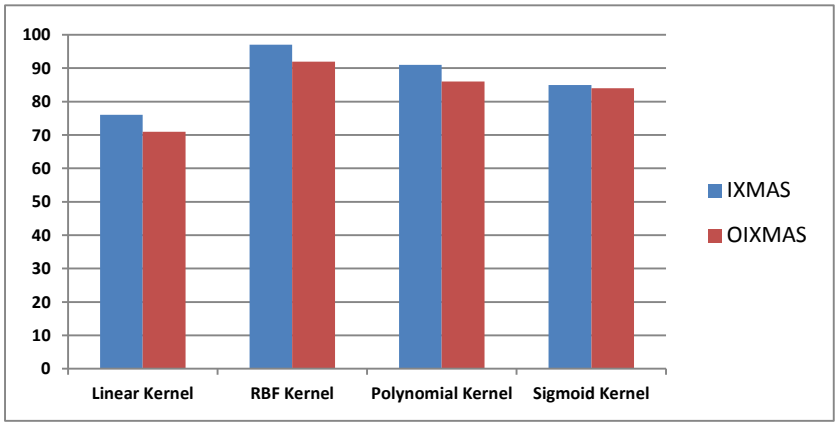

Figure 3.10: Comparison of SVM kernels for action recognition on IXMAS and OIXMAS datasets

3.6 Summary

In this chapter, a novel appearance based approach HOG-MDCLBP is proposed for the representation of actions in multiple camera environments. Such environments are prone to illumination variations, occlusions and view and scale variations. MDCLBP uses sign difference and the difference from median value to extract the texture features. The median is less affected by external effects, so the combination of sign difference and median value threshold eliminates the impact of illumination variations and occlusions. Gradient features are extracted by using HOG that is invariant to illumination variations. HOG and MDCLBP in combination provide the feature representation robust to partial occlusions and illumination variations. An action is represented by concatenating the histogram of HOGs and the histogram of MDCLBP features. Recognition is performed using SVM classifier. Experiments are done on multi-view dataset and occluded dataset to demonstrate the robustness of the proposed action representation technique. The experimental results show the sought-after performance not only in comparison to other state-of-the-art approaches but also validating the desired robustness against multiple views and occlusions. The average accuracy of 96.58% is achieved on multiple views and un-occluded dataset and 91.58% on occluded dataset. This leads us to believe that this approach can be used in surveillance applications under multiple camera environments. In next chapter, we have proposed a method for the recognition of small unit interactions by considering individual human actions along with trajectory information.

Chapter 4

Multi-Feature Small Unit Human Interaction Recognition in Multiple Camera Environments

4.1 Introduction

A concentration of works has been seen in recent years revolving around single person actions and group activity recognition [129]. The area of modelling small unit interactions between two persons under multiple cameras is still relatively less explored. Only recently, some studies have focused on the recognition of social interactions under single and multiple camera scenarios [10], [19], [21]. It is observed that activities in a public environment are rarely performed in isolation [22] ; mostly the people tend to engage in joint activities i.e. waiting in queue, talking together, walking together and physical violence etc. Recognizing human activities in public places has always been a challenging task due to many factors i.e. poor surveillance footage and indistinct actions. Analysis of nearby person's action is also important to distinguish between ambiguous interactions like *talking* and *queueing*. These two types of interactions are mostly confused when analysed individually; reaction of nearby person can be very useful to differentiate such types of interactions. Moreover, the action of nearby persons can also be a useful cue to detect alarming situations like physical violence and falling down etc.

In this thesis, the term ***person-to-person interaction*** is used for analysing the **interaction behaviour** of two persons. Joint modelling of actions, interactions and group activities [22], [89], [91] have been focused in some studies, but that research is restricted to single camera view. In the proposed approach, the concept similar to [22] is used, which suggests that analysis of individual person cannot provide reliable results in surveillance applications. Further, the concept similar to [91] is employed, which utilized poses and individual actions for interaction recognition. Different from [90], [91], the collective pose information *(same direction, opposite direction and facing each other)* of both persons has been extracted using HOG-MDCLBP descriptor. Individual person action recognition is performed using HOG-MDCLB descriptor to make the representation robust to partial occlusions. Individual human actions and collective poses are combined with trajectory features for the recognition of person-to-person interaction. Small unit social interactions are focused in this research. These

interactions may occur in public environments like *talking*, *stand together*, *walk separately* and *walk together*.

Trajectory features include: relative distance between two persons, collective pose, and the distance between current position and the previous position of each person. Instead of using raw features (HOG-MDCLBP), individual human actions identified in Chapter 3 are concatenated with trajectory features to keep the feature size small. Extracted features are then fed into SVM classifier for learning and then testing is performed on test sequences.

Block diagram of proposed method for small unit interaction recognition is depicted in Figure 4.1. The part under rounded rectangle depicts the process of individual person's action recognition (Chapter 3) and remaining part illustrates the process of small unit person-to-person interaction recognition.

This chapter is organized as follows: Section 4.2 discusses the motivation of recognizing small unit person-to-person interactions. Section 4.3, describes the presented method for person-to-person interaction recognition. Section 4.4 is about experimental setup and results.

4.2 Motivation

In this thesis, individual person actions are incorporated with trajectory features (called spatio-temporal features) for the recognition of small unit person-to-person interactions. Person-to-person interactions are recognized by perceived individual actions, distance between persons, collective pose, and the distance between the current position and the previous position of each person. This work is focused on the recognition of small unit social interactions between two persons i.e. *walk together*, *walk separately*, *stand together*, *talking* etc. which is termed as person-to-person interaction. Long term complex interactions: handshake, hug, bend, faint, kick, punch and push are also recognized and anticipated in Chapter 5. These interactions are defined as high-level events which contain the long term spatial and temporal interaction between objects of interest [130]. Analysis of individual action is crucial to recognize high-level events.

Instead of using action descriptors of both persons, only the action labels of both persons have been used for interaction recognition. In this way, the feature vector size is reduced.

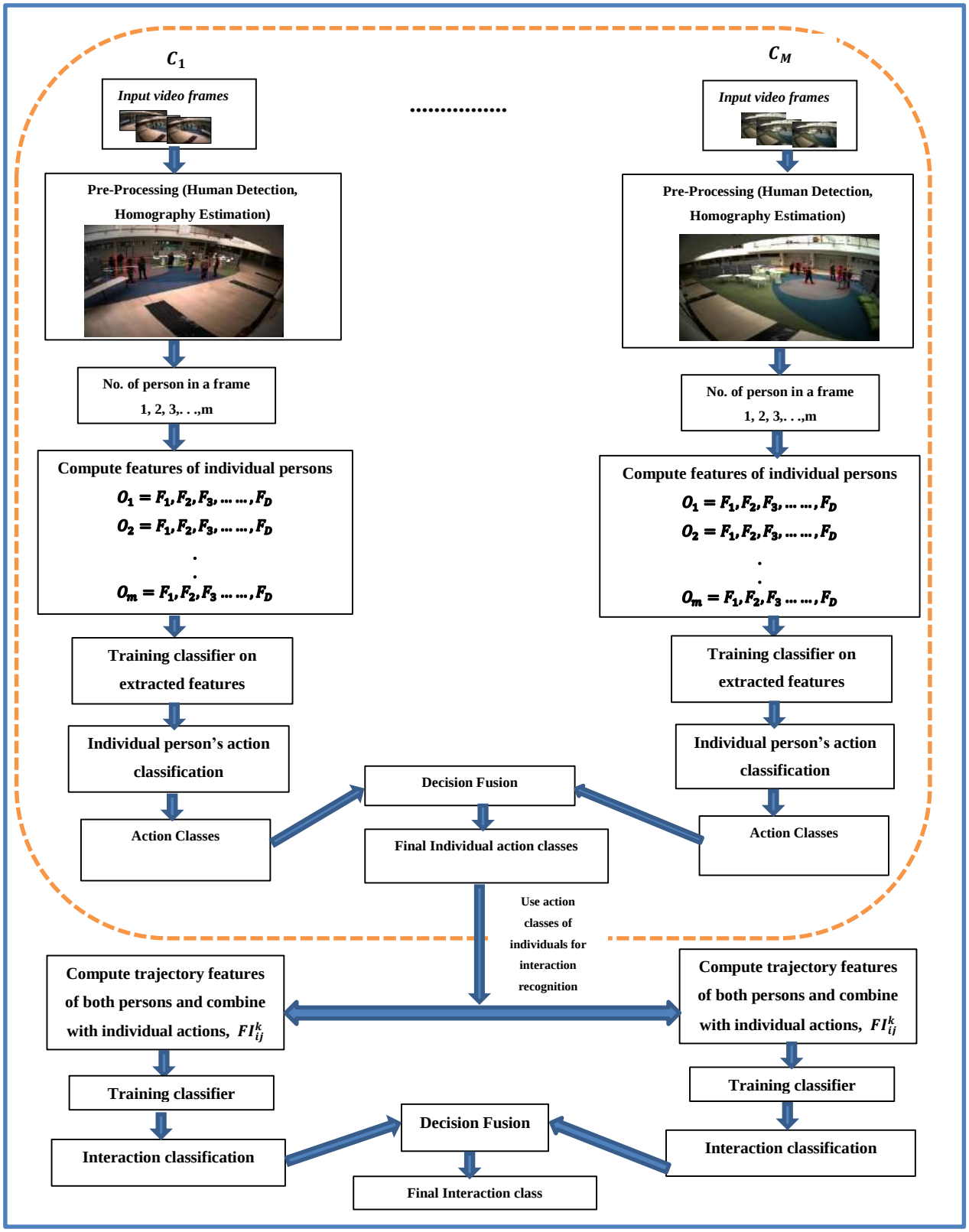

Figure 4.1: Block diagram of small unit sub interaction recognition

The relative pose feature used in this thesis is similar to [90] that helps in determining the type of interaction. For example, the *talking* and *standing* in a queue can be truly classified by the classifier if both persons *facing each other*.

Position differences of persons are used to help differentiate between *walk together* and *pass by*; position difference of two persons will be same if they *walk together* and different if one person *pass by another*.

4.3 Feature Extraction for the recognition of small unit person-to-person interactions

Let P_i^k be the focal person i detected in frame k, a distance threshold d^{th} is defined around P_i^k; the value of d^{th} is set empirically under each camera view. The persons detected under d^{th} are assumed to be the interacting persons as presented in Figure 4.2. The interaction of focal person with nearby persons is considered one by one for the recognition of person-to-person interaction.

Figure 4.2: illustration of selection of interacting persons under d^{th} . Green rectangle is drawn around focal person. The threshold area around focal person is represented with yellow rectangle. The persons detected under d^{th} are marked with red rectangle.

If P_j^k indicates j^{th} person identified in the chosen region, we consider the predicted action labels of P_i^k and P_j^k, location differences of P_i^k, location difference of P_j^k , relative distance between P_i^k and P_j^k and pose $\mathcal{P}_{i,j}$. The postures such as bend, walk, sit and stand of interacting persons can be easily differentiated with the help of predicted actions of each person. The relative distance between both persons is used to assist differentiate among *stand together*, *walk together*, *stand separately* and *walk separately*. Euclidean Distance is computed as:

$$D_{ij}^{k} = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2} \quad . \tag{4.1}$$

Where (x_a, y_b) and (x_b, y_b) are positions of P_i^k and P_j^k at frame k. Figure 4.3 depicts the tracked persons in three cameras. The ground positions of the person are available with HALLWAY dataset.

Figure 4.3: Images showing person's positions and trajectories in HALLWAY dataset.

In order to discern the interactions in which both persons have same individual actions, pose of persons is considered collectively. The collective pose is denoted with $\mathcal{p}_{i,j} \mid \mathcal{p}_{i,j} \in \mathrm{U}$ where U is the set with all poses among both persons. These poses are acquired through analysing both individual's poses which includes: *back to back*, *facing each other*, and *facing same side*. To esitmate $\mathcal{p}_{i,j}$, appearance information is acquired by extracting HOG-MDCLBP features from the bounding box which is defined around the individuals together. Histograms of HOG and MDCLBP features of three collective poses are shown in Figure 4.4. Collective poses are classified by training a SVM classifier on HOG-MDCLBP descriptors. Collective pose plays a key role in the analysis of interaction between two persons. Consider, the individual action of both persons is recognized as *walking* and the relative distance between them is small because they are very near to each other. The interaction between them will be considered as *walking together* if $\mathcal{p}_{i,j}$ is *same direction* otherwise *walking separately*. Similarly $\mathcal{p}_{\mathrm{i,j}}$ also differentiates among *talking* and *standing together*.

Next, the distance between the current position and the previous position of each person is computed. The position distance helps to accurately differentiate between *walk together* and *stand together* i.e. the position distance will change at each frame in *walk together*, otherwise the difference will be zero in *stand together*.

$$v\left(P_i^{k,k-1}\right) = T_{P_i^k}(x, y, k) - T_{P_i^k}(x, y, k-1) \ , \tag{4.2}$$

$$v\left(P_j^{k,k-1}\right) = T_{P_j^k}(x, y, k) - T_{P_j^k}(x, y, k-1)\ . \tag{4.3}$$

Here $v\left(P_i^{k,k-1}\right)$ and $v\left(P_j^{k,k-1}\right)$ are location differences of P_i^k and P_j^k . Finally, the output feature vector comprises: the labels of individual actions, collective pose, distance between two persons and the location differences of both persons. The form of final feature vector is:

$$FI_{ij}^k = \left(l_i^k, l_j^k, D_{ij}^k, \wp_{i,j}, v\left(P_i^{k,k-1}\right), v\left(P_j^{k,k-1}\right)\right). \tag{4.4}$$

Here l_i^k and l_j^k are the predicted labels of P_i^k and P_j^k respectively.

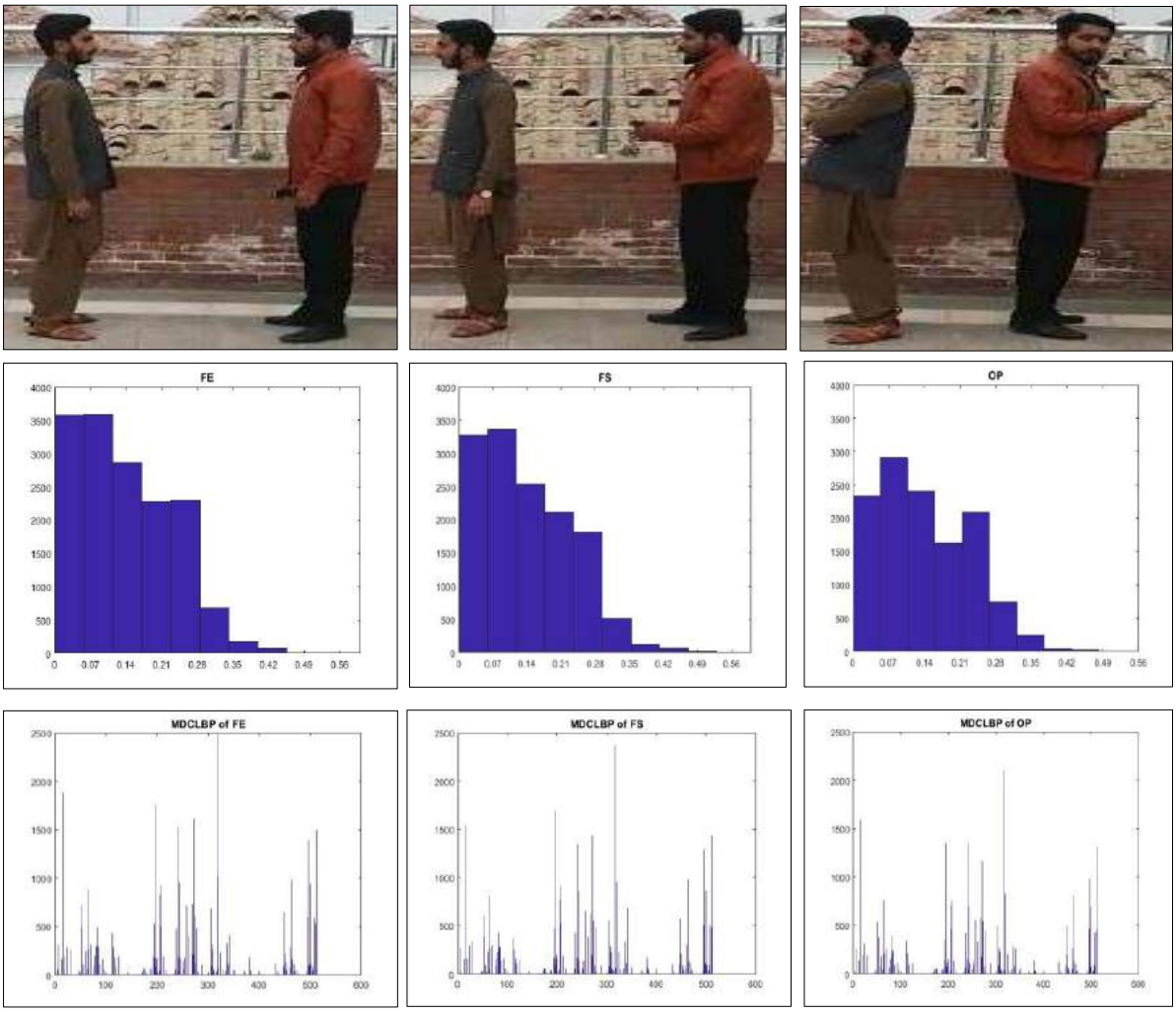

Figure 4.4: Histograms of HOG and MDCLBP features of three collective poses. (First row) Images of three collective poses: facing each other (FE), facing same direction (FS), and facing opposite (OP). (Second row) Histogram of HOG features. (Third row) Histograms of MDCLBP features.

4.4 Recognition of Small Unit Person-to-Person Interactions

Given a set $S = \{FI_{ij}^{k}, Y\}_{k=1}^{q}$ s.t $(FI_{ij}^{q}, q = 1 \dots Q)$, where Q is the total number of input examples and Y is the set containing labels of all interaction classes; $\alpha_q \in Y$, a multiclass SVM classifier [131] is trained on input examples i.e.

$$min\frac{1}{2}\|w\|^2 + ¢\sum_{q=1}^{Q}\xi_q \ subject\ to\ \alpha_q\left(\vec{w}.K(FI_{ij}^{q}) + b\right) \geq 1 - \xi_q \quad , \quad (4.5)$$

where w is margin, ξ_q is slack variable, ¢ is regularization parameter that controls the trade-off between margin and error and K is the kernel function. RBF kernel with gamma=0.07 and ¢=0.15. The training and testing is carried out on randomly selected data by using 60% of entire data for training and remaining 40% for testing.

4.4.1 Decision Fusion

Late fusion or decision level fusion is performed in this thesis to obtain the final decision on interaction. Majority voting [128] based fusion is performed in which an interaction class is chosen if that class is correctly identified by all classifiers or by majority classifiers.

4.5 Experimental Setup

Experiments are performed on partially overlapping camera views. Person-to-person interactions are analysed in a social environment having multiple persons performing different interactions.

4.5.1 Dataset

The efficacy of proposed multi-features interaction recognition method is tested on publically available HALLWAY dataset.

4.5.1.1 HALLWAY Dataset

HALLWAY dataset includes four individual actions: *walk*, *stand*, *sit* and some examples of *bend*; small unit interactions comprise *walk separately*, *stand together*, *walk together* and *talking*. Multiple cameras are used to record the sequence in an open area having day light illumination and the area is equipped with wall posters, chairs and desks. This sequence is 5 minutes long with maximum nine persons moving and interacting with each other. Frame rate for sequences is 15fps and the resolution is 800×600 pixels.

The positions of persons are provided along with the dataset. Homography matrix is computed to associate persons in different camera views. A bounding rectangle is drawn

around two nearby persons by merging the bounding boxes of both persons. For experiments, three cameras have been chosen. Some example frames are shown in Figure 4.5.

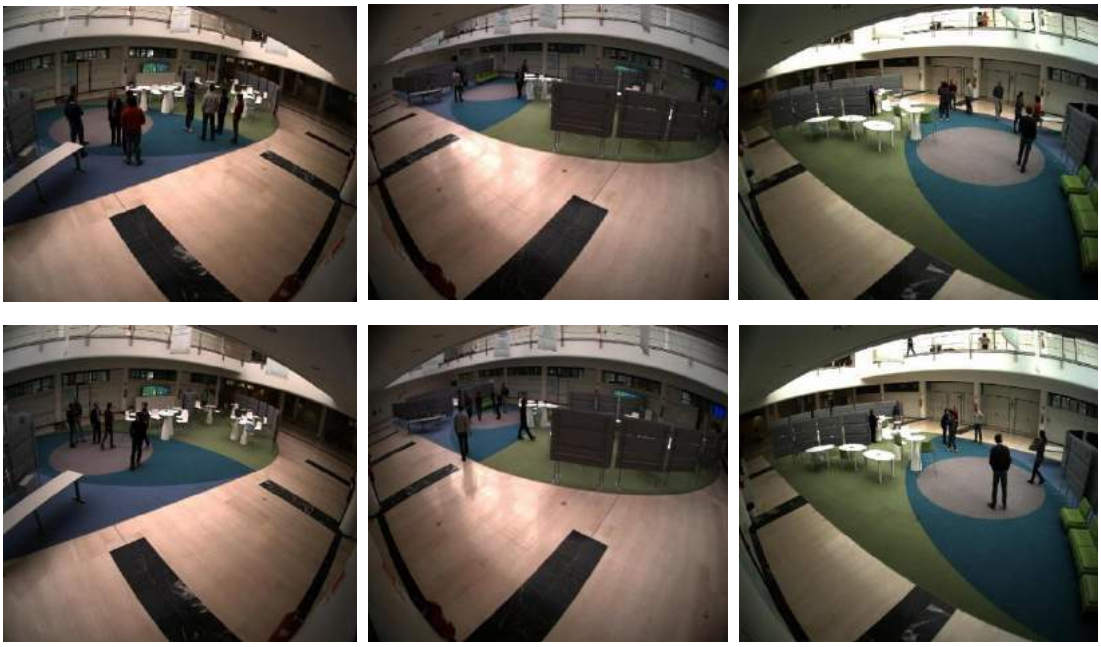

Figure 4.5: Example frames of HALLWAY dataset

4.5.2 Evaluation Method

In this section the proposed method for the recognition of small unit interactions is evaluated. The implementation of proposed interaction recognition is done in two steps: individual human action recognition (Chapter 3) and person-to-person interaction recognition. Experimental results in Chapter 3 have demonstrated the efficacy of proposed HOG-MDCLBP in the process of reconizing individual human actions. Therefore, before recognizing the interaction between two persons, the individual actions are recognized using HOG-MDCLBP features. Person's locations are provided with HALLWAY dataset. Once the persons are detected and labels are assigned under each camera view, the next step is to recognize individual action labels using HOG-MDCLBP. For the recognition of person-to-person interactions; a person is selected as focal person and spatio-temporal features are extracted from the focal person and nearby person. In HALLWAY dataset, interaction recognition is performed after every one second (15 frames).

4.5.3 Experimental Results

Small unit person-to-person interactions are analysed by using individual actions (l_i^k and l_j^k) of both persons in interaction, the distance among both individuals , location differences of interecting individuals and the information of relative pose. Four small unit interaction

classes are defined which includes: *talking, stand together, walk separately and walk together*. Classification is done under each camera view and the results of classifiers altogether are fused to obtain final decision.

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	99	0	0	1
Stand Together	0	96	0	4
Walk Together	1	0	99	0
Talking	0	10	0	90

(a)

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	98	0	0	2
Stand Together	0	96	4	0
Walk Together	10	10	80	0
Talking	0	0	2	98

(b)

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	100	0	0	0
Stand Together	0	95	5	0
Walk Together	1	0	99	0
Talking	0	1	1	98

(c)

Figure 4.6: Confusion matrices of proposed method for the recognition of small unit interactions on HALLWAY dataset. (a) Camera 1, 96% accuracy (b) Camera 2, 93% accuracy (c) Camera 3, 98% accuracy

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	100	0	0	0
Stand Together	0	96	0	4
Walk Together	1	0	99	0
Talking	0	1	1	98

Figure 4.7: Confusion Matrix for small unit interaction recognition on HALLWAY dataset after fusion

Confusion matrices in Figure 4.6 show that the proposed interaction recognition method attains satisfactory results under each camera observing small rate of misclassification. Fusion results are visualised in Figure 4.7 and Figure 4.8 shows the performance of separate classification under all cameras and decision fusion which demonstrates the improvement in system accuracy when classification decisions from multiple cameras are fused.

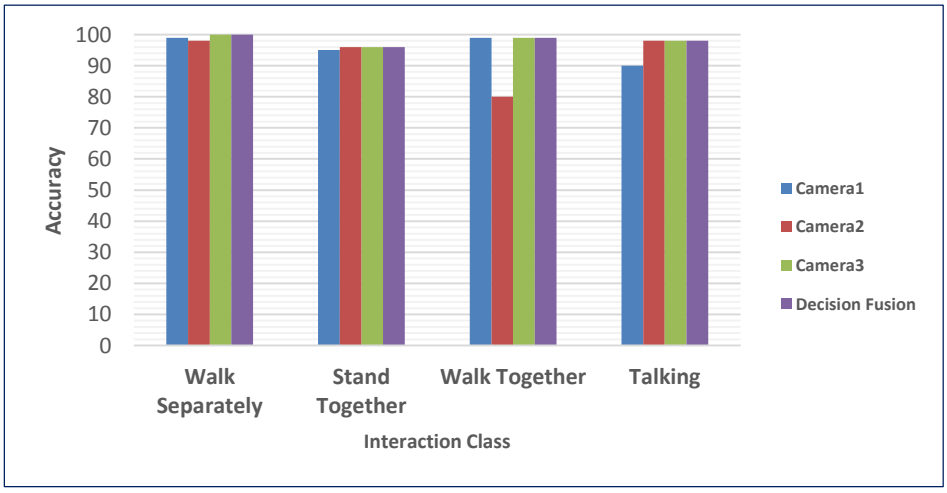

Figure 4.8 Comparison of the results of individual cameras for person-to-person small unit interaction recognition with decision fusion

To demonstrate the significance of the elements in proposed feature vector, tests have been accomplished by removing the elements of feature vector ($[l_i^k, l_j^k]$ and $\mathcal{p}_{i,j}$) one after another. firstly, individual activity labels l_i^k and l_j^k are excluded from F_{ij}^k .

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	100	0	0	0
Stand Together	0	9.1	90.9	0
Walk Together	0	8	34.5	57.5
Talking	0	0.2	1.6	98.2

(a)

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	100	0	0	0
Stand Together	0	7.9	92.1	0
Walk Together	0	4.4	59.3	36.3
Talking	0	0.2	1.2	98.6

(b)

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	100	0	0	0
Stand Together	0	19.2	80.8	0
Walk Together	0	40.2	30.6	29.2
Talking	0	10.0	20.9	69.1

(c)

Figure4.9: Confusion matrices for the recognition of small unit interactions from HALLWAY dataset by eliminating individual action's labels from the feature vector (a) Camera 1, 66.4% accuracy (b) Camera 2, 60.4 % accuracy (c) Camera 3, 39.72% accuracy

Confusion matrices in Figure 4.9 show the degradation of accuracy when the information of individual actions is removed from the feature vector. The misclassification rate of *Walk together* and *stand together* is particularly high which shows that individual actions are advantageous in discriminating the person's posture throught out the interaction. T*alking* and *walk together* are also confused when the labels of individual actions are removed from the feature vector. *Talking* is misclassified as *stand together* and in some cases it is misclassified as *walk together*. This is because individual actions play important role in recognizing interactions between persons.

The second test is conducted to demonstrate the significance of $\mathcal{p}_{i,j}$ in proposed feature set F_{ij}^k exclusive of pose information. Confusion matrices in Figure 4.10 show that all *standing together* cases are confused with *talking*. It is noted that during conversation, typically people face twards each other. In order to distinguish *stand together* and *talking* interactions, pose is therefore very essential. The same applies to *walk together* and *walk separately*, if both people walk in opposite directions, it will be regarded as a separate walk (*walk separately),* otherwise it will be misclassified if information of pose in not added in feature set. These experiments are carried out to demonstrate the significance of the proposed elements in feature set.

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	60.4	0.1	39.5	0
Stand Together	0	2	0	98
Walk Together	40.9	0	50.1	0
Talking	11	1	2	86

(a)

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	53.9	0	46.1	0
Stand Together	0	0	0	100
Walk Together	17	0	83	0
Talking	0.8	0	1.2	98

(b)

	Walk Separately	Stand Together	Walk Together	Talking
Walk Separately	86.3	0	13.7	0
Stand Together	0	0	0	100
Walk Together	27.1	0	72.9	0
Talking	0.9	0	1.4	97.7

(c)

Figure4.10. Confusion matrices for the recognition of small unit interactions from HALLWAY dataset by eliminating pose information from the feature vector (a) Camera 1, 64 % accuracy (b) Camera 2, 57.7% accuracy (c) Camera 3, 49.6%

Table 4.1 shows the statistical measures (average precision and recall) of all tests which demonstrates that proposed feature vector have achieved high values of precision and recall. The precision on proposed method is 98% and decreased by 39% in both scenarios when individual person actions and relative pose is removed from feature vector. 97% Recall rate is observed on proposed method and 68% when in dividual actions are excluded. The recall rate is reduced to 57% when the information of pose is removed from feature vector. Which indicates that the higher number of true positives are returned by proposed method.

Table 4.1: Comparison of precision and recall rates of proposed features set F_{ij}^k for the recognition of person-to-person interactions with $F_{ij}^k - [l_i^k, l_j^k]$ and $F_{ij}^k - p_{i,j}$ on HALLWAY dataset

	F_{ij}^k	$F_{ij}^k - [l_i^k, l_j^k]$	$F_{ij}^k - p_{i,j}$
Precision	0.98	0.59	0.59
Recall	0.97	0.68	0.57

The comparison of proposed method is performed with two state-of-the art approaches [35], [132] which utilized trajectories for recognition of small unit interactions. Experiments are performed on HALLWAY dataset to compare the performance of proposed method with previous approaches. The comparison of accuries is done in Table 4.2 which shows that proposed method outperforms other two approaches. These approaches [35], [132] utilized only trajectory based features for interaction recognition. In contrast, our proposed method considers individual actions of both persons along with trajectory features for interation recognition.

Table 4.2: Comparison of proposed method with state-of-the-art approaches on HALLWAY dataset

Ref.	Accuracy
Blunsden et al. **[132]**	90%
Lin et al. **[35]**	95%
Proposed method	**98.25%**

4.6 Summary

In this chapter, a method for recognition of small unit social (person-to-person) interactions is proposed. Person-to-persons interaction recognition is performed by incorporating individual person actions, collective pose information, location differences and distance. A multi-feature approach for interaction representation is presented which overcomes the issues of partial occlusions in public areas. A simple interaction representation method based on trajectory features is proposed and the importance of feature elements is validated experimentally. HALLWAY dataset is used for experiments.

Chapter 5

High-Level Human Interactions Anticipation using CNN-TOFCs Features

5.1 Introduction

This chapter describes the proposed method to anticipate high-level human interactions (complex activities) in multiple camera environments. Many researchers have focused on this active research topic [3], [23], [111], such studies however are restricted to a single camera view. A new methodology is proposed in this chapter for the anticipation of person-to-person interactions (two person's behaviour) under multiple camera environments. The task of behaviour anticipation requires the use of partial observations for early recognition of ongoing activities. It is a challenging problem to make the machine able to recognize unfinished activities. This problem becomes more challenging when the scene is monitored with multiple cameras having illumination variations and cluttered background.

Most of the previously proposed methods for human behaviour and activity prediction have focused on hand-crafted features i.e. trajectories [133], space-time features [31], [134] and motion capture data [32], which are aimed to capture visual properties of input image. These hand-crafted features alone are, though, not powerful to extract strong discriminative features for anticipation [23].

Recently, deep learning has been a new trend in computer vision and successfully employed by many researchers for the anticipation of human activities and interactions [25], [105], [112] and also in many other applications [104], [107]. The recent work in interaction recognition and anticipation tasks show that deep networks perform better than conventional hand crafted features [105], [107], [135]. The pre-trained CNNs can also be used as a feature extractor and the features extracted from CNN can be used to train other models [104], [106].

The interaction between two persons in a video can be viewed as a set of temporal frames. Temporal information should be analysed to anticipate the interactions. The deep features (using CNN) are extracted from each frame; these features do not provide temporal information that is crucial for the recognition of activities. Some researchers have solved this problem by providing the temporal information as input to the CNN for high-level tasks such as prediction and classification [23], [136], [137].

This chapter proposes to combine the hand-crafted (temporal) features and the deep features extracted from pre-trained Alexnet for the anticipation of human interactions in multiple camera scenarios. The extracted deep CNN features and temporal features are concatenated and presented to the SVM classifier for training. Training and testing are performed under each camera view separately and fusion is performed at decision level. That is, each camera anticipates the ongoing interaction and only the anticipation results are combined.

5.2 Motivation

This chapter presents a method that is used for the anticipation of ongoing high-level human interactions. The emphasis of this chapter is two-fold: the first emphasis is mainly concerned with the representation of human interactions in which we have presented to combine CNN features with Hand-crafted features for interaction representation. The second emphasis is concerned with the analysis of human interactions for the purpose of interaction anticipation at its early stages on the basis of learned feature descriptors.

Deep learning models have been successfully used in many computer vision applications such as action recognition [72], [137]–[139], person re-identification [104], action and interaction prediction [23], [25], [112], [140] and in face recognition [141]–[143]. These models have also been used for feature extraction from input images. Extracted features are then fed to any traditional classifier (like SVM) for training [144], [145]. Literature shows that deep features outperformed handcrafted features in some applications but the performance of pure deep learning models is still not adequate [146], which requires the use of hybrid features i.e. combination of deep and handcrafted features.

This chapter proposes to combine deep features with handcrafted (temporal) features for the anticipation of interactions. Deep features provide spatial information of every single frame of input video. Along with spatial information, temporal information is very crucial for interaction anticipation as it is difficult to anticipate the interaction class by looking at a single frame. Spatial information from each frame of input video is extracted using deep CNN and handcrafted (temporal) information is represented with optical flow components.

This chapter proposed to apply second order difference method on consecutive optical flow components (magnitude and orientation). Thresholding is applied on resultant components to remove the effects of small variations in background. Second order difference is preferred

over first order because the later detects very small variations in background. Second order difference provided fine magnitude with minimum background variations.

5.3 Proposed Method for Complex Human Interactions Anticipation

The proposed method for human interaction anticipation is depicted in Figure 5.1. Same process as depicted in figure is applied under each camera view and classifier decision fusion is performed to get the final results.

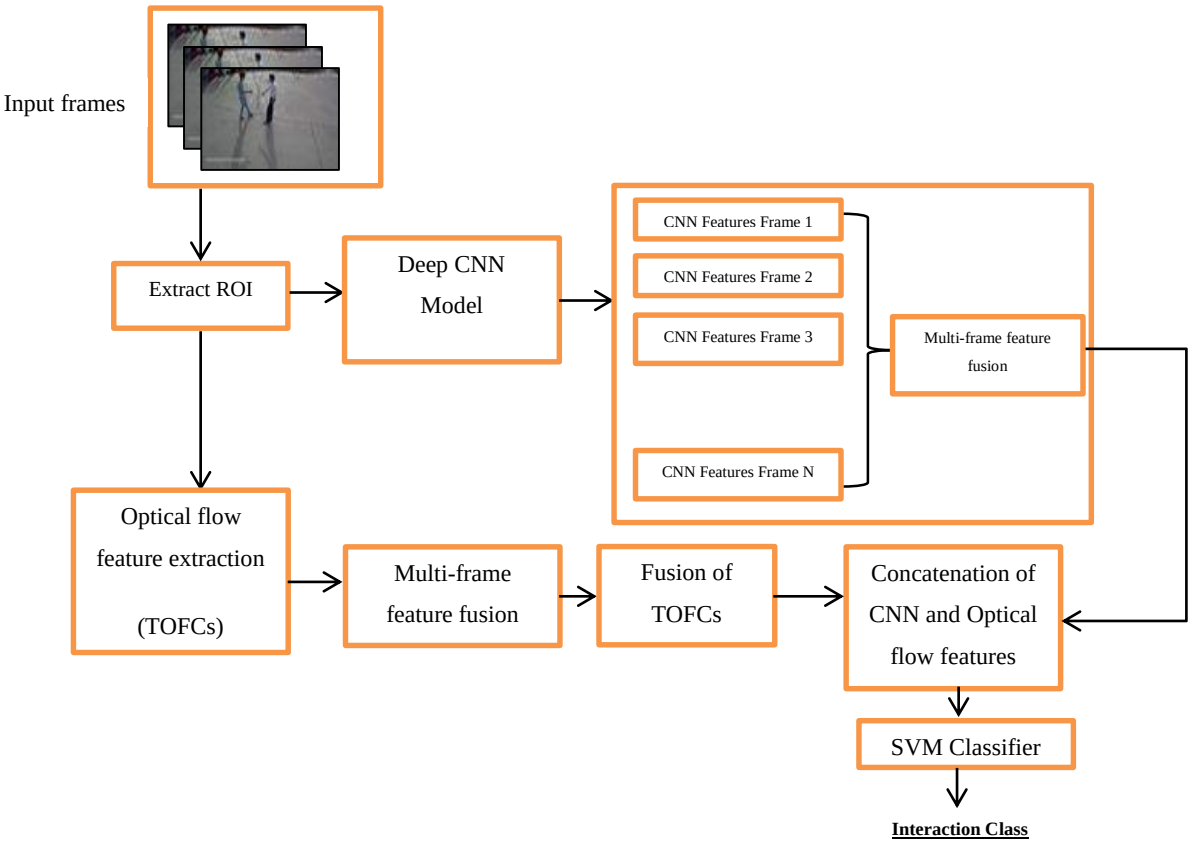

Figure 5.1: Depiction of proposed approach for interaction anticipation under one camera

5.3.1 Deep CNN Model for Feature Extraction

CNN features are extracted from input frames having only the detected ROI. The pre-trained Alexnet model is used for feature extraction. Alexnet is a deep network which is trained on a large data ImageNet , it won the Large-scale visual recognition challenge (ILSVRC) 2012 [147] by attaining the highest classification performance. The basic Alexnet architecture contains 8 layers: 5 convolution layers and 3 fully connected layers as depicted in Figure 5.2 and working of convolution layers is depicted in Figure 5.3.

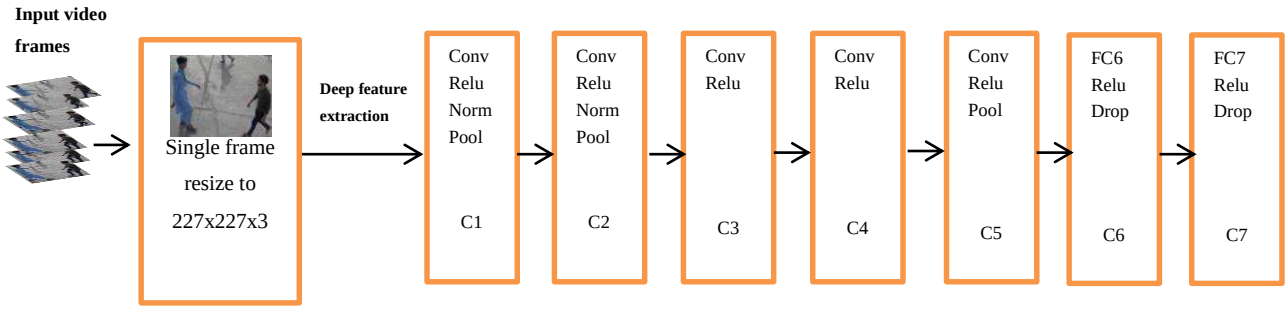

Figure 5.2: Alexnet net architecture for feature extraction

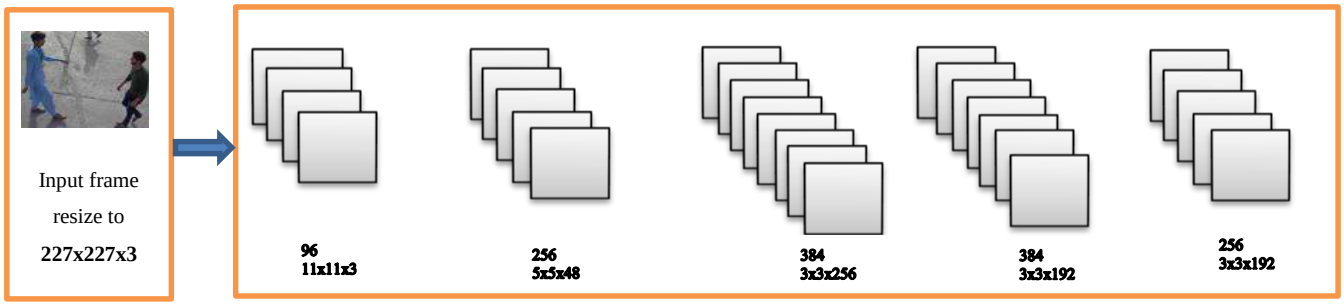

Figure5.3: Five convolutional layers in Alexnet

Each convolution layers is followed by a rectified linear unit (ReLU) layer and a local response linear layer (LRN) follows only first two convolution layers. Alexnet contains three max pooling layers: 2 after the first two LRN layers and 1 after the 5^{th} convolutional layer. ReLU is an activation function and used to rectify a signal, the LRN layer normalizes the convolution output and the max pooling layer applies window operation to select maximum value to reduce the size of layer output. C6 and C7 in Figure5.3 are the fully connected layers and each layer outputs 4096D feature vector. So, if there are N frames in input video, Deep CNN produced $N \times 4096$ features at C7 layer. In the proposed method, the output of C7 layer is used as feature vector for representing human interactions with deep CNN.

CNN features are extracted from each frame resulting $N \times 4096$ dimensional features. The output features of all frames of a video V are concatenated temporally by applying Median Absolute Deviation (MAD) on extracted deep features.

$$feat_{cnn}(V) = |\{fc(k)\}_{k=1}^{N} - median(fc)|. \quad (5.1)$$

Where fc is the matrix of CNN features and $feat_{cnn}$ is the resultant deep feature vector after applying MAD.

5.3.2 Temporal Feature Extraction

For the hand-crafted features, temporal information is extracted from four consecutive frames to get the temporal variations from input observations. For this, optical flow is computed and a new method called Transformed Optical Flow Components (TOFCs) is proposed to represent optical flow magnitude and orientation for human interaction anticipation. Process of temporal feature extraction is depicted in Figure 5.4

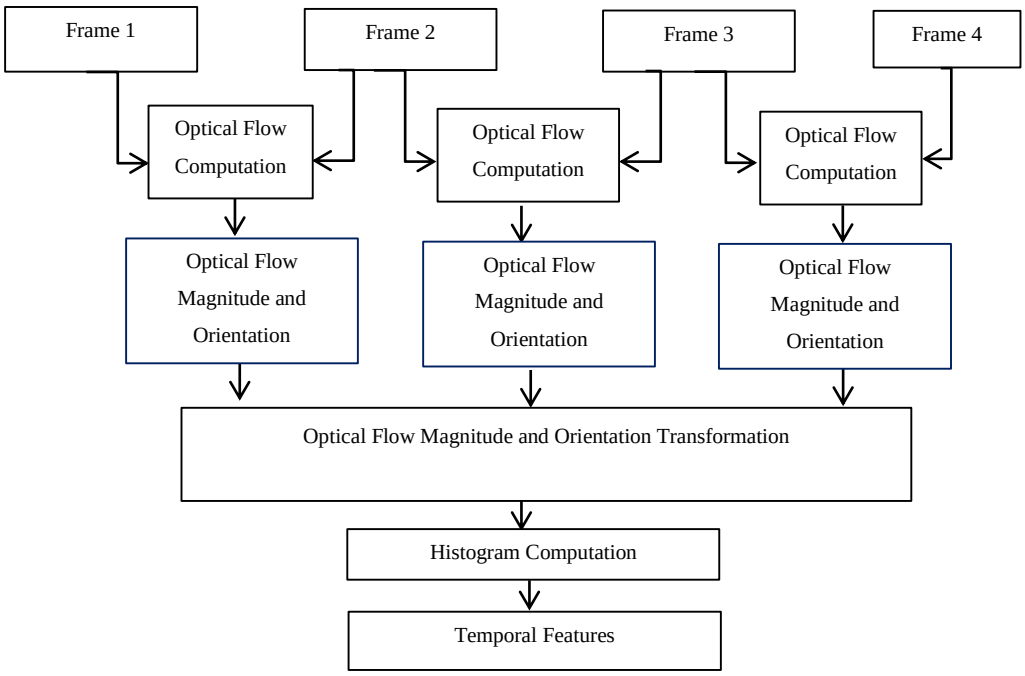

Figure 5.4: Diagram illustrating the process of temporal feature extraction

Horn Schunck optical flow [148] is the most widely used method to compute flow vectors. Let R_{ij}^{k} and R_{ij}^{k+1} be the region of interests around two persons in two consecutive frames k and $k+1$. Let $\mathfrak{r}_x$ and $\mathfrak{r}_y$ be the horizontal and vertical flow vectors computed by applying Equations (5.2), (5.3) with an iterative solution:

$$\mathfrak{r}_x^i = \mathfrak{r}_x^{i-1} - E_x\left[E_x\mathfrak{r}_x^{i-1} + E_y\mathfrak{r}_y^{i-1} + E_t\right]/\left(\alpha^2 + E_x^2 + E_y^2\right), \tag{5.2}$$

$$\mathfrak{r}_y^i = \mathfrak{r}_y^{i-1} - E_y\left[E_x\mathfrak{r}_y^{i-1} + E_y\mathfrak{r}_y^{i-1} + E_t\right]/\left(\alpha^2 + E_x^2 + E_y^2\right). \tag{5.3}$$

Where $\mathfrak{r}_x^{i-1}$ and $\mathfrak{r}_y^{i-1}$ are the smooth flow vectors of previous iteration,E_x, E_y and E_t are the partial derivative with respect to x,y and t, α is the weighting factor.

Optical flow magnitude is computed from flow vectors as follows:

$$\mathfrak{m}_{x,y} = \sqrt{\left(\mathfrak{r}_x^i\right)^2 + \left(\mathfrak{r}_y^i\right)^2}, \tag{5.4}$$

$$\theta_{x,y} = \tan^{-1}\left(\mathfrak{r}_y^i \Big/ \mathfrak{r}_x^i\right). \tag{5.5}$$

Where $\mathfrak{m}_{x,y}$ and $\theta_{x,y}$ are the magnitude and orientation at location(x, y). Since $\mathfrak{m}_{x,y}$and $\theta_{x,y}$ are extracted from the flow vectors which can have many noisy observations due to the illumination variations in videos. It is proposed to transform optical flow magnitude and orientation by applying second order difference on both components in consecutive frames. Caetano et al. [137] proposed to threshold magnitude and orientation values extracted from optical flow of two consecutive frames. This could be the early decision to threshold magnitude and orientation by looking at only the current information. More precisely, it is assumed that the temporal information can be effectively specified by the difference of optical flow components.

To extract TOFCs, second order difference of optical flow magnitudes and orientations in four consecutive frames is computed. A linear transformation is applied on optical flow magnitude to scale the values between 0-255. Thresholding is then applied on resultant optical flow magnitude and orientation.

Let m^1, m^2, m^3 and θ^1, θ^2, θ^3 be the magnitudes and orientations extracted from optical flow vectors of(I_{k-2}, I_{k-1}) (I_{k-1}, I_k) and (I_k, I_{k+1}) respectively. Element wise second order difference [149] of magnitudes and orientations is computed as follows:

$$\mathfrak{m}''_{x,y} = \mathfrak{m}^1_{x,y} - 2\mathfrak{m}^2_{x,y} + \mathfrak{m}^3_{x,y}, \tag{5.5}$$

$$\theta''_{x,y} = \theta^1_{x,y} - 2\theta^2_{x,y} + \theta^3_{x,y}. \tag{5.6}$$

Equation 5.5 and 5.6 are applied on optical flow components to enhance the temporal information in a frame by considering magnitudes and orientations of previous and next frames. The resultant magnitude is linearly scaled between 0 and 255 using linear transformation.

Next, the thresholding is applied on linearly transformed magnitude and orientation to get the temporal images. A threshold τ is selected empirically and filtering is performed on $\mathfrak{m}''$ on the basis of threshold value.

$$\mathfrak{m}''_{x,y} = \begin{cases} 0, & \mathfrak{m}''_{x,y} < \tau \\ \mathfrak{m}''_{x,y}, & otherwise \end{cases}, \tag{5.7}$$

The orientation component is thresholded as follows:

$$\theta''_{x,y} = \begin{cases} 0, & \mathfrak{m}''_{x,y} < \tau \\ \theta''_{x,y}, & otherwise \end{cases}, \tag{5.8}$$

Figure 5.5 shows an example of optical flow vectors detected in the background and the resultant magnitude after applying second order difference on corresponding pixels of four consecutive frames.

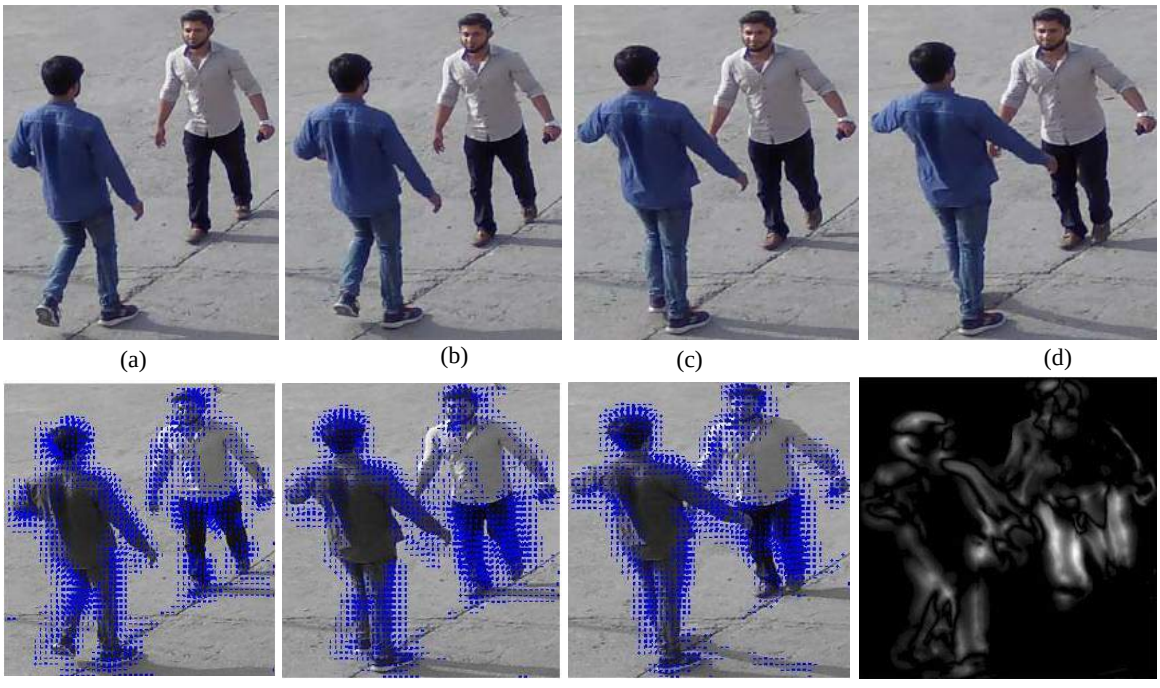

Figure 5.5: Illustration of temporal features with second order difference. (a-d) reference images at the frames $k-2, k-1, k$ and $k+1$. (e-g) are the optical flow vectors. (h) magnitude image showing the result of applying 2nd order difference.

5.3.3 Temporal Feature Representation

Extracted temporal features are represented in two ways:

1. By representing both transformed magnitude and orientation with histogram separately and then simply concatenating both histograms i.e. $[H(\mathfrak{m}''),H(\theta'')]$

2. By representing transformed components with histogram of oriented magnitudes called Histogram of Transformed Oriented Magnitude (HTOM), like HOFM in [150]. Different from HOFM, second order difference and thresholding is applied on optical flow components before computing oriented magnitudes from overall region of interest.

The steps to compute HTOM are as follows:

i. Orientations are represented with 8-bins in the range -12 to +12 using

$$-8 * pi/2 : 2 * pi/2 : 8 * pi/2$$

ii. The histogram is computed by looking at the magnitude and orientation value at each pixel location.

iii. Histogram bins are selected from orientations and the values of histogram are selected on the basis of magnitude.

iv. Figure 5.6 depicts the process of computing HTOM.

v. If the orientation at any location is greater than 12, magnitude will be added to last bin i.e. 12. For the orientation values less than -12, magnitude will be added to the first bin.

vi. The histograms of all frames of a video are fused by applying Equation (5.1) on all histograms of a video. HTOMs of *kick* interaction are displayed in Figure 5.7.

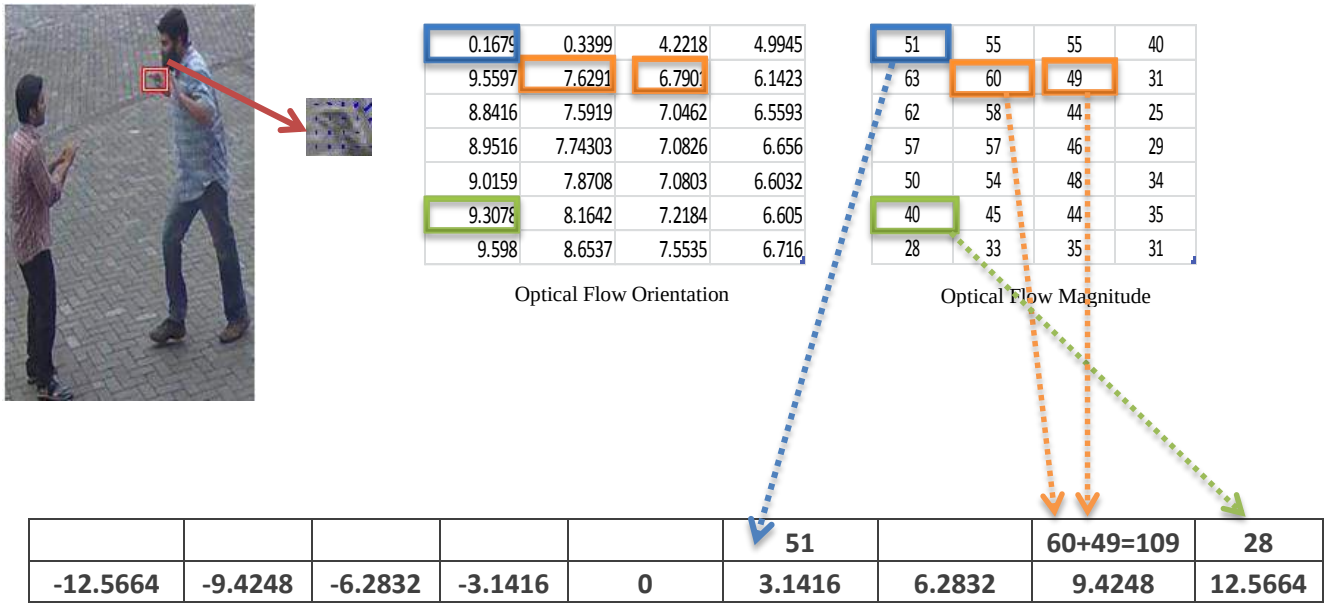

Figure 5.6: Process of computing HTOM.

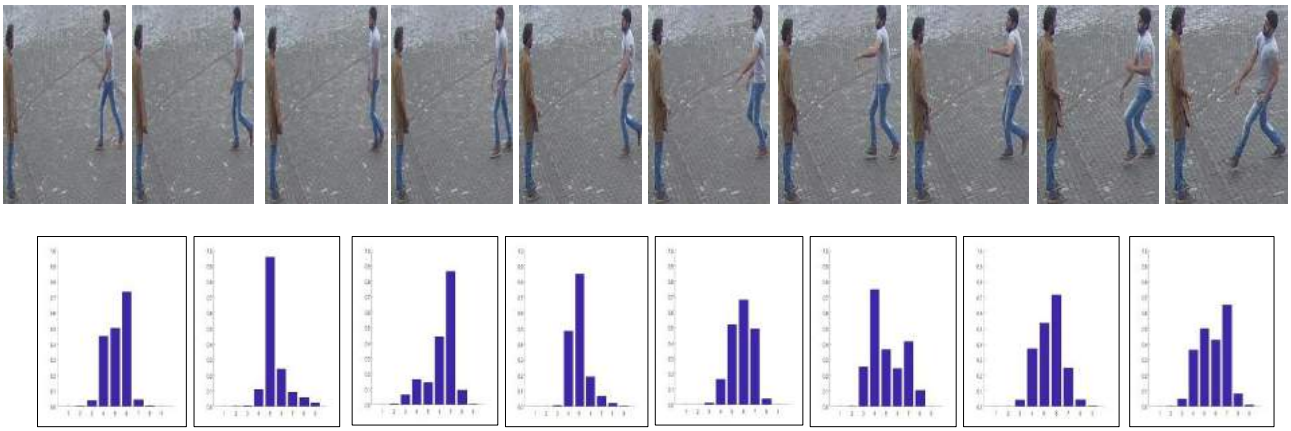

Figure 5.7: HTOM of kicking

5.3.3.1 Human Interaction Representation for Anticipation

The proposed human interaction representation is based on combining Deep CNN features and temporal features to detect unfinished interactions. To achieve this, we have concatenated Deep CNN features and temporal features for the representation of interaction between two humans.

$$feature(V) = [feat_{cnn}, feat_{tofc}]. \quad (5.9)$$

Where $feat_{cnn}$ are the CNN features and $feat_{tofc}$ are temporal features (either represented with concatenated histograms of oriented magnitudes or by concatenating both components).

Once the feature representation is done, the next step is to train the classifier. Training is performed on feature vectors which are represented using Equation (5.9).

5.3.4 Human Interaction Anticipation

After feature extraction, SVM classifier is used to recognize the interactions between persons using a subset of frames of full interaction. Classification is performed under each camera view separately and the decision results are fused to get the final anticipated interaction class. Majority voting technique [128] is used to decide the final class i.e. the correct interaction class is the one chosen by maximum classifiers. If different interaction classes are chosen by all classifiers then class with the highest accuracy is chosen as the final class [151].

5.4 Experiments

In the following section, the experimental results of the proposed interaction anticipation method are presented. Experiments are performed on two newly created (multi-view) datasets and a standard (single view) dataset. Section 5.4.1 details the experimental setup and the datasets used for evaluation. The pre-processing on MU-Interaction dataset is described in Section 5.4.1.2. Section 5.4.2 presents the evaluation method and the results of human behaviour anticipation method are presented in Section 5.4.3. The proposed technique is compared with some baseline methods and results are presented in Section 5.4.4

5.4.1 Experimental Setup

5.4.1.1 Datasets

The two datasets utilized in the experiments of this chapter are created in multi-view outdoor environments. To the best of our knowledge, no dataset containing high-level interactions in multi-view outdoor environment is available. These datasets are recorded with multiple IP cameras in Mirpur University. Detailed description of both datasets is given as follows:

5.4.1.1.1 MU-Interaction1

This dataset is recorded in the front of secretariat building of Mirpur University by using three IP cameras and comprises the videos of 7 interaction classes: *Bend, Faint, Handshake, Hug, Kick, Punch and Push*; performed by 8 different persons. 15-25 samples are recorded and the actors are free to exchange their positions i.e. for the abnormal interactions, like *kick*, the attacker is free to attack from any direction. The frame rate is 10 and resolution is 1920×1080 pixels. A snapshot of each camera view is depicted in Figure 5.8

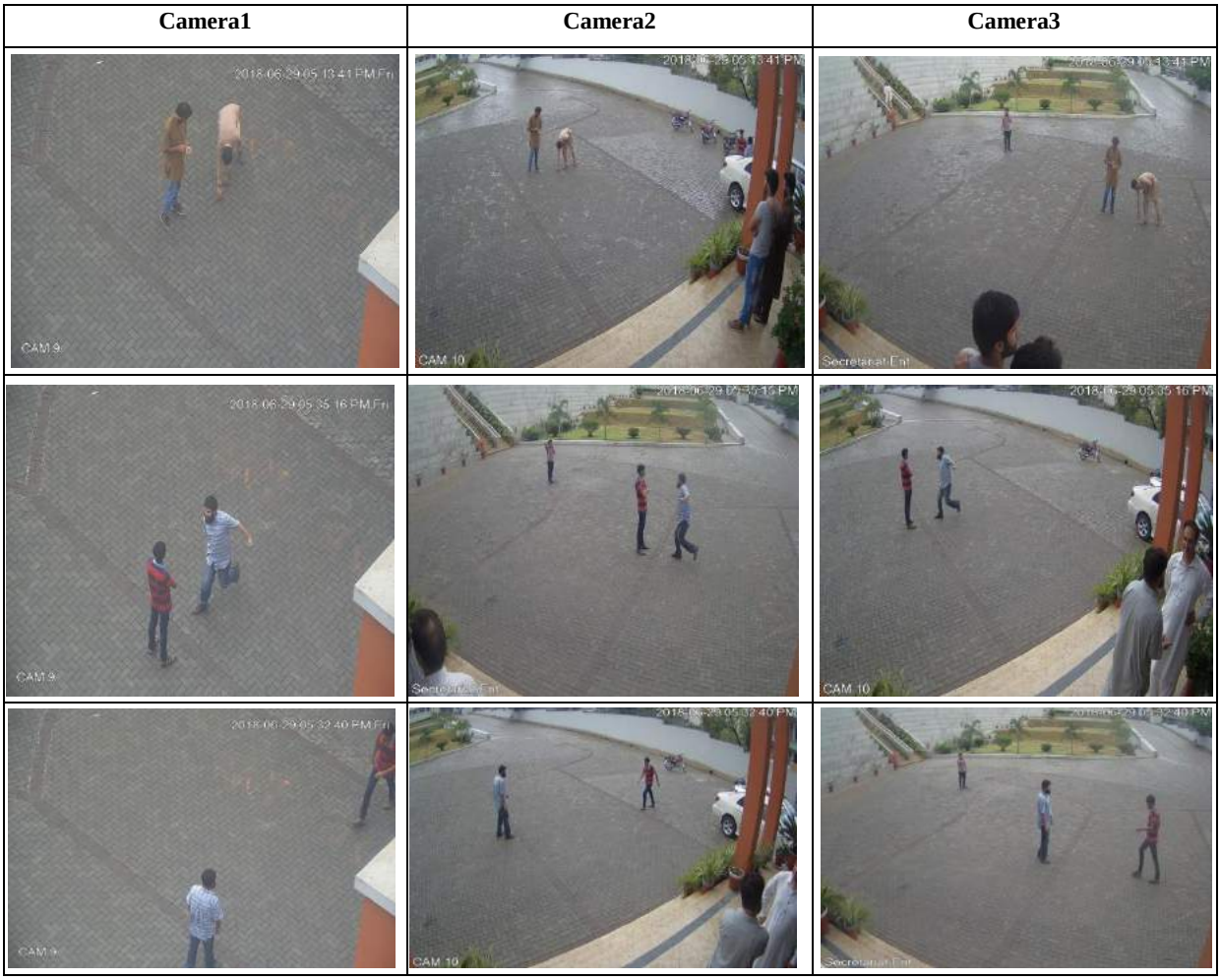

Figure 5.8: Example snapshots of each camera view

5.4.1.1.2 MU-Interaction2

MU-Interaction2 is very challenging dataset recorded outdoor at university entrance by using a network of three IP cameras and comprises the videos of 5 interaction classes: *Hug, Handshake, Kick, Punch, and Push*; performed by 8 different persons and several other persons also appear in the scene. Total 69 samples are collected under each camera view and the actors freely exchange their locations. Illumination variations and shadow are very prominent in this dataset. A snapshot from the dataset is provided in Figure 5.9, three Dahua IP cameras, with 1920×1080 pixel resolution and 10fps.

Figure 5.9: Snapshots from MU-Interaction2 dataset

5.4.1.2 Pre-processing on MU-Interaction Dataset

5.4.1.2.1 Human Detection

MU-Interaction dataset is captured with multiple cameras. People are detected under each camera view separately. The Aggregate Channel Feature [152] method is used for people detection. In this method, feature channel scaling is performed on input image by using Equation 5.10.

$$Ch = \text{ð}(I) \tag{5.10}$$

Where I input image and đ is a low level shift invariant function. Ch is a channel and the pixels in Ch are computed from parallel patches of input image I. Channels Ch are computed at every scale to attain feature pyramid representation. Channels include: HOGs, normalized gradient magnitude and LUV colour channel. All computed channels are divided into 4×4 blocks and pixels are summed in each block. Channels are finally smoothed with a smoothing filter. The weighted sum of channels in Ch is calculated to make the representation of channels quite general. Decision trees are learned over extracted features by using boosting approach to separate objects from background.

5.4.1.2.2 Homography Transformation

Homography transformation is used to transform a position (x, y, t) in an image plane to a position (x_G, y_G, t)in ground plane by applying Homography matrix (Equation (5.11)).

$$(x_G, y_G, t) = \grave{H}(x, y, t) \tag{5.11}$$

Where, (x_G, y_G, t) is the ground plane projection of $(x, y, t)$and $\grave{H}$ is the Homography matrix which is created by selecting control points to establish correspondence between image and ground plane. Figure 5.10 and Figure 5.11 shows control points **(marked red)** under both scenarios respectively.

Figure 5.10: Selected control points for Homography transformation in MU-Interaction1 dataset

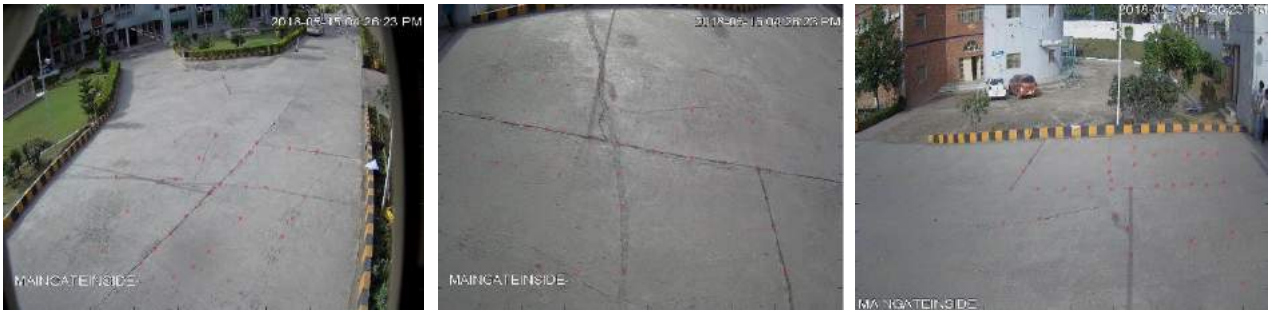

Figure 5.11: Selected control points for Homography transformation in MU-Interaction2 dataset

The points in image plane are selected from Camera 1 by detecting the people in subsequent frames and projected to Camera2 and Camera 3, a bounding box is drawn around the persons detected under all cameras.

5.4.2 Evaluation Method

Persons are detected using the algorithm presented by [152] and bounding boxes are drawn around all detected persons. The ROI contains both interacting person, which is selected by merging the bounding boxes of two persons. The ROI is the cropped and features are extracted from the selected ROI.

60% data is used for training and remaining 40% is used for testing the efficacy of proposed human interaction anticipation method. Leave-one-out cross validation is also used to assess the performance of proposed method i.e. testing is performed on one sample and training is performed on remaining samples. This process is repeated K times (the sample size), with each of the interaction sample is used in testing exactly once. Classification is performed under each camera view separately and the results of all classifiers are fused to get the final decision.

5.4.3 Evaluation of proposed CNN-TOFCs Human Interaction Anticipation Method

In this section, the proposed interaction anticipation method is evaluated on MU-Interaction dataset and the comparison is performed with some state-of-the-art approaches. MU-Interaction dataset is captured with three partially overlapping cameras and we have focused on the interactions performed in overlapping field of view.

Deep CNN features are extracted from ROI having two persons (Figure 5.12) by using Alexnet model. Alexnet is convolutional neural network which is trained on large ImageNet database. Alexnet is utilized for feature extraction and the output of the fully connected layer (C7) just before the final classification layer is used as feature vector. The input image is first resized to $227 \times 227 \times 3$ by the Alexnet model. The size of output feature vector for each image is 1×4096 and a video is represented by $N \times 4096$ features, where N is the length of video. To represent a video with a single feature vector, the features of all frames are combined by applying MAD.

Temporal features are extracted from optical flow components i.e. the magnitude and orientation of four consecutive frames. Second order difference of consecutive magnitudes and orientations is computed. The resultant components are further thresholded on the basis

of a threshold value τ that is set empirically ***(τ =20 for MU-Interaction1 and τ =35 for MU-Interaction2)***. Finally, the temporal features are represented with histograms and the features of all frames are combined by applying MAD.

Interaction between two persons is represented by combining Deep CNN and Temporal features. SVM is used for recognition of ongoing interactions.

Figure 5.12: Region of Interest (ROI). The blue box is the selected ROI around two persons. Each input frame is cropped into ROI before feature extraction

5.4.3.1 Experimental Results

Separate classification have been performed in each camera view by selecting the same video from each view. The results of classifier are fused to get the final accuracy of the proposed method. Fusion is performed on the basis of maximum score i.e. the class that is anticipated by maximum classifiers is depicted as final class. The performance of the proposed method is evaluated on the basis of accuracy, precision, recall and f-measure.

5.4.3.1.1 Experiments on MU-Interaction1

First experiment is performed on **MU-Interaction1** dataset captured with three cameras. This dataset comprises 136 samples from which 60% samples are used for training and remaining 40% for testing. The performance of the proposed approach is validated by applying leave-one-out cross validation. For anticipation, partial information is provided instead of providing full video frames. Experiments are performed on different observation ratios, from 0.2 to 1.0, with step size of 0.1 following the same procedure as in [23]. If N is the total number of frames in a video, $[1, round(0.3 * N)]$ frames means that 30% frames are used for anticipating the interactions. Eexperiments are performed repeatedly by randomly selecting training and testing sequences. Anticipation results of the proposed approach are evaluated using both methods of temporal feature representation i.e. (1) concatenating

histograms of magnitudes and orientations, (2) representing optical flow components with HTOM.

5.4.3.1.1.1 Evaluation of proposed interaction anticipation method when TOFCs are represented with concatenated histograms

Experiments are performed on different observation ratios using concatenation method. The accuracy of the proposed method on observation 0.2 is 30%; the performance of the proposed method is improved 20% when 30% observations are used for anticipation. An improvement of 15% is observed with the observation ratio 0.4. The accuracy achieved on observation ratio 0.6 is more significant **(92.59%)**, so this is selected as final accuracy because the next observations are closer towards the interaction completion. **94.5%** accuracy is achieved (2.5% is improved) when using full observations (that turns out to be the recognition of full interaction). An error rate of 8% is observed with leave-one-out cross validation.

Table 5.1 shows the confusion matrix of anticipation accuracies using concatenation method with observation ratio 0.6 after fusing the scores from all cameras. Recognition results of each interaction class under each camera view are computed and compared with other cameras. For example, the first class is Bend, all instances of Bend are correctly recognized in C1, so the recognition accuracy of Bend in Camera1 100%, in C2 and C3 the accuracy of Bend is 25% and 75% respectively; hence the decision made by C1 is selected as final decision. Likewise, the accuracy of Push is 100% in C1 and C3 and on the basis of majority voting all classes of push are correctly identified. **92.59%** accuracy is achieved by the proposed method; Table 5.2 depicts average accuracy, precision, recall and f-measure of proposed method.

Table 5.1: Confusion Matrix showing interaction anticipation accuracies after fusion of the scores of all cameras in MU-Interaction1 dataset when TOFCs are represented with concatenated histograms (Average accuracy=91.5%).

	Bend	**Faint**	**Handshake**	**Hug**	**Kick**	**Punch**	**Push**
Bend	**85.71**	14.29	0	0	0	0	0
Faint	0	**100**	0	20	0	0	0
Handshake	0	0	**100**	0	0	0	0
Hug	0	0	0	**100**	0	0	0
Kick	0	0	0	0	**91.67**	8.33	0
Punch	0	0	0	14.285	0	**71.43**	14.285
Push	0	0	0	0	0	8.33	**91.67**

Table 5.2: Average Precision, Recall and F-measure of proposed anticipation method when TOFCs are represented with concatenated histograms in MU-Interaction1 dataset

Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F-measure
0.91	0.90	0.92	0.91

5.4.3.1.1.2 Evaluation of proposed interaction anticipation method when TOFCs are represented with HTOM

The computation of HTOM is similar to [150], except from second order difference that is applied on optical flow components before computing oriented magnitudes. The process to compute HTOM is described in section (5.2.2.1).

In this method, temporal features that are extracted from ROI are represented with 9-bin histogram. When these features are combined with CNN features, it returned feature vector of length 1×4105. SVM is trained on these features and evaluation is performed on different observation ratios. Performance is improved 1.5% to 2% in each observation as compared to temporal features concatenation method. The proposed method achieved **94%** accuracy on observation ratio 0.6. Confusion matrix in Table 5.3 shows that the accuracy of individual interaction classes is improved with this representation method. Precision, recall and F-measure are shown in Table 5.4. It can be observed from precision and recall that the proposed method achieved acceptable performance in interaction anticipation.

Leave-one-out cross validation is applied using both representation methods and the error rate of 0.08 is observed with temporal features representation method and 0.06 with HTOM based representation.

Table 5.3: Confusion matrix showing anticipation accuracy after fusing the scores of all cameras in MU-Interaction1 dataset when TOFCs are represented with HTOM (average accuracy= 92.72%)

	Bend	Faint	Handshake	Hug	Kick	Punch	Push
Bend	**100**	0	0	0	0	0	0
Faint	0	**100**	0	0	0	0	0
Handshake	0	0	**80**	20	0	0	0
Hug	0	0	16.67	**83.33**	0	0	0
Kick	0	0	0	0	**100**	0	0
Punch	0	0	0	0	0	**85.71**	14.29
Push	0	0	0	0	0	0	**100**

Table 5.4: Average Precision, Recall and F-measure of Proposed Anticipation Method when TOFCs are represented with HTOM in MU-Interaction1 dataset

Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. F-measure
0.927	0.93	0.93	0.93

5.4.3.1.2 Experiments on MU-Interaction2

Second Experiment is performed on **MU-Interaction2** which is very challenging dataset, recorded in outdoor environment having shadows and cluttered background. 69 clips are recorded with each camera for five interaction classes. Learning and testing is performed under each camera and the results of all classifiers are fused to get the final decision. Leave-one-out training approach is used for this dataset i.e. each time 68 videos are used for training and testing is performed on single observation. The test is performed K times (K is the total number of samples) under each camera view and fusion is applied to get the final decision. Finally, the results are averaged to get the performance of proposed method.

5.4.3.1.2.1 Evaluation of proposed interaction anticipation method when TOFCs are represented with concatenated histograms

Experiments are performed with different observation ratios. It is observed from experiments that accuracy is improved in first 9 observations; accuracy is decreased by 0.5% when all frames are used for testing. The proposed method achieved **88%** accuracy in anticipating complex interactions from MU-Interaction2 dataset. The results of anticipations are visualized in Table 5.5 by using 60% information of overall interaction sequence in each video. Average accuracy, precision, recall and f-measure are displayed in Table 5.6. Anticipation accuracy is slightly decreased in this dataset as compared to MU-Interaction1. This is primarily due to challenging video sequences with waving trees and illumination variations and also a different number of samples per class.

Table 5.5: Confusion matrix showing anticipation accuracies after fusing the scores of all cameras in MU-Interaction2 dataset when TOFCs are represented with concatenated histograms (Average accuracy=86.34%)

	Handshake	Hug	Kick	Punch	Push
Handshake	**71.42**	14.29	0	14.29	0
Hug	11.11	**88.89**	0	0	0
Kick	0	0	**100**	0	0
Punch	7.14	0	0	**92.86**	0
Push	0	7.14	0	14.29	**78.57**

Table 5.6: Average Precision, Recall and F-measure of Proposed Anticipation Method when TOFCs are represented with concatenated histograms in MU-Interaction2 dataset

Avg. Accuracy	Avg. Precision	Avg. Recall	F-measure
0.86	0.87	0.86	0.86

Results show that the proposed method can anticipate complex interactions in some challenging conditions e.g. clutter background and shadows.

5.4.3.1.2.2 Evaluation of proposed interaction anticipation method when TOFCs are represented with HTOM

Same experiment setting is used as in previous experiments. MU-Interaction2 is very challenging dataset with illumination variations and some cluttered background. Although the concatenated histograms achieved accuracy up to 88%, the accuracy on this dataset is also improved when temporal features are represented with HTOM. Overall accuracy improvement of 3.3% is observed with HTOM and accuracy improvement ratio is also improved when experiments are performed on different observation ratios. Table 5.7 shows the confusion matrix of overall accuracies after fusing the results from all camera views. Proposed approach achieved overall **91.30%** of system accuracy when tested with leave-one-out approach. Performance measures are displayed in Table 5.8.

Experimental results clearly demonstrate the efficacy of the proposed method for the anticipation of complex interactions in outdoor challenging multiple camera environments.

Table 5.7: Confusion matrix showing anticipation accuracy after fusing the scores of all cameras in MU-Interaction2 dataset when TOFCs are represented with HTOM (Average accuracy=90.95%).

	Handshake	Hug	Kick	Punch	Push
Handshake	**85.71**	14.29	0	0	0
Hug	5.56	**83.33**	0	0	0
Kick	0	0	**100**	0	0
Punch	0	0	0	**100**	0
Push	0	0	0	14.29	**85.71**

Table 5.8: Average Precision, Recall and F-measure of proposed anticipation method on MU-Interaction2 dataset when TOFCs are represented with HTOM

Accuracy	Avg. Precision	Avg. Recall	F-measure
0.90	0.91	0.91	0.91

Experiments are also performed on both features (CNN and TOFCs) separately to demonstrate the importance of combining both features and results are displayed in Table 5.9. Deep features attained 65% accuracy on MU-Interaction1 and 60% on MU-Interaction2. The accuracy is decreased by 4% and 3% with TOFCs on both datasets respectively. Results demonstrate that although deep features extract strong high level feature. They alone are not strong enough for the representation of high level interactions for anticipation. Deep features along with hand crafted features can detect the salient motion information for interaction anticipation.

Table 5.9: Comparison of proposed features with separate feature elements

	Accuracy	
	MU-Interaction1	**MU-Interaction2**
Deep Features	65%	60%
TOFCs	61%	57%
Deep Features+TOFCs	92.72%	90.95%

5.5 Significance test

T-test is performed to measure the significance of results of different feature elements on both datasets. The Sig. value (p value) on Deep features + TOFCs is 0.006 and Mean is 91.53500 which shows that the results on combined features are significantly different from other feature components.

Table 5.10: T-test results to measure the significance of proposed features

	Sig.	Mean	95% confidence interval of the difference	
			Lower	Upper
Deep Features	0.025	62.50000	30.7345	94.2655
TOFCs	0.027	58.50000	26.7345	90.2655
Deep Features + TOFCs	0.006	91.83500	80.5900	103.0800

5.6 Comparison

To further validate the effectiveness of the interaction anticipation module, the proposed approach is compared with some state-of-the-art approaches on well-known publically

available UT-Interaction [153] dataset. This dataset is captured with single camera and the background is simple with slight camera jitter. This dataset is recorded with little different zoom rate. It consists of 6 interaction classes i.e. *Handshake*, *Hug*, *Point*, *kick*, *Punch* and *Push* and 10 sequences of each class. Table 5.9 displays the comparisons of proposed method with other approaches. The accuracies are achieved on 50% of each input video i.e. $[1, round(0.5 * N)]$ frames of each video are used for training and testing. The results explicate that our proposed method outperforms the currently available [23] highest results by 6% when tested on 50% observations. Ryoo's [134] is an earlier work which achieved 70% accuracy based on spatio-temporal features for interaction representation of UT-Interactions dataset. The accuracy is increased by 5% by [140] which utilized hierarichical movements for interaction representation. Further, the use of deep temporal features [23] proved to be more powerful as compared to handcrafted features. Finally, the experimental results of our proposed method on UT-Interaction and MU-Interaction datasets show that the interaction prediction rate is increased when handcrafted features are combined with deep features.

Table 5.9: Performance evaluation on UT-Interaction dataset

The classification accuracy of proposed approach for interaction/behaviour anticipation is compared with previous methods. The proposed method achieves state-of-the-art performance when compared with other methods

Ref.	Accuracy
Proposed Method	**94%**
Ke et al. [23]	88.3%
Lan et al. [140]	83.1%
Ryoo [134]	70%

5.7 Summary

In this chapter, a method for person-to-person interaction anticipation under multiple camera views has been presented. Specifically, a new feature is presented for the representation of interactions for anticipation. Deep features and temporal features have been used for interaction representation. Deep features are extracted by using pre-trained Alexnet model and temporal features are extracted by computing optical flow of four consecutive frames. A transformation is applied on optical flow magnitude and orientation. The purpose to apply this transformation is to enhance useful flow values and discard noisy observations. MAD is used to combine features extracted from all frames. Both deep and temporal features are

concatenated and SVM is used to recognize ongoing human interactions. Experiments are performed on a newly created dataset; MU-Interaction. Experimental results proved the efficacy of proposed method under multiple camera views.

Chapter 6

Conclusion and Future Work

6.1 Introduction

This thesis has set to propose computer vision methods to be able to recognize and anticipate human behaviours in public scenarios. Mainly, the thesis is geared towards solving the challenges in multiple camera environments (partial occlusions and illumination variations) while performing: individual human action recognition, human behaviour recognition and human behaviour anticipation.

Analysis of human activities is nontrivial in multiple camera scenarios due to occlusions, illumination variations, scale and orientation variations, and cluttered background. Comprehensive literature survey (Chapter 2) concluded that although many approaches have been proposed for vision based surveillance applications. There is still room for improving human behaviour analysis. Human Behaviour anticipation has gained the attention of computer scientists to make the system able to recognize the activities from half observations.

6.2 Contributions

1. The problem of individual human action recognition is resolved by proposing a new action descriptor HOG-MDCLBP. The proposed method achieved 96.58% accuracy on un-occluded dataset and 91.58% accuracy on occluded dataset under multiple camera views. Experimental results proved that the proposed feature is applicable to multiple camera views having partial occlusions and illumination variations .
2. Multi-feature based human behaviour/interaction recognition technique is presented. State/action of individual person and collective poses along with trajectory features are used for behaviour representation. Multi-feature representation approach has shown promising results. Small unit interactions are recognized with 98.25% accuracy.
3. Human behaviour anticipation is performed by combining Deep CNN and temporal features. A new method (CNN-TOFM) to represent temporal information is proposed. Optical flow components are transformed by applying second order difference and thresholding the resultant components. This transformation has reduced shadows and cluttered background effects from input video frames.

Human behaviour anticipation is relatively less explored problem and no dataset is available in multiple camera scenarios. We have created a new dataset for the evaluation of proposed anticipation module. The proposed method achieved 92% and 96% accuracy on MU-Interaction1 and MU-Interaction2 datasets.

Training and testing are performed using SVM classifier. The efficacy of the proposed method is compared with previous approaches and the results proved that the proposed method is able to anticipate ongoing behaviours/interactions with higher accuracy than other methods.

6.3 Future Enhancements

This study is focused on Individual Human Action Recognition, Human Behaviour Recognition and Human Behaviour anticipation under multiple camera views. Although acceptable accuracy is achieved in all modules, however, there are some future directions that need to be considered.

1. We are intended to include more complex and crowded scenarios to validate the performance of our proposed methods.
2. Handcrafted features are extracted for recognition of individual actions and interactions. Further research can be carried out to explore deep features and models for action and interaction recognition.
3. In future, we are planning to extend Behaviour Anticipation problem on multiple non-overlapping cameras. Non-overlapping cameras are installed in many public places which make the anticipation task more challenging. Person tracking and re-identification are required to re-identify a person when it exits from one camera and enters into another camera view. The methods in such scenarios should also be able to predict the action of a person performed between one camera view to another camera view.
4. Online Human Behaviour Prediction is necessary for many applications such as robotics etc. which requires real time processing. Alexnet is used in this research for the extraction of features from input sequences. Other deep learning moduels can be used for feature extraction and classification to provide real time processing.

References

[1] L. Mazerolle, D. Hurley, and M. Chamlin, “Social behavior in public space: An analysis of behavioral adaptations to CCTV,” *Secur. J.*, vol. 15, no. 3, pp. 59–75, 2002.

[2] M. P. Jadhav, M. S. Suryawanshi, and M. D. Jadhav, “Automated Video Surveillance,” 2017.

[3] M. S. Ryoo, “Human activity prediction: Early recognition of ongoing activities from streaming videos,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 1036–1043.

[4] P. V. K. Borges, N. Conci, and A. Cavallaro, “Video-based human behavior understanding: A survey,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 11, pp. 1993–2008, 2013.

[5] I. Bayer and T. Silbermann, “A multi modal approach to gesture recognition from audio and video data,” in *Proceedings of the 15th ACM on International conference on multimodal interaction*, 2013, pp. 461–466.

[6] V. Pitsikalis, A. Katsamanis, S. Theodorakis, and P. Maragos, “Multimodal gesture recognition via multiple hypotheses rescoring,” in *Gesture Recognition*, Springer, 2017, pp. 467–496.

[7] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, 2001.

[8] A. Amir *et al.*, “A Low Power, Fully Event-Based Gesture Recognition System.,” in *CVPR*, 2017, pp. 7388–7397.

[9] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local SVM approach,” in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, vol. 3, pp. 32–36.

[10] H.-I. Suk, A. K. Jain, and S.-W. Lee, “A network of dynamic probabilistic models for human interaction analysis,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 7, pp. 932–945, 2011.

[11] M. S. Ryoo and J. K. Aggarwal, “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,” in *Computer vision, 2009 ieee 12th international conference on*, 2009, pp. 1593–1600.

[12] B. Zhang, Y. Yan, N. Conci, and N. Sebe, “You Talkin’to Me?: Recognizing Complex Human Interactions in Unconstrained Videos,” in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 821–824.

[13] F. Baumann, J. Lao, A. Ehlers, and B. Rosenhahn, “Motion Binary Patterns for Action Recognition.,” in *ICPRAM*, 2014, pp. 385–392.

[14] J. R. Cózar, J. M. González-Linares, N. Guil, R. Hernández, and Y. Heredia, “Visual words selection for human action classification,” in *High Performance Computing and Simulation (HPCS), 2012 International Conference on*, 2012, pp. 188–194.

[15] D. Weinland, R. Ronfard, and E. Boyer, “Free viewpoint action recognition using motion history volumes,” *Comput. Vis. Image Underst.*, vol. 104, no. 2, pp. 249–257, 2006.

[16] B. B. Amor, J. Su, and A. Srivastava, “Action recognition using rate-invariant analysis of skeletal shape trajectories,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 1–13, 2016.

[17] O. R. Murthy, I. Radwan, A. Dhall, and R. Goecke, “On the effect of human body parts in large scale human behaviour recognition,” in *Digital Image Computing: Techniques and Applications (DICTA), 2013 International Conference on*, 2013, pp. 1–8.
[18] N. D. Rodríguez, M. P. Cuéllar, J. Lilius, and M. D. Calvo-Flores, “A fuzzy ontology for semantic modelling and recognition of human behaviour,” *Knowl.-Based Syst.*, vol. 66, pp. 46–60, 2014.
[19] M. Taj and A. Cavallaro, “Interaction recognition in wide areas using audiovisual sensors,” in *Image Processing (ICIP), 2012 19th IEEE International Conference on*, 2012, pp. 1113–1116.
[20] X. Ji, C. Wang, X. Zuo, and Y. Wang, “Multiple Feature Voting based Human Interaction Recognition,” *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 9, no. 1, pp. 323–334, 2016.
[21] S. Motiian, F. Siyahjani, R. Almohsen, and G. Doretto, “Online human interaction detection and recognition with multiple cameras,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 649–663, 2017.
[22] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, “Discriminative latent models for recognizing contextual group activities,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1549–1562, 2012.
[23] Q. Ke, M. Bennamoun, S. An, F. Boussaid, and F. Sohel, “Human interaction prediction using deep temporal features,” in *European Conference on Computer Vision*, 2016, pp. 403–414.
[24] Q. Sun, H. Liu, M. Liu, and T. Zhang, “Human activity prediction by mapping grouplets to recurrent Self-Organizing Map,” *Neurocomputing*, vol. 177, pp. 427–440, 2016.
[25] S. Choi, E. Kim, and S. Oh, “Human behavior prediction for smart homes using deep learning,” in *RO-MAN, 2013 IEEE*, 2013, pp. 173–179.
[26] R. Alazrai, Y. Mowafi, and C. G. Lee, “Anatomical-plane-based representation for human–human interactions analysis,” *Pattern Recognit.*, vol. 48, no. 8, pp. 2346–2363, 2015.
[27] M. S. Ryoo and J. K. Aggarwal, “UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA),” in *IEEE International Conference on Pattern Recognition Workshops*, 2010, vol. 2, p. 4.
[28] C. C. Loy, “Activity understanding and unusual event detection in surveillance videos,” PhD Thesis, 2010.
[29] R. Poppe, “A survey on vision-based human action recognition,” *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010.
[30] T. Hu, S. Messelodi, and O. Lanz, “Wide-area Multi-camera Multi-object Tracking with Dynamic Task Decomposition,” in *Proceedings of the International Conference on Distributed Smart Cameras*, 2014, p. 7.
[31] H. Wang, W. Yang, C. Yuan, H. Ling, and W. Hu, “Human activity prediction using temporally-weighted generalized time warping,” *Neurocomputing*, vol. 225, pp. 139–147, 2017.
[32] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, “Ongoing human action recognition with motion capture,” *Pattern Recognit.*, vol. 47, no. 1, pp. 238–247, 2014.
[33] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, “Leveraging structural context models and ranking score fusion for human interaction prediction,” *IEEE Trans. Multimed.*, vol. 20, no. 7, pp. 1712–1723, 2018.
[34] D. Weinland, M. Özuysal, and P. Fua, “Making action recognition robust to occlusions and viewpoint changes,” in *European Conference on Computer Vision*, 2010, pp. 635–648.

[35] W. Lin, Y. Chen, J. Wu, H. Wang, B. Sheng, and H. Li, "A new network-based algorithm for human activity recognition in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 826–841, 2014.
[36] T. Huynh-The, B.-V. Le, S. Lee, and Y. Yoon, "Interactive activity recognition using pose-based spatio–temporal relation features and four-level Pachinko Allocation Model," *Inf. Sci.*, vol. 369, pp. 317–333, 2016.
[37] R. Kavi and V. Kulathumani, "Real-time recognition of action sequences using a distributed video sensor network," *J. Sens. Actuator Netw.*, vol. 2, no. 3, pp. 486–508, 2013.
[38] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 2, pp. 236–243, 2013.
[39] T. Hu, X. Zhu, W. Guo, S. Wang, and J. Zhu, "Human action recognition based on scene semantics," *Multimed. Tools Appl.*, pp. 1–22, 2018.
[40] P. Y. Han, K. E. Yee, and O. S. Yin, "Localized Temporal Representation in Human Action Recognition," in *Proceedings of the 2018 VII International Conference on Network, Communication and Computing*, 2018, pp. 261–266.
[41] M. Singh, A. Basu, and M. K. Mandal, "Human activity recognition based on silhouette directionality," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 9, pp. 1280–1292, 2008.
[42] C.-H. Lin, F.-S. Hsu, and W.-Y. Lin, "Recognizing human actions using NWFE-based histogram vectors," *EURASIP J. Adv. Signal Process.*, vol. 2010, p. 9, 2010.
[43] D. K. Vishwakarma and R. Kapoor, "Hybrid classifier based human activity recognition using the silhouette and cells," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 6957–6965, 2015.
[44] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1932–1939.
[45] S. Mukherjee, S. K. Biswas, and D. P. Mukherjee, "Modeling sense disambiguation of human pose: Recognizing action at a distance by key poses," in *Asian Conference on Computer Vision*, 2010, pp. 244–255.
[46] F. Baumann, A. Ehlers, B. Rosenhahn, and J. Liao, "Recognizing human actions using novel space-time volume binary patterns," *Neurocomputing*, vol. 173, pp. 54–63, 2016.
[47] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using volume local binary patterns," in *Dynamical Vision*, Springer, 2007, pp. 165–177.
[48] O. Kihl, D. Picard, and P.-H. Gosselin, "Local polynomial space–time descriptors for action classification," *Mach. Vis. Appl.*, vol. 27, no. 3, pp. 351–361, 2016.
[49] Z. Cai, H. Neher, K. Vats, D. A. Clausi, and J. Zelek, "Temporal Hockey Action Recognition via Pose and Optical Flows," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.
[50] S.-R. Ke, H. L. U. U. Thuc, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "Human Action Recognition Based on 3D Human Modeling and Cyclic HMMs," *ETRI J.*, vol. 36, no. 4, pp. 662–672, 2014.
[51] K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image for human action recognition," *Pattern Recognit.*, vol. 45, no. 7, pp. 2562–2572, 2012.
[52] J. Ben-Arie, Z. Wang, P. Pandit, and S. Rajaram, "Human activity recognition using multidimensional indexing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1091–1104, 2002.
[53] S. Danafar and N. Gheissari, "Action recognition for surveillance applications using optic flow and SVM," in *Asian Conference on Computer Vision*, 2007, pp. 457–466.

[54] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, 2005, vol. 1, pp. 886–893.
[55] X. Cao, B. Ning, P. Yan, and X. Li, “Selecting key poses on manifold for pairwise action recognition,” *IEEE Trans. Ind. Inform.*, vol. 8, no. 1, pp. 168–177, 2012.
[56] E. Adeli Mosabbeb, K. Raahemifar, and M. Fathy, “Multi-view human activity recognition in distributed camera sensor networks,” *Sensors*, vol. 13, no. 7, pp. 8750–8770, 2013.
[57] F. Murtaza, M. H. Yousaf, and S. A. Velastin, “Multi-view Human Action Recognition Using Histograms of Oriented Gradients (HOG) Description of Motion History Images (MHIs),” in *Frontiers of Information Technology (FIT), 2015 13th International Conference on*, 2015, pp. 297–302.
[58] S. P. Sahoo, R. Silambarasi, and S. Ari, “Fusion of histogram based features for Human Action Recognition,” in *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2019, pp. 1012–1016.
[59] H. Lin, L. Chaisorn, Y. Wong, A.-A. Liu, Y.-T. Su, and M. S. Kankanhalli, “View-invariant feature discovering for multi-camera human action recognition,” in *Multimedia Signal Processing (MMSP), 2014 IEEE 16th International Workshop on*, 2014, pp. 1–6.
[60] R. Mattivi and L. Shao, “Human action recognition using LBP-TOP as sparse spatio-temporal feature descriptor,” in *International Conference on Computer Analysis of Images and Patterns*, 2009, pp. 740–747.
[61] “A Multiple Kernel Learning Based Fusion Framework... - Google Scholar.” [Online]. Available: https://scholar.google.com.pk/scholar?hl=en&as_sdt=0%2C5&q=A+Multiple+Kernel+Learning+Based+Fusion+Framework+for+Real-Time+Multi-View+Action+Recognition&btnG=. [Accessed: 23-Sep-2017].
[62] Y. Wang and G. Mori, “Human action recognition by semilatent topic models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 10, pp. 1762–1774, 2009.
[63] T. Guha and R. K. Ward, “Learning sparse representations for human action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 8, pp. 1576–1588, 2012.
[64] L. Xia, J. Huang, and L. Tan, “Human action recognition based on chaotic invariants,” *J. Cent. South Univ.*, vol. 20, no. 11, pp. 3171–3179, 2013.
[65] A. Iosifidis, A. Tefas, and I. Pitas, “Discriminant bag of words based representation for human action recognition,” *Pattern Recognit. Lett.*, vol. 49, pp. 185–192, 2014.
[66] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva, and A. C. Kot, “Skeleton-based human action recognition with global context-aware attention LSTM networks,” *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, 2018.
[67] F. Gu, F. Flórez-Revuelta, D. Monekosso, and P. Remagnino, “A Multiple Kernel Learning Based Fusion Framework for Real-Time Multi-View Action Recognition,” in *International Workshop on Ambient Assisted Living*, 2014, pp. 26–33.
[68] R. Khemchandani and S. Sharma, “Robust least squares twin support vector machine for human activity recognition,” *Appl. Soft Comput.*, vol. 47, pp. 33–46, 2016.
[69] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, “Human activity recognition using multi-features and multiple kernel learning,” *Pattern Recognit.*, vol. 47, no. 5, pp. 1800–1812, 2014.
[70] M. H. Kolekar and D. P. Dash, “Hidden Markov Model based human activity recognition using shape and optical flow based features,” in *Region 10 Conference (TENCON), 2016 IEEE*, 2016, pp. 393–397.

[71] T. Lu, L. Peng, and S. Miao, “Human Action Recognition of Hidden Markov Model Based on Depth Information,” in *Parallel and Distributed Computing (ISPDC), 2016 15th International Symposium on*, 2016, pp. 354–357.
[72] H. Rahmani, A. Mian, and M. Shah, “Learning a deep model for human action recognition from novel viewpoints,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.
[73] R. Gross and J. Shi, “The cmu motion of body (mobo) database,” 2001.
[74] I. Laptev, M. Marsza\lek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” 2008.
[75] S. Singh, S. A. Velastin, and H. Ragheb, “Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods,” in *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010, pp. 48–55.
[76] J. D. Shutler, M. G. Grant, M. S. Nixon, and J. N. Carter, “On a large sequence-based human gait database,” in *Applications and Science in Soft Computing*, Springer, 2004, pp. 339–346.
[77] J. Liu, J. Luo, and M. Shah, “Recognizing realistic actions from videos in the wild,” 2009.
[78] M. Rodriguez, “Spatio-temporal maximum average correlation height templates in action recognition and video summarization,” 2010.
[79] C.-C. Chen, M. S. Ryoo, and J. K. Aggarwal, *UT-Tower dataset: aerial view activity classification challenge*. Online, 2010.
[80] R. Kavi and V. Kulathumani, “Real-time recognition of action sequences using a distributed video sensor network,” *J. Sens. Actuator Netw.*, vol. 2, no. 3, pp. 486–508, 2013.
[81] N. M. Oliver, B. Rosario, and A. P. Pentland, “A Bayesian computer vision system for modeling human interactions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 831–843, 2000.
[82] C.-W. Chen and H. Aghajan, “Multiview social behavior analysis in work environments,” in *Distributed Smart Cameras (ICDSC), 2011 Fifth ACM/IEEE International Conference on*, 2011, pp. 1–6.
[83] D. Das, “Activity recognition using histogram of oriented gradient pattern history,” *Int. J. Comput. Sci. Eng. Inf. Technol.*, vol. 4, no. 4, pp. 23–31, 2014.
[84] P. Antonakaki, D. Kosmopoulos, and S. J. Perantonis, “Detecting abnormal human behaviour using multiple cameras,” *Signal Process.*, vol. 89, no. 9, pp. 1723–1738, 2009.
[85] K. Fujimura, Y. Yoshimitsu, T. Naito, and S. Kamijo, “Behavior understanding at railway station by postures and the pseud-trellis analysis of trajectories,” in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, 2010, pp. 1116–1122.
[86] S. Calderara, U. Heinemann, A. Prati, R. Cucchiara, and N. Tishby, “Detecting anomalies in people’s trajectories using spectral graph analysis,” *Comput. Vis. Image Underst.*, vol. 115, no. 8, pp. 1099–1111, Aug. 2011, doi: 10.1016/j.cviu.2011.03.003.
[87] F. Chen and A. Cavallaro, “Detecting group interactions by online association of trajectory data,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 1754–1758.
[88] O. Ouyed and M. S. Allili, “Recognizing Human Interactions Using Group Feature Relevance in Multinomial Kernel Logistic Regression.,” in *FLAIRS Conference*, 2018, pp. 541–546.
[89] W. Choi, K. Shahid, and S. Savarese, “What are they doing?: Collective activity classification using spatio-temporal relationship among people,” in *Computer Vision*

Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, 2009, pp. 1282–1289.

[90] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 3273–3280.

[91] W. Choi and S. Savarese, "Understanding collective activitiesof people from videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 6, pp. 1242–1257, 2014.

[92] Y. Kong and Y. Fu, "Close human interaction recognition using patch-aware models," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 167–178, 2016.

[93] W. Ahmed and M. H. Yousaf, "Robust Activity Recognition Model Via Motion Templates," in *2018 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*, 2018, pp. 1–6.

[94] O. Brdiczka, M. Langet, J. Maisonnasse, and J. L. Crowley, "Detecting human behavior models from multimodal observation in a smart home," *IEEE Trans. Autom. Sci. Eng.*, vol. 6, no. 4, pp. 588–597, 2009.

[95] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 1, pp. 91–101, 2013.

[96] "Image and Signal Processing Group (UCL) | Softwares / APIDIS browse." [Online]. Available: https://sites.uclouvain.be/ispgroup/Softwares/APIDIS. [Accessed: 15-Sep-2019].

[97] Y. Kong, Y. Jia, and Y. Fu, "Learning human interaction by interactive phrases," in *European conference on computer vision*, 2012, pp. 300–313.

[98] R. B. Fisher, "The PETS04 surveillance ground-truth data sets," in *Proc. 6th IEEE international workshop on performance evaluation of tracking and surveillance*, 2004, pp. 1–5.

[99] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2556–2563.

[100] "Page for JAIST Multi-View Surveillance (MVS) Video Database." [Online]. Available: http://www.jaist.ac.jp/~chen-fan/multivision/jaistmvsdb.html. [Accessed: 16-Sep-2019].

[101] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, "Two-person interaction detection using body-pose features and multiple instance learning," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 28–35.

[102] A. Patron-Perez, M. Marszalek, I. Reid, and A. Zisserman, "Structured learning of human interactions in TV shows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 12, pp. 2441–2453, 2012.

[103] E. Acar, F. Hopfgartner, and S. Albayrak, "Understanding affective content of music videos through learned representations," in *International Conference on Multimedia Modeling*, 2014, pp. 303–314.

[104] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng, "An enhanced deep feature representation for person re-identification," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, 2016, pp. 1–8.

[105] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 98–106.

[106] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, “Combining deep learning and hand-crafted features for skin lesion classification,” in *Image Processing Theory Tools and Applications (IPTA), 2016 6th International Conference on*, 2016, pp. 1–6.

[107] B. Chen, B. Marlin, and N. de Freitas, “Deep Learning of Invariant Spatio- Temporal Features from Video,” in *NIPS 2010 Deep Learning and Unsupervised Feature Learning Workshop*, 2010.

[108] V. Dutta and T. Zielinska, “Predicting human actions taking into account object affordances,” *J. Intell. Robot. Syst.*, vol. 93, no. 3–4, pp. 745–761, 2019.

[109] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, “Activity forecasting,” in *European Conference on Computer Vision*, 2012, pp. 201–214.

[110] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, and A. Oliver-Albert, “A predictive model for recognizing human behaviour based on trajectory representation,” in *Neural Networks (IJCNN), 2014 International Joint Conference on*, 2014, pp. 1494–1501.

[111] K. Xu, Z. Qin, and G. Wang, “Human activities prediction by learning combinatorial sparse representations,” in *Image Processing (ICIP), 2016 IEEE International Conference on*, 2016, pp. 724–728.

[112] J. S. Hartford, J. R. Wright, and K. Leyton-Brown, “Deep learning for predicting human strategic behavior,” in *Advances in Neural Information Processing Systems*, 2016, pp. 2424–2432.

[113] “Datasets - visint.org.” [Online]. Available: http://www.visint.org/datasets. [Accessed: 17-Sep-2019].

[114] E. M. Tapia, S. S. Intille, and K. Larson, “Activity recognition in the home using simple and ubiquitous sensors,” in *International conference on pervasive computing*, 2004, pp. 158–175.

[115] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition.,” in *CVPR*, 2008, vol. 1, p. 6.

[116] L. Sun, H. Ai, and S. Lao, “Localizing activity groups in videos,” *Comput. Vis. Image Underst.*, vol. 144, pp. 144–154, 2016.

[117] F. Ahmed, E. Hossain, A. Bari, and M. S. Hossen, “Compound local binary pattern (clbp) for rotation invariant texture classification,” *Int. J. Comput. Appl.*, vol. 33, no. 6, pp. 5–10, 2011.

[118] A. Hafiane, G. Seetharaman, and B. Zavidovique, “Median binary pattern for textures classification,” in *International Conference Image Analysis and Recognition*, 2007, pp. 387–398.

[119] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[120] C. Cortes and V. Vapnik, “Support-vector networks,” *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[121] J. Zhang, M. Marsza\lek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, 2007.

[122] R. Rifkin and A. Klautau, “In defense of one-vs-all classification,” *J. Mach. Learn. Res.*, vol. 5, no. Jan, pp. 101–141, 2004.

[123] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, “A survey of decision fusion and feature fusion strategies for pattern classification,” *IETE Tech. Rev.*, vol. 27, no. 4, pp. 293–307, 2010.

[124] C. G. Snoek, M. Worring, and A. W. Smeulders, “Early versus late fusion in semantic video analysis,” in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.

[125] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[126] W. Wang, Y. Yan, L. Nie, L. Zhang, S. Winkler, and N. Sebe, “Sparse code filtering for action pattern mining,” in *Asian Conference on Computer Vision*, 2016, pp. 3–18.

[127] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, “View-independent action recognition from temporal self-similarities,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 172–185, 2011.

[128] L. Xu, A. Krzyzak, and C. Y. Suen, “Methods of combining multiple classifiers and their applications to handwriting recognition,” *IEEE Trans. Syst. Man Cybern.*, vol. 22, no. 3, pp. 418–435, 1992.

[129] J. K. Aggarwal and M. S. Ryoo, “Human activity analysis: A review,” *ACM Comput. Surv. CSUR*, vol. 43, no. 3, p. 16, 2011.

[130] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah, “High-level event recognition in unconstrained videos,” *Int. J. Multimed. Inf. Retr.*, vol. 2, no. 2, pp. 73–101, 2013.

[131] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Trans. Intell. Syst. Technol. TIST*, vol. 2, no. 3, p. 27, 2011.

[132] S. Blunsden, E. Andrade, and R. Fisher, “Non parametric classification of human interaction,” in *Iberian Conference on Pattern Recognition and Image Analysis*, 2007, pp. 347–354.

[133] J. Azorin-Lopez, M. Saval-Calvo, A. Fuster-Guillo, and J. Garcia-Rodriguez, “A novel prediction method for early recognition of global human behaviour in image sequences,” *Neural Process. Lett.*, vol. 43, no. 2, pp. 363–387, 2016.

[134] M. S. Ryoo, “Human activity prediction: Early recognition of ongoing activities from streaming videos,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 1036–1043.

[135] W. Y. Kwon, Y. Park, S. H. Lee, and I. H. Suh, “Human Activity Recognition Using Deep Recurrent Neural Networks and Complexity-based Motion Features.”

[136] Z. Ge *et al.*, “Exploiting temporal information for DCNN-based fine-grained object classification,” in *Digital Image Computing: Techniques and Applications (DICTA), 2016 International Conference on*, 2016, pp. 1–6.

[137] C. A. Caetano, V. H. C. De Melo, J. A. dos Santos, and W. R. Schwartz, “Activity Recognition based on a Magnitude-Orientation Stream Network,” in *Graphics, Patterns and Images (SIBGRAPI), 2017 30th SIBGRAPI Conference on*, 2017, pp. 47–54.

[138] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[139] G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[140] T. Lan, T.-C. Chen, and S. Savarese, “A hierarchical representation for future action prediction,” in *European Conference on Computer Vision*, 2014, pp. 689–704.

[141] O. M. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition.,” in *BMVC*, 2015, vol. 1, p. 6.

[142] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, “Unravelling robustness of deep learning based face recognition against adversarial attacks,” *ArXiv Prepr. ArXiv180300401*, 2018.

[143] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[144] I. Omara, X. Wu, H. Zhang, Y. Du, and W. Zuo, “Learning pairwise SVM on hierarchical deep features for ear recognition,” *IET Biom.*, 2018.

[145] Z. Zhu *et al.*, “Deep learning-based features of breast MRI for prediction of occult invasive disease following a diagnosis of ductal carcinoma in situ: preliminary data,” in *Medical Imaging 2018: Computer-Aided Diagnosis*, 2018, vol. 10575, p. 105752W.

[146] A. B. Sargano, P. Angelov, and Z. Habib, “A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition,” *Appl. Sci.*, vol. 7, no. 1, p. 110, 2017.

[147] “ImageNet.” [Online]. Available: http://www.image-net.org/. [Accessed: 10-Jul-2018].

[148] B. K. Horn and B. G. Schunck, “Determining optical flow,” *Artif. Intell.*, vol. 17, no. 1–3, pp. 185–203, 1981.

[149] K. Wainwright, *Fundamental methods of mathematical economics*. McGraw-Hill, 2005.

[150] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, “Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, 2017.

[151] N. Wanas, *Feature-based architectures for decision fusion*. University of Waterloo [Department of Systems Engineering], 2003.

[152] P. Dollár, R. Appel, S. Belongie, and P. Perona, “Fast feature pyramids for object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, 2014.

[153] M. S. Ryoo and J. K. Aggarwal, “UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA),” in *IEEE International Conference on Pattern Recognition Workshops*, 2010, vol. 2, p. 4.