Pose-invariant action recognition for automated behaviour analysis

Manoj Ramanathan

2017

Manoj Ramanathan. (2017). Pose-invariant action recognition for automated behaviour analysis. Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/70099>

<https://doi.org/10.32657/10356/70099>

Pose-Invariant Action Recognition for Automated Behaviour Analysis

Manoj Ramanathan

School of Electrical and Electronic Engineering

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

2016

Statement of Originality

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

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

MANOJ RAMANATHAN

Date

To the all mighty and my loving family

Acknowledgements

Thank you God for the countless blessings!

First of all, I would like to express my sincere thanks and gratitude to my supervisors Associate Professor Teoh Eam Khwang and Dr. Yau Wei-Yun, who have constantly guided, supported and encouraged me throughout the duration of research study. Their technical insights, suggestions and experiences have helped me both academically and personally. Without their constant support and research direction, this research would not have been possible.

I would like to thank staff members of Internet of Things Lab, Mr Sow Peck Heng and Mr Yuen Sien Huan, at the School of Electrical and Electronic Engineering, Nanyang Technological University for their technical support and help. I would like to thank all staff members of Institute for Infocomm Research (I2R), A*STAR for their kind support and assistance. Many thanks to all the students who helped me to create the dataset. I would like to thank all my friends, who have encouraged, supported and provided insights.

I'd like to thank my family, relatives and friends who have so far stood by me. Through their love, my parents have encouraged, supported and guided me in this whole life and research. They have been my source of strength and determination throughout my life. I am forever indebted to them. I would like to thank my brother who has been a source of inspiration and motivation to me.

Finally, I would like to acknowledge the School of Electrical and Electronic Engineering, Nanyang Technological University and Institute for Infocomm Research (I2R) for providing the research facilities for my work. Also I would like to thank the school of Electrical and Electronic Engineering, Nanyang Technological University for awarding me research scholarship during the study.

Summary

Computer vision deals with providing visual capabilities to a computer so that it can understand its surrounding environment. This has given rise to numerous applications such as human-computer interactions, object detection, scene understanding, surveillance etc. Understanding human behaviour is a difficult task due to its fickle nature. A person's action is one of the visible clues towards one's behaviour or intention. Due to this, research in action recognition has gained momentum in recent years.

Actions can be mainly characterized by the motion of body parts and evolution of the pose of the person. In an unconstrained setting, achieving a clear and distinct action representation is very difficult due to pose variations caused as a result of variation in camera view angle changes, occlusion of body parts, cluttered background etc. Also, traditional action recognition datasets portray actions that have subjects mainly in upright postures such as walking, running, boxing etc. There is lack of study related non-upright posture action recognition. Thus in this research, we develop a novel pose-invariant action recognition which combines both motion of body parts and pose of the person in a mutually reinforcing manner. The proposed method is generalized to handle possible human posture variations (including non-upright postures). The assumption made is that the neck point of the person and major viewing direction of the person during motion are available in the frame.

Our pose-invariant action recognition framework is inspired from closed loop framework of control system theories with two components, namely, pose-invariant motion feature extraction component (Forward path) and canonical pose component (Feedback path) with the output of each used as input to the other to reinforce the initial prediction in a manner that improves the overall final result. The two components of our proposed action recognition framework are

- **Pose-Invariant Motion Component or Propagation Motion Forward (PMF) Path**

In this component, we aim to capture motion of different body parts that characterizes an action. To achieve this, a propagation mechanism is used to divide the region of interest is divided into 3 grids corresponding to the head, torso and legs respectively based on the estimated body orientation of the person in the frame. To capture the motion in each grid, we use a set of kinematic features derived from optical flow and encode them in a pose-invariant manner by converting them to a human-body centric space. Using these initial set of extracted invariant kinematic motion features, we provide initial hypothesis of possible actions.

- **Canonical Pose Component or Canonical Pose Feedback (CPF) Path**

The initial action recognized based on invariant kinematic motion features are used as input by this component. Each action is characterized by specific poses, we call these as canonical stick poses. Given an action and available training videos, we propose an algorithm to extract normalized canonical stick poses that can be compared with the test videos. For comparing these stick poses with the visual data, we propose a novel pose hypothesis generation scheme that compares each of the extracted canonical sticks of the action recognized by the first component with the video frame to identify the most likely canonical stick pose in it. The identified canonical stick pose in the frame is used to improve the pose-invariant motion feature extraction by realigning the grids used in the first component and also compute another kinematic motion feature of each body part with respect to the center of that body part. To capture temporal dynamics of action, identified stick poses are used to compute a new set of temporal stick features. The new set of pose-invariant motion features extracted from both forward and feedback paths using the realigned grids serves as the final action representation. Using these features, action hypothesis is refined until the action converges to yield the recognized action.

In our canonical pose feedback path, to compare each canonical stick pose with the test frame, we propose a canonical pose hypothesis scheme that uses a body part

detector as it's basis. Most body part detectors or person detectors available work based on the assumption that the person is in an upright posture. Therefore, we propose our own body part detector that can work even when the person is in non-upright postures. In our body part detector, we do not impose any connectivity or shape constraints between the parts allowing us to detect body parts in non-upright postures. We divide the human body into 4 body parts head, torso, arm and leg. For detecting these body parts, training images were collected from the web using a python crawler. Using the training images, body part classifiers were trained. Based on the Bayes' theorem for conditional probability, we propose an algorithm to compute the likelihood scores for each of the detected parts in the frame. The algorithm uses the developed classifier's output to determine a likelihood score for the detected body parts. This likelihood score is used by the pose hypothesis generation scheme of the canonical pose component.

To test our method's effectiveness under non-upright posture, we introduce a new action dataset comprising of 8 actions in different postures captured from 3 views. The dataset comprises of visual data for 35 subjects. We have conducted several experiments on publicly available benchmark action datasets such as Weizmann, KTH, UCF Sports, Hollywood, HMDB51 and this new dataset to test the effectiveness of our recognition framework. The results show that the proposed reinforcement framework for action recognition is pose-invariant, partially view-invariant and is able to work even if there is partial occlusion of the person performing the action.

Contents

Acknowledgement	i
Summary	ii
List of Figures	ix
List of Tables	xi
List of Abbreviations	xii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Challenges	3
1.3 Objectives	5
1.4 Our Contributions	6
1.5 Organization of the Thesis	8
2 A Literature Review	11
2.1 Challenges in Vision-based Action Recognition	12
2.1.1 Variation in View point	12
2.1.2 Occlusion	16
2.1.3 Execution Rate	18
2.1.4 Anthropometric Variations	20
2.1.5 Camera Motion	23
2.1.6 Cluttered Background	25
2.2 Action Classification	27
2.2.1 Generalizability	28
2.2.2 Abnormality Detection	30
2.2.3 Classifier Training	31

2.3	Deep Learning Approaches	33
2.4	Datasets	36
2.5	Concluding Remarks	38
3	Forward Path for Pose-Invariant Action Recognition	40
3.1	Introduction	40
3.2	Overview of Complete Framework	42
3.3	Propagation Motion Forward Path	43
3.3.1	Propagation Mechanism	44
3.3.2	Kinematic Feature Computation	45
3.3.3	Human-body Centric Space	46
3.4	Experimental Results and Discussion	49
3.4.1	Weizmann Dataset	50
3.4.2	KTH Dataset	51
3.4.3	UCF Sports Dataset	52
3.4.4	Hollywood Dataset	54
3.4.5	HMDB51 dataset	55
3.4.6	Discussion	57
3.5	Non-Upright Action Dataset (NUAD)	58
3.6	Concluding Remarks	60
4	Likelihood Score Computations for Body Part Detection	61
4.1	Introduction	61
4.2	Related work	62
4.3	Body part classifiers	64
4.3.1	Collection of Training Images	65
4.3.2	Feature Extraction and Classifier training	65
4.4	Body Part Detection Framework	68
4.4.1	Foreground Segmentation	68
4.4.2	Likelihood score computation	71
4.4.3	Computation of Prior probabilities	72
4.4.4	Inter-frame Voting Strategy	73
4.5	Experimental Results and Discussion	75
4.6	Concluding Remarks	77
5	Feedback Path for Pose-Invariant Action Recognition	79
5.1	Introduction	79

5.2	Canonical Pose Feedback path	80
5.2.1	Canonical Stick Pose Generation and Representation	80
5.2.2	Canonical Stick Pose Hypothesis	82
5.2.3	Computation of Body Part Referenced Kinematic Features .	85
5.3	Temporal Stick Features	87
5.4	Experimental Results and Discussion	89
5.4.1	Weizmann Dataset	90
5.4.2	KTH Dataset	90
5.4.3	UCF Sports Dataset	92
5.4.4	Hollywood Dataset	96
5.4.5	HMDB51 dataset	98
5.4.6	NUAD	101
5.4.7	MSR Action Dataset	102
5.4.8	Discussion	103
5.5	Concluding Remarks	113
6	Conclusion and Recommendations for Future Work	115
6.1	Conclusion	115
6.2	Recommendations for Future Work	118
6.2.1	Automatic Estimation of Viewing Direction and Reference Points (Neck)	118
6.2.2	Canonical Stick Pose Extraction and Availability	119
6.2.3	Stick Post Processing - Handling Body Part Detection Errors	120
6.2.4	Deep learning approaches	120
6.2.5	Action Similarity Measures	121
	Author's Publications	122
	Bibliography	124

List of Figures

2.1	Multi-view images from i3DPost [1] dataset of walking action	13
3.1	Proposed pose-invariant action recognition framework. Each video passes through PMF and CPF paths. In PMF path, based on neck point and foreground, body is divided into head, torso, leg grids and kinematic features are computed and re-aligned in the human body centric space. Using these features, an initial action hypothesis is estimated. In CPF path, canonical pose hypothesis scheme uses canonical stick poses for initial estimated action to determine the pose in the frame. Using the estimated pose, two additional kinematic features are computed (re-aligned in human body centric space) and grid division is corrected. This forms the new set of features for action recognition. The loop is repeated until the recognition converges to an action	41
3.2	Proportionality of body parts used in Propagation Mechanism. Sample Images from Weizmann [2] and UCF Sports [3]	44
3.3	Neck points marked for sample dataset images	46
3.4	Human-body centric space cylinder	47
3.5	Division of grids into smaller cells & Labelling of cells based on the change in view, body orientation, Only labelling of Grid 2 cells are shown	48
3.6	Incorrect head size determined by propagation mechanism	51
3.7	Confusion matrix for UCF Sports dataset with only PMF path . . .	53
3.8	Sample Images from NUAD dataset	59
4.1	Sample training images collected using web crawler for arm, head, leg and torso	66
4.2	Proposed Body part detection framework	68

4.3	(a) Sample video frame (b) Averaged background subtraction (c) Mean shift segments generated	69
4.4	Consistent Motion Segments identified in 2 frames (a) Waving action (b) Walking action, black is the background region, rest denote different identified segments	70
4.5	Arms detected in a frame with their respective likelihood scores (ELM is used)	73
4.6	Body part detection results obtained using (a) ELM (b) DBN. . . .	78
4.7	Erroneous body part detections (a) Partial Head Occlusion (b) Variation in appearance of Torso and Head	78
5.1	Canonical stick pose marked on the obtained canonical pose image .	81
5.2	Canonical Sticks extracted for different actions in different datasets	83
5.3	Computation of Temporal Stick Features from video	88
5.4	Poses identified using all canonical sticks (Top) and using lesser canonical sticks (Bottom), <i>Head - Green, Torso -Yellow, Arms - Blue & Red, Legs - Cyan & Magenta</i>	94
5.5	UCF Sports dataset Confusion Matrix (a) Complete Framework (PMF + CPF) with all canonical sticks (b) Complete Framework with lesser canonical sticks	95
5.6	No motion & neck visible in Hollywood dataset	97
5.7	Poses identified in Hollywood dataset when part of body is not visible or occluded, <i>Head - Green, Torso -Yellow, Arms - Blue & Red</i>	99
5.8	Poses identified in HMDB51 using available canonical sticks, <i>Head - Green, Torso -Yellow, Arms - Blue & Red, Legs - Cyan & Magenta</i>	101
5.9	Pose identified in CPF path with (a) Groundtruth Neck (b) Tolerance Range Neck (c) Outside Tolerance Range neck	105
5.10	Groundtruth & Tolerance Range for Neck reference position	110
5.11	Erroneous stick pose identified by CPF path	111

List of Tables

2.1	Classification of methods robust to anthropometric variations based on features	22
2.2	Classification of Approaches based on Robustness to challenges, Generalizability, Abnormality Detection	29
3.1	Pose-invariant Motion Feature Extraction in PMF path	49
3.2	Performance of PMF path in different datasets	50
3.3	Required Grids for actions in Hollywood dataset	55
3.4	Performance of PMF path in NUAD	59
4.1	4-Fold Cross validation results using LBP, LGP and HOG features with ELM	67
4.2	Accuracy (in %) Comparison of 4-Fold Cross validation results between ELM and DBN	67
4.3	Identifying Consistent Motion Segments Algorithm	70
4.4	Performance (in%) comparison of the four different variations mentioned, ELM Mean, DBN Mean, ELM Motion and DBN Motion . .	76
5.1	Canonical Pose Generation for a Given Action	82
5.2	Canonical Pose Hypothesis for a Given Frame	86
5.3	Temporal Dynamics of Stick Pose Joint and Body part	89

5.4	Performance Comparison for Weizmann dataset, Proposed method uses foreground masks given with dataset	91
5.5	Performance Comparison for KTH dataset	92
5.6	Performance Comparison for UCF Sports dataset, Proposed method uses manually marked bounding box	93
5.7	Effect of feature in pose-invariant action recognition using UCF sports dataset (a) div + curl, (b) proj + rot, (c) BodyProj + Body- Rot, (d) temporal stick features	96
5.8	Performance Comparison for Hollywood dataset, Proposed method uses manually marked bounding box	98
5.9	Performance Comparison on HMDB51 dataset, proposed method run on subset of actions, others are run on whole dataset	98
5.10	Performance Comparison for HMDB51 dataset for General Body Movements category, Proposed method uses manually marked bound- ing box	102
5.11	Performance of Proposed framework in NUAD, Proposed method uses ground truth person location	102
5.12	Action Recognition Performance in MSR Action dataset, Proposed method uses manually marked bounding box	103
5.13	Experiments for various cell configurations in Weizmann dataset . .	104
5.14	Effect of varying τ_d and τ_θ in Canonical pose hypothesis using Weiz- mann dataset	105
5.15	Performance of proposed action recognition framework with pro- posed view estimation algorithm in Weizmann dataset	110
5.16	Pose estimation error due to body part detection and canonical pose hypothesis in subset of Weizmann dataset	112

List of Abbreviations

AdaBoost	Adaptive Boosting
API	Application Programming Interface
BoW	Bag of Words
CNN	Convolutional Neural Networks
CPF	Canonical Pose Feedback
DBN	Deep Belief Networks
Div	Divergence
ELM	Extreme Learning Machines
h-LDA	Hybrid latent Dirichlet allocation
HMM	Hidden Markov Models
HoF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradients
HoMB	Histogram of Motion Boundaries
IXMAS	INRIA Xmas Motion Acquisition Sequences
LBP	Local Binary Pattern
LGP	Local Gradient Pattern
MACH	Maximum Average Correlation Height
McFIS	Meta Cognitive Neuro-Fuzzy Inference System
MEI	Motion Energy Image

MHI	Motion History Image
MuHAVi	Multicamera Human Action Video Data
NUAD	Non-Upright Action Dataset
OF	Optical Flow
PCA	Principal Component Analysis
PMF	Propagation Motion Forward
PNF	past-now-future
Proj	Projection
RFID	Radio-Frequency Identification
RMS	Root Mean Square
Rot	Rotation
SIFT	Scale-invariant Feature Transform
SVM	Support Vector Machines
UCF	University of Central Florida
ViHASi	Virtual Human Action Silhouette

Chapter 1

Introduction

1.1 Background and Motivation

Mankind's ability to think, evolve and innovate has paved way to several historical inventions of great significance. This inherent ability of mankind to adapt and innovate creatively is also what makes them a complex and complicated entity to interpret/ understand. Behavioural analysis can be regarded as a psychological study that aims to understand human and make meaningful decisions or predictions based on the same. In the context of this thesis, behaviour of a person can be viewed as the way in which he/she responds or reacts in a given scenario or specific stimulus. Behaviour of a person is considered as a higher order neurological function that involves making meaningful interpretation of the situation observed and assesses how to respond. In general, such human behaviour is very complex and fickle in nature making it more difficult to understand. For instance, what will a person do when he sees a football on the road. One might just walk off without doing anything or one might join the game being played nearby or one can just take the ball with himself back home. These are only a few ways one

might react in a situation. There are several situations in which behaviour of a person cannot be understood or predicted like the loss of loved one etc. It is a very complex topic to analyse as these are governed by several factors such as attitude, mood of person, previous life experiences etc which are very difficult to quantify. Understanding human behaviour in a variety of situations will help computers to make meaningful, useful decisions and predictions.

Over the years, computer vision has received much research attention as it aims to analyse and interpret visual data to help in understanding the environments, people in the field of view etc. With recent advances in camera technology, smart phones, cloud storage and multimedia, a whole lot of visual data is at the disposal of everyone. Due to this computer vision research is much more necessary and important. The goal of several computer vision related research works such as object recognition, pedestrian detection, interpretation of scene etc is to make the computer understand it's operating environment. One important aspect of every environment is the human-computer interaction. Thus several vision based research works such as pedestrian detection, emotion recognition, action recognition etc are focused on people in the operating environment. For instance, [4, 5] deal with robots built to function with patients and elderly people. These robots try to understand human interactions and behaviour to help them with daily chores.

Even though behaviour of a person is not directly visible, there are a few visual clues left behind like action or emotion of a person. Actions are a direct indication or visible effect of a person's behaviour. For instance, when the person move towards a coffee maker, we can say he/she intends to have a coffee. This has given rise to a vast research community focusing on action recognition from visual data. Several applications such as surveillance, human-computer interaction, interactive smart classrooms, healthcare industry would benefit from recognition of action performed by a subject of interest. Actions can be defined as movements of one

or more body parts in a certain way with the intention of completing a task. This might involve interactions with objects and other humans. But action recognition in an unconstrained setting is very difficult due to several challenges such as view angle changes, occlusion, cluttered background etc. One main challenge in action recognition is the amount of pose variations observed as a result of the change in camera view, occlusion or anthropometric variations. Thus developing a pose-invariant action recognition method is of prime importance.

In most of the traditional action recognition datasets and methods, focus has mainly been on actions performed by a person in upright posture. But human body is articulated in nature and can be in a variety of postures. As far as we are aware of, there is no research study yet on action recognition in such non-upright postures. In this thesis, we aim to develop a novel pose-invariant action recognition framework by combining motion and pose clues in a mutually reinforcing manner that can handle challenges like pose variations, occlusion etc. In this method, we aim to capture the motion of each body part in a pose-invariant manner and recognize actions in any posture (upright and non-upright).

1.2 Challenges

Human body is articulated in nature with several degrees of freedom that allows each body part to move independently of each other. Due to this articulation, human body is capable of performing actions in a variety of postures. And for actions that involve human-human or object interactions, the pose or movement of body parts is determined by the other person or object location. Due to all this, actions captured in videos or images exhibit huge amount of pose variations. Apart from this, when actions are performed in a realistic environment it is not necessary for them to be in upright posture, which means action recognition methods must

be able to cater for other postures as well.

Actions are mainly characterized by the motion of body parts and unique poses observed. With this idea, several researches are based on either motion [6] or pose/appearance clues [7,8] or a combination of both [9]. Besides the main challenge of pose variations, several other challenges in automatic action recognition include view angle variations, occlusion, anthropometric variations, speed of executing an action, camera motion, cluttered background etc. These factors directly affect either motion or pose that would be visible in the video, making action recognition very difficult. Factors such as view angle variation, occlusion, anthropometric variations and cluttered background also contribute to pose variations observed in a video. For instance, the change in view would change the pose visible. Thus it requires methods to extract view-invariant features or view-specific classification [10,11]. Self-occlusion or occlusion due to nearby objects severely affect action recognition as they alter appearance of a person or hide motion of body parts. Anthropometric variations specifically deals with how an action being performed varies from one person to another. In addition to these, videos captured in the wild tend to have cluttered background and moving cameras making it harder to classify actions. Thus in order to handle pose variations in an action recognition framework, it would be necessary to extract discriminative action representation that is capable of handling the above mentioned challenges as well.

Another important aspect of recognition framework is its ability to learn actions. The system's ability to learn an action is directly related to the discriminative capacity of its features and number of training videos available for learning. Another important consideration in recognition, is how well can it be extended to other actions for which training videos were not provided. This can be termed as generalizability of the framework. Ability to generalize for different actions determine how well it can be used for applications like surveillance, abnormal action

detection etc. Learning an action with very few training videos and be able to cater for the above mentioned challenges is an unsolved problem.

As mentioned above, actions can be performed in a variety of postures. But available benchmark datasets focus on actions being performed in upright posture of a person. Even though a few datasets like UCF sports [3] show a few action videos in non-upright postures, they do not test the ability of recognition framework to handle these postures completely. For example, actions like diving, Swing-Bench and Swing-side in UCF sports include both upright and non-upright postures. However several actions can be done only in non-upright actions like exercise, sports tutorial etc. Also several actions can be done in either upright or non-upright actions such as wave, turn etc. Just as observed motion varies with view point, it also varies with body posture. There is a lack of study on actions being performed in non-upright postures. Since this deals with the posture of the person, it would be necessary for pose-invariant framework to be generalized to handle these non-upright postures too.

1.3 Objectives

Performing behavioural analysis using visual data is a very challenging task as behaviour cannot be easily measured. To quantify behaviour, in this thesis, we rely on action being performed. Action recognition in realistic videos is very difficult to several challenges like pose variations, camera view, occlusion etc.

In addition, there is a lack of research study on non-upright human posture actions. In this thesis work, we have two main objectives:

1. To develop a novel pose-invariant action recognition framework that is able to handle partial occlusion, anthropometric variations and view angle changes.

2. To extend the action recognition framework to support non-upright posture actions and show that the performance attained is reasonable.

1.4 Our Contributions

Our main contributions in this thesis work are

1. Novel action recognition framework combining motion and pose cues observed in a video. Proposed reinforcement framework comprises of two components, propagation motion forward (PMF) path and canonical pose feedback(CPF) path to combine motion and pose cues. PMF path performs an initial motion based action recognition to estimate canonical sticks that will be used in the pose hypothesis stage of CPF path. The identified canonical pose in CPF path is used to realign the grids used in PMF path. Subsequently a new set of pose-invariant kinematic features that capture the motion of each body part is recomputed. The proposed framework forms a loop structure between motion and pose component that is repeated until the action recognition result converges.
2. To handle pose variations, we develop a novel propagation mechanism, that computes kinematic motion features ($Div, Curl, Proj, Rot$) based on the assumption that a reference (neck) point is available and major viewing direction of the person is available for estimated body regions. The computed features are encoded in a novel pose-invariant manner. These motion features represent action as motion of body parts with respect to the human body itself making them pose-invariant and handle different postures. The features are also computed at two stages. At the stage of motion forward path where the features are computed from motion cues with respect to esti-

mated reference point and at the pose feedback path where the features are computed for the body parts and stick pose representation. The propagation mechanism uses the natural proportionality of body parts thereby allowing us to determine how much of the body is visible and how much is occluded.

3. Novel body part detection scheme that does not impose any skeletal constraints thereby allowing us to detect body parts in non-upright postures. Our body part detection scheme provides a novel likelihood score based on Bayesian inference which can be used as a certainty value for detection. This likelihood score forms the basis of our canonical pose feedback path. To study the robustness of proposed body part detection scheme and action recognition framework under non-upright human postures, a new dataset is introduced with 8 non-upright actions performed by 35 subjects and captured in 3 views.
4. Our canonical pose feedback path begins with an initial recognized action. For each action, we devise an algorithm to generate canonical stick poses from training videos and a normalized stick representation that allows scale invariant comparison with test videos. A novel pose hypothesis generation scheme forms the core of CPF path and uses the above mentioned body part likelihood score, available canonical sticks for initial action. The scheme generates a pose hypothesis for every canonical stick to be present in the test frame and based on the hypothesis we determine the pose in frame. This scheme is able to handle partial occlusion and can generate hypothesis with available partial canonical sticks. The obtained pose is used to compute another novel set of kinematic features to describe the motion of every body part and temporal dynamics of an action that are encoded in a pose-invariant manner.

1.5 Organization of the Thesis

The rest of the thesis is organized as follows

Chapter 2 provides a literature review of existing methods in the field of action recognition. Challenges in a research domain defines the research and existing methods in the domain. A detailed discussion of each challenge and existing methods to overcome them are presented. This helps to identify gaps and unexplored areas in research so far. A brief introduction to action recognition method's learning capability and extension to different applications is also presented. Benchmark action recognition datasets are discussed with respect to the research challenges as well.

Chapter 3 introduces our pose-invariant action recognition. To begin with, an overview of the complete framework is provided. Our propagation motion forward path is explained in detail. With assumption that neck point and major viewing direction of the person is available, pose-invariant motion features are extracted. With the help of the neck point, PMF path uses the proportionality of body parts and divides the foreground into 3 grids that roughly corresponds to head, torso and legs respectively. Pose-invariant kinematic motion features are extracted for each of these grids. Using the major viewing direction of the person, the extracted features are converted to a novel human-body centric space and encoded in a view-invariant manner. Our features compute motion with respect to the body itself thereby making the framework pose-invariant. These features are used to recognize an initial action for the video. Experimental results and discussion on benchmark datasets are provided for our framework with only PMF path. Non-Upright Action Dataset (NUAD) is introduced and initial results using PMF path are provided.

Chapter 4 deals with body part detection likelihood score computations. Several

available pose detection or body part detections are limited in nature that they require the person to be in upright postures. To detect body parts in non-upright postures, a scheme is introduced with no shape or skeletal constraints imposed. Detection of body parts in low resolution visual data also poses difficulty. First, a training database of images is collected for each body part and a classifier is trained for detecting 4 body parts (head, torso, arm, leg) and 'None' classes. The scheme begins with a segmentation of the test frame and apply the body part detector to these segments. Using Bayesian inference, a body part likelihood score is defined that uses classifier's output value and 'None' classifier output as events. When a segment is classified as a body part a likelihood score is also associated to it. The detected set of segments and corresponding likelihood scores are used to detect a final location of body parts in the image. Experiments have been conducted using ELM and deep learning approaches combined with either mean-shift color or RMS error motion segmentation. Results and discussions are provided to understand the best possible combination for body part detection under posture variation and low resolution data.

Chapter 5 introduces our canonical pose feedback path in the framework. A canonical pose hypothesis generation scheme forms the basis of this component. The pose hypothesis component is in turn dependent on the body part detection likelihood score in chapter 4 and initial action recognized using features mentioned in chapter 3. Every action is characterized by a set of poses observed, termed as canonical poses. First, an algorithm to generate these poses from training videos is presented. These poses are converted to stick poses and a normalized stick representation is introduced that allows us to compare with any test frames in a scale-invariant manner. The pose hypothesis scheme uses the body part likelihood score for every segment in the test frame and all stick poses for initial recognized action and generates a hypothesis for every stick pose to be in the test frame.

Thus it determines the best fitting pose for each test frame. The estimated pose is used to compute another set of kinematic features that describe each body part's motion and also re-align the grids generated by PMF path thus completing a loop-like structure that continues until action recognition result converges. The temporal dynamics of an action are also captured in the CPF path using temporal stick features. Experimental results and discussion for the complete framework on benchmark and our own datasets are provided. Tests to show pose-invariance of the framework, insights learned and discussion are provided.

Chapter 6 concludes the thesis by summarizing the contributions made and provides recommendations for the future work.

Chapter 2

A Literature Review

Automatic action recognition has attracted quite a number of research works. Action recognition can be simply tackled using 3D markers [12] and wearable sensors [13,14] easily but their applicability in day to day activities is limited as they require user to wear markers or sensors. Action recognition using visual data has therefore been a field with lot of research potential. Even though progress has been tremendous in this domain, there are still key challenges preventing the realization of automatic action recognition such as pose-variations, view angle changes, occlusion, speed at which action is done, anthropometric variations, intra-class variations, moving cameras and cluttered backgrounds. Few important concerns in action recognition include generalizability to recognize any action, spatial and temporal action localization in videos and collection of enough training samples.

In this chapter, we look at how each of these challenges affect recognition capacity in detail and methods being proposed to tackle them. Pose variations could be a result of variation in viewpoint, occlusion, anthropometric variations, intra-class variations etc. Therefore, we do not devote a separate section for pose-variations as it will be covered while discussing other challenges. A few previous

surveys in this domain focus mainly on features and classifiers used [15, 16] or framework adopted [17]. [15] discussed how representation and classification are influenced by the challenges in the domain. Focusing only on view-invariant action recognition, [18] reviewed available methods, with focus on view-invariant pose detection, human detection and behaviour understanding. In this chapter, our objective is to understand if a challenge has been adequately solved in action recognition domains. Thus the methods are classified on how well they can handle the challenges. By looking at the available methods from the perspective of an action recognition challenge, we come to know about different ways to handle the challenges. This will help to identify main challenges that still require more research focus.

2.1 Challenges in Vision-based Action Recognition

Each of the bottlenecks in action recognition are discussed and methods proposed to handle them are outlined in this section. A direct quantitative comparison of the different methods is not possible as they employ different datasets and testing strategies. But certain criteria such as training data required and generalizability of the method to detect any actions can be used to compare. To use any action recognition method for an application, it would be necessary to compare available methods using proper benchmark criterion.

2.1.1 Variation in View point

Figure 2.1 shows a few example images taken from multiple views of the same action. As shown in the figure, the person's posture and location varies in each

view. Apart from that, motion features or patterns obtained from each view would be different making action recognition considerably difficult. Thus for proper action recognition most methods assume a fixed camera view.

Figure 2.1: Multi-view images from i3DPost [1] dataset of walking action

Using multiple camera views to train a classifier ([6, 19–26]), is the simplest solution to view angle change. The entire space of viewing angle is divided into evenly spaced segments. A separate camera captures each segments to provide visual data. Either separate classifiers for each view point or a single classifier that handles changes in view point can be trained to recognize actions. View-dependent representations of body poses are obtained by dividing the viewpoints into segments and this results in limitations during implementation [18].

Such method attempts to solve view angle variations by extending the solution of one view to enough number of views that will encapsulate the entire set of views. The method's performance are entirely dependent on the type of feature extracted and classifier used for training. MHI and MEI [19], [6] capture human motion in the video volume. They are susceptible to movement of background objects or other people in the camera's view. Methods like [21–23] for correct detection

require exact extraction of silhouettes. Assuming that foreground or silhouettes are available limits applicability of methods. Thus these methods are tested on datasets with controlled environment setting with simple backgrounds, no illumination changes etc. This limits their usage in a normal environment. Another important concern is the complexity and computation time of these methods as multiple cameras have to be synchronized and processed together.

By modelling the body posture either 3D or 2D observable in an action ([7, 21, 23, 27–38]) a view invariant representation can be obtained. Articulated nature of human body makes it highly deformable in nature. The movement of each body part is controlled by several degrees of freedom, kinematic and skeletal constraints of the body. As the previous method, features or representation used to represent the 3D or 2D posture model directly determines the performance. For example, visual hulls ([33, 34]), envelope shape ([36–38]) and silhouettes ([21, 23]) have been used. Some researchers [29, 32] have also proposed detecting separate body parts and then combining the results to obtain the body model.

With modelling approaches, there is a requirement to track posture changes and create models to reflect these changes. To track body parts methods like object interactions [30], Kalman filtering [21], motion and appearance model [32] have been employed. Tracking based on contextual clues like objects are applicable only to a small number of action classes where the objects are already known. To track the dynamics of an action [7, 31] use a selected set of key frames that exhibit maximum pose variations. However, action representation in such cases is entirely dependent on key frames chosen and defining a criterion for choosing these key frames are not easy. Also dynamics of the complete action needs to be encapsulated by these key frames.

For performing different actions and complex movements, the human body needs

to be flexible. Articulated nature of human body comprising of several degrees of freedom allows for such movements. But creating a model that can comprehensively capture these movements from just images/videos is very difficult since it poses a problem in a very high dimensional space. Any methods with lower dimensional representations such as [27] (5 body joints), [32](10 body parts), [29] (5 body parts) would be an approximation and fail to model complex movements. Also, an unique representation cannot be obtained with [27], which shows results only on walking action.

Lastly, researchers try to use view-invariant features or features cast in a space or model where view-invariant matching is possible ([2, 35, 39–49]). Temporal order of action sequences [46], temporal self-similarities and dissimilarities of an action [45], Radon transform and Radon transform surface ([47, 49]) and fundamental ratios for fixed cameras [35] have been used successfully for recognition of certain limited actions. These methods are also limited based on the features. Actions are characterized using only shape information in [47, 49]. Since motion or temporal dynamics are not considered, differentiating similar actions like running, jogging etc would be difficult as subtle differences are not captured. The descriptor used in [45] is not strictly view invariant but experiments have been conducted to show that they can handle large view angle changes. Fundamental ratios, which are the ratios among the elements in the upper left 2x2 sub matrix of the fundamental matrix of a camera, and is invariant to camera parameters are used in [35]. But this method is limited because it identifies actions by looking for similar planar motions from varying viewpoints.

The popular choices for view-invariant matching are Fourier transform and cylindrical co-ordinate systems ([44, 48]). This is because a change in viewpoint is converted into a much simpler translation that can be easily computed and matched. Even though wavelets filters in [48] are good in background segmentation and im-

age denoising, they require a separate MACH filter synthesized for every action that needs to be recognized. In each of these methods focus is only achieving view invariance while designing the features and other challenges are mostly not considered.

Among all these approaches, 3D modelling of the human body postures shows more robustness towards change in viewpoint. This could be attributed to the fact that features extracted with respect to the human body itself. When human body is used as a reference, the action representation is relative rather than absolute making matching easier. But to infer 3D poses accurately from 2D video frames is very challenging as a number of parameters have to be estimated and ambiguous poses may be recovered due to perspective projection [18]. Another interesting point to be noted is that most available multi-view datasets on which researchers ([7, 23, 33, 46, 49]) have reported around 80-90% performance are in controlled setting with relatively simple, uniform or fixed background. To truly gauge the performance and robustness of any method, it would have to be tested in real world settings.

2.1.2 Occlusion

When body is not completely visible in the videos, action recognition is very difficult. The occlusions can either be a result of another person or object in the field of view. Due to view of the camera and human body articulated nature, self-occlusions are also possible where one body part occludes another. This poses a big challenge to the research community because not all the body parts performing the action are visible in the video sequence.

Some limitations caused by occlusion are avoided when considering videos as space time volumes [2]. Volumetric analysis ([2, 34, 50–55]) of actions are robust to self-

occlusions because of two reasons. Firstly, the body part need to be occluded throughout the video. Secondly, features from other body parts in the space-time volumes when they are not occluded might be discriminative enough to identify the action. To reduce the effect of cluttered background, volumetric analysis methods are combined with local features like edge maps [56], local appearance model [57] and space time interest points ([58–67]). Local representations create appearance models that can characterize the action in small patches. Compared to global features, these local representations can withstand pose, shape and illumination changes [68]. The main concern would be selection of appropriate interest points or patches to represent action in a discriminative manner. A normal interest point detector [60] might wrongly identify local patches which are not in the foreground object. For matching space time volumes in presence of outliers like occlusion, [50] used the Bhattacharyya coefficient.

A posture can be considered as a complete system with body parts as modular components. Following this as principle, probabilistic methods such as HMM ([26,30,69,70]) and Bayesian networks ([30,69,71,72]) create appearance models for limbs, heads and torso separately. Each body part's configurations is considered as a state in their model and change of states is governed by a probability value. Such methods have an added advantage of local representation as body parts are considered separately. [69] modified HMMs to handle missing data (occluded body part information) and used them to divide higher level activities into component sub-actions. From the available data, a probable body part configuration can be obtained. HMMs are good to represent simple actions but such a flat model cannot characterize the hierarchy and shared structure in all classes of actions [70]. Using the available data, similar to probabilistic methods, Kalman ([21,32,73]) and particle filters ([25]) also estimate the location of body parts. Thus these filters are employed for body part tracking.

Similar to probabilistic methods, pose based constraints ([52,74,75]) try to identify body parts individually. Based on available or visible evidences, these methods try to estimate parts which are occluded. Use of poselets ([8,76–79]) also follows the principle of body part detection. Originally proposed for action recognition in static images, extension of poselets [76] to videos is not so trivial. In addition, they require manually annotated training images to train the detectors for each body part, which limits their usage for action recognition.

One possible approach to handle occlusions are object interactions ([80,81]) and multiple camera setup ([6,19,37]). Using objects as contextual clues [80,81] estimated the location of body parts. Such methods are limited since all actions do not require object interaction. Similar to [80,81], using contextual or scene related clues to estimate occluded regions tend to make the approach biased to the scene background, hence cannot be separately to determine location of the occluded parts. In the case of multiple camera set-up methods, [6,19] are dependent on silhouettes which are not effective under self-occlusions.

Probabilistic methods and pose based approaches offer most acceptable performance when actions are being occluded. These methods tend to estimate the best possible location for occluded parts by taking into account the image evidences and articulated human body constraints. Investigation into classifiers that are occlusion-resistant is necessary. Features extraction is not possible from occluded parts, making it necessary to devise robust classifiers that can handle occlusions.

2.1.3 Execution Rate

This challenge mainly deals with how fast person is doing the action. The speed at which a person does an action can also vary depending upon several factors. Even the same person cannot guarantee same speed of execution when repeating

an action. Thus action recognition systems have to account for this variation in speed.

Methods that include HMM ([26, 30, 47, 69, 70, 72, 82–87]), Bayesian networks ([29, 30, 69, 71–73, 88–93]), conditional random fields ([94–96]) and fuzzy based systems ([97–100]) provide a probabilistic framework that is better suited to handle this challenge. In a probabilistic framework, system’s state is determined by the probability values and these values decide when the state has to be changed. Thus when actions are performed at different speeds, the model will be updated only when the probability values indicate a state change. Dynamic processes in nature can be modelled by HMMs [101]. Due to rigid flat structure of HMMs, it is extremely difficult to model simultaneous interactions and complex interactions [70]. These models only provide architecture to represent action, but the performance of the system will depend on the effectiveness of the features extracted. Experiments conducted in [102] show that HMMs are sensitive to training samples and tend to require more training data for better performance.

Simultaneous activities can be represented by employing Allen’s predicate logic and PNF networks ([103, 104]). These networks allow researchers to explicitly model temporal relationships, for example, ‘*during*’, ‘*before*’, ‘*after*’ etc. An interval algebra constraint network is employed in PNF networks [104] to represent temporal structure of actions. The employed network uses time intervals as the variables and binary temporal constraints between intervals correspond to arcs. The network aims to feasible values for the variable, which is a NP-hard problem. With only three states (*Past*, *Now*, *Future*), calculations are much easier in PNF networks. [63] combined AND-OR graphs with Allen’s interval logic to create a graphical model that encodes variations in activities. Both causal and temporal relationships can be encoded but requires weakly labelled data.

To generate models, the above mentioned approaches use temporal domain. Several researchers consider the whole video length and model actions in spatial domain. Time warping techniques ([7, 9, 20, 28, 45, 48, 74, 105–107]) tend to convert the template and the given input video data into a common time scale that allows for easy comparisons. A classic example of such methods are histogram based methods, specifically bag of features, dictionary or codebook based approaches ([8, 54, 64, 75, 77, 79, 108–116]) which consider the features from the complete video segments. Since these methods describe the whole action without considering how the action is being performed making it difficult to identify subtle change or temporal ordering of motion. For action segmentation and localization, these methods cannot be used directly. Also, codebooks generated suffer from quantization errors as all possible intra-class variations are generalized into codewords or visual words.

Considering temporal variations as intra-class variations, researchers ([6, 19, 27, 31, 35, 40, 42, 46, 117–122]) rely on the features and classifier used for action representation and training data. As most datasets contain more than one person performing an action, change in execution rate are implicitly captured by them. Due to this most methods are robust to this challenge.

2.1.4 Anthropometric Variations

These variations manifest due to differences between one person and another. Such variations are an integral part of huge amount of pose variations observed in an action. This is because each person might do an action differently. For instance, not all people swing their arms while walking. Apart from this, body size and context also determines how a person performs an action. To handle these variations, researchers are required to develop a generalized method.

A classification of several methods is provided in Table 2.1. As mentioned earlier, the pose or appearance of a person exhibits most variations during an action. It is very difficult to learn all these variations without enough training data. Thus researchers tend to prefer other features like optical flow, motion, frequency domain etc. For discriminative and unique motion representation, the whole foreground must be visible. Due to this such methods cannot be directly used for handling occlusions. Since motion features are coarse in nature, similar flow patterns can be seen for several number of scenes over a short time [120]. Such methods require proper foreground detection to avoid effects of background motion [9, 56]. Common practices to handle background changes include, bag of features [108], motion trajectories ([42, 118]), local feature representation [121] and space time interest points [105]. Features extracted by these methods are restricted to small patches in the frame, but these patches may or may not contain desired person's action.

Shape or appearance features are the simplest means of capturing anthropometric variations by determining the shape of the subject performing the actions. From the determined shapes, classifier is made invariant to these variations. Silhouettes, appearance models, pose based constraints and polygonal shapes are popular ways to capture the shape information. These methods are robust to variations in clothing and lighting [120]. For reliable and accurate performance, methods based on shape features requires reliable tracking system [9], accurate background model [9, 120] and stationary cameras [120]. Shape based methods mainly suffer in the presence of complex background, shadows and moving objects as accurate background subtraction is very difficult [123]. Silhouettes can only be used to represent whole body shapes and any body part motion within the body contour cannot be captured [120]. Even though discrete Fourier transform of image blocks [124] and power spectrum features [22] are robust to these varia-

Table 2.1: Classification of methods robust to anthropometric variations based on features

Features	Robust Methods
Shape, Appearance	[2, 7, 21, 23, 27, 32, 35, 41, 47, 52, 56, 57, 61–63, 71, 73, 94, 103, 126–130]
Motion, frequency and others	[3, 6, 22, 28, 29, 31, 33, 34, 39, 40, 42, 82, 84, 85, 88, 90, 105, 108, 118, 119, 121, 124, 131–135]
Hybrid	[9, 30, 43, 44, 46, 51, 54, 58–60, 64, 69, 77, 91, 109, 120, 136]

tions, they tend to be susceptible in the presence of camera motion and complex backgrounds.

As mentioned above, there are advantages and disadvantages when either shape or motion feature is used. To combine the advantages of both representation, hybrid features were applied. For instance, methods based space time interest points ([43, 58–60, 64, 91]) and poselets [77] encode features based on pose based constraints. Local patches thus generated are characterized by motion patterns [64, 109], optical flow [77, 125] etc. As poselets [76, 77] can determine orientation of body parts as well, they are able to handle internal motion. Methods employing hybrid features show much better performance than methods employing shape or motion separately. Such methods are robust to dynamic background changes and moving cameras [109].

Table 2.2 shows that most approaches can handle this challenge very well. It can also be seen that hybrid features perform better when compared to using only shape or motion feature. Since most available datasets contain more than one person performing the actions, anthropometric variations are very well captured. But most of the datasets, actions are captured in a controlled environment. It would be interesting to see the performance of these methods under unrestricted settings.

2.1.5 Camera Motion

It can be observed in videos, the cameras are not static always to make sure the foreground is in focus always. Also in amateur videos and unconstrained systems [66], cameras tend to be moving. Motion features related to only the action are difficult to extract as additional and erroneous motion patterns are introduced by moving cameras. Shape features generally require a good tracking mechanism, background model and stationary cameras [9, 120]. Tracking of required features and proper background subtraction is very difficult under the influence of camera motion.

Epipolar geometry and camera system are very common techniques to handle camera motion. In [41] tracking matrix is factorized into two; with one describing relative poses of the camera and foreground while the other describes the shape itself, thus yielding invariance to camera motion. Standard epipolar geometry was generalized to dynamics scenes in [39] using multi view geometry to handle moving cameras. They achieve this by deriving a temporal fundamental matrix and also analysing the rotation and translation motions. The main concern with these methods is that they require reliable and accurate camera calibrations with proper synchronization between all cameras. For such methods, intrinsic and extrinsic camera parameters are required to be known.

Another approach used to account for moving cameras is to include a motion compensation or subtraction component ([9, 19, 85]). [9] uses a set of shape-motion prototypes for action recognition. They determined if a pixel belongs to the desired person or not by assigning probability based on appearance likelihood maps. To remove camera motion components, estimated median of flow fields is deducted from the motion descriptor. [19] proposed an algorithm to use body centered motion field, where they subtract the motion caused by camera from the image. To

account for camera motion, a feature-based image alignment (homography from SIFT correspondences) is applied to all videos in [85]. Motion compensation is simple to implement but it might remove motion features that belong to the foreground object.

Researchers have tried to derive features that can be made invariant to camera motion. Optical flow [137,138], and local space time features [58] use observed velocity patterns to account for erroneous motion arising from moving cameras. [58, 137] achieve scale invariance as they use velocity patterns in scale-space. Effects of camera motion are cancelled out in the interest points generated by [138] as it uses the difference in partial derivatives of the flow field. To handle camera movements, [58] (Space time interest points), [138] (motion boundary histogram) and [137] (optical flow) use local features.

Codebook based methods ([61,62,91,139]) rely on the generated dictionaries. Such methods tend to be dependent on training data. Newly observed local features are assigned the same memberships of the codebook only if they contain patterns of scale changes and camera motion similar to those in training data [62]. Due to this they require extensive training and also suffer from quantization errors as a result of clustering.

On the whole, epipolar geometry based methods seem to provide the optimum solution when moving cameras are used. This is mainly because action representation obtained incorporates camera system and its geometry. A major hurdle in comparing approaches is the non-availability of a standard dataset that contains camera motion [39]. Hence, most of the above mentioned methods ([58,61,62] etc) either report their performance on stationary camera datasets or create their own dataset [39,139] for testing.

2.1.6 Cluttered Background

Dynamic or cluttered background is a form of distraction in the video sequence from the original action of interest as it introduces ambiguous information [62]. The presence of some disturbance in the background can induce changes either in appearance or motion of the foreground object. This disrupts flow based methods as unwanted background motion are also detected. Some approaches apply segmentation approaches to either segment or track foreground region, which is difficult in the presence of background clutter. Due to this, many approaches assume that the background is either uniform or fixed or accurate background segmentation is possible before processing [22, 23].

The most simplest solution to this challenge is adding a pre-filter or segmentation process ([9, 56, 131]) or normalize and threshold to separate foreground from background ([47, 60, 73, 82, 121, 140, 141]). These approaches can handle simple backgrounds which are uniformly distributed rather than complex backgrounds in unconstrained setting. As pointed out earlier volumetric analysis based on spatio-temporal features ([58, 114, 117, 120, 132, 142]) can be immune to background clutter. Such analysis do not rely on background subtraction or human body-part segmentation, and are relatively immune to noise, camera jitter, changing background, and variations in size and illumination [103].

To separate foreground region accurately, researchers have explored background modelling ([32, 57, 86, 143]). Some commonly used modelling paradigms include Gaussian mixture models ([36, 85, 111, 124]), Random Forests [144] and graphical models like latent semantic analysis etc ([61, 62, 91, 126]). Being robust to shadows and other artefacts [143], these approaches offer reliable background subtraction. The main challenge would be modelling a complex and dynamic background over a long period of time. Modelling complex and non-linear data in Gaussian mixture

models can result in large fitting error due to the linear nature of these models [145]. Some of these methods are limited by modelling constraints imposed while model generation. [32] is limited to work only on gray images.

Features that are robust to background clutter have been explored such as space time saliency ([43, 50, 59, 103]), pruning of motion features ([55, 109, 138]), BoW approach ([66, 130]), PbHOG [90] and pose based representation ([52, 53, 75, 77, 78, 113, 146]). Pose-based representations can cope with background variations, occlusions and shifts in global representations if key poses are selected well [52]. Performance of such approaches are dependent on the algorithm used for choosing key poses. Sometimes relevant key poses might not be included, reducing the accuracy of the method. Apart from this, using infra-red imagery [147] and depth [139] to remove clutter would limit the method's applicability to all scenarios.

For controlled environments and fixed background cases, a simple normalization and threshold approach based on either color or region can offer good results. Reduction in computation time and complexity is an added advantage of these methods. But for real life scenarios and dynamic backgrounds, modelling paradigms are most robust and reliable as they can adapt easily to any background. Even though these approaches provide good background segmentation, the final performance is dependent on the features extracted to represent the action. This can be seen from the results reported by these methods. In the presence of challenging backgrounds, the performance reported by all methods is in the range of 70-75% ([9, 50, 53, 64–66, 91, 109, 140]) indicating that more research has to be done to improve the performance. Combining model based methods with local feature representations can help in reducing the effect of cluttered backgrounds.

So far, some of the main challenges faced in action recognition were discussed. The performance of an approach can be affected even if just one of these challenges

is present in the videos. Thus solving these bottlenecks form the mainstay of researchers in action recognition. But most of the time, methods are developed to handle more than one challenge, which makes it difficult to classify them according to the challenges they are robust to. Table 2.2 classifies various methods according to the challenge it can handle. In Table 2.2, ‘I’ represents an integrated approach and ‘S’ denotes a separate block is added to make the method robust to the challenge. For example, using HMM that models temporal constraints inherently to tackle execution rate changes will be ‘I’, whereas including a time warping on the features or video sequences separately will be ‘S’. ‘√’ means that the approach can be generalized and ‘×’ represents that the approach is not robust to the challenge.

2.2 Action Classification

An important aspect of recognition framework is the classifier. It learns the actions using the features extracted and identifies actions in test videos. A classifier and feature combination should complement each other. The overall system’s performance depends on how well a classifier and feature complement and compensate for each other’s limitations. To get a reliable action recognition, classifier training process and data play a significant role. The training process is meant for the classifier to learn various intra and inter-class action variations. To learn all possible variations and obtain an accurate classifier, it is necessary to train with adequate and diverse training data. The effectiveness of an approach in terms of classification can be assessed by 3 aspects, generalizability, detection of abnormal actions and amount of labelled training data required (preferably lesser training data).

2.2.1 Generalizability

Usually, classifiers are trained only for a set of actions that need to be identified in the visual data. They cannot detect any other actions without prior knowledge or training. Some methods can easily learn another action with very few training data. This ability to learn actions other than what they are trained for increases its applicability to several day-to-day tasks. Therefore, researchers aim to develop methods that can easily learn actions and can be used for any purpose.

For generalizability, it is important that the action representation obtained is unique and discriminative. For instance, canonical poses [27], person dependent features ([20, 83]) rigid shape formations [41] cannot provide unique representation. With MACH filters ([3, 48]) extracted for each action and classifiers for gait patterns ([21, 27, 148] etc), generalization to activities with subtle variations is difficult. Region based methods that rely on negative spaces ([74, 107]) are dependent on background subtractions and shapes of negative spaces making it difficult to generalize.

Methods that require wearable sensors like Moving Lights Display [28], 3D Markers [12], solid coloured gloves [83] and RFIDs [14] cannot be generalized easily for any action in any real-life scenario. Similarly, using human object interactions as contextual cues ([30, 69, 72, 80, 81, 89, 130]) also cannot be extended to actions without objects. Apart from this a few methods are restricted to work only in certain conditions, like W^4 [32] and [147] built for monocular gray and infra-red imagery. Likewise, [139] requires depth information.

Table 2.2: Classification of Approaches based on Robustness to challenges, Generalizability, Abnormality Detection

Method	VI ¹	Oc ²	Er ³	An ⁴	Cam ⁵	C B ⁶	Gen ⁷	Ab ⁸
[82, 141]	×	×	I	I	×	S	×	×
[103, 114, 140, 143, 149]	×	×	I	I	×	I	✓	×
[150]	×	×	I	I	×	I	✓	✓
[125, 142]	×	×	I	I	×	S	✓	×
[146, 147, 151]	×	×	I	I	×	I	×	×
[152]	×	×	I	S	×	I	✓	×
[153]	×	×	S	S	×	I	✓	×
[20]	I	×	S	×	×	×	×	×
[105, 118]	×	×	S	I	×	×	×	×
[87, 111, 116, 122]	×	×	S	I	×	×	✓	×
[108]	×	×	S	I	×	×	✓	✓
[27, 28]	I	×	S	I	×	×	×	×
[7, 38, 45]	I	×	S	I	×	×	✓	×
[84, 88, 94–100, 110, 115, 129, 154, 155]	×	×	I	I	×	×	✓	×
[92, 93, 119]	×	×	I	I	×	×	✓	✓
[89, 107, 135, 156]	×	×	I	I	×	×	×	×
[29]	S	×	I	S	×	×	×	×
[8, 56]	×	I	×	I	×	I	✓	×
[57, 78, 81]	×	I	×	I	×	I	×	×
[72]	×	S	I	I	×	×	×	×
[73]	×	S	I	I	×	S	✓	×
[131]	×	S	I	I	×	S	×	×
[32]	I	S	×	I	×	S	×	×
[21]	I	S	×	I	×	×	×	×
[19]	I	I	I	×	S	×	✓	×
[31, 35, 42, 46]	I	×	I	I	×	×	✓	×
[40, 48]	I	×	I	I	×	×	×	×
[39]	I	×	I	I	I	×	✓	×
[86, 121]	×	×	S	I	×	S	✓	×
[90, 113]	×	×	S	I	×	I	✓	×
[23, 33, 44]	I	×	×	I	×	×	✓	×
[41]	I	×	×	I	I	×	×	×
[43]	I	×	×	I	×	S	✓	✓
[133, 134, 136, 157]	×	×	×	I	×	×	✓	×
[3, 128]	×	×	×	I	×	×	×	✓
[127, 158, 159]	×	×	×	I	×	×	×	×
[59]	×	S	×	I	×	S	✓	×
[60]	×	S	×	I	×	S	×	×
[58]	×	S	×	I	S	S	×	×
[51, 63, 70, 71]	×	I	I	I	×	×	✓	×
[69, 74]	×	I	I	I	×	×	×	×
[132, 160]	×	×	×	I	×	I	✓	×
[120]	×	×	×	I	×	I	×	×
[124]	×	×	×	I	×	S	✓	×
[126]	×	×	×	I	×	S	✓	✓
[2, 22, 34]	I	I	×	I	×	×	✓	×
[6]	I	S	×	I	×	×	✓	×
[30, 37]	I	I	I	I	×	×	×	×
[26]	S	I	I	I	×	×	✓	×
[25]	S	I	S	I	×	×	✓	×
[61], [62]	×	I	×	I	I	I	✓	×
[117]	×	×	I	×	S	I	×	×
[9, 109, 137, 139]	×	×	S	I	I	I	✓	×
[91]	×	×	I	I	I	I	✓	×
[85]	×	×	S	I	S	S	✓	×
[50]	×	I	×	×	×	I	✓	×
[47]	I	×	I	I	×	S	✓	✓
[36]	I	×	S	I	×	I	✓	×
[52–55, 64–67, 75, 77]	×	I	I	I	×	I	✓	×
[130]	×	×	×	I	I	I	×	×
[24]	I	×	×	×	×	×	✓	×
[49]	S	×	S	S	×	×	✓	×

2.2.2 Abnormality Detection

Abnormal action detection is an important activity in a surveillance scenario. It involves identifying abnormal or suspicious behaviour of a person. Usually, this task is handled by an operator, who monitors the person's movement throughout. Apart from these suspicious activities, abnormalities can also be group coordination in a dance, people trying to correct their postures, athletes trying to improve their performance in training sessions etc. Detecting such abnormalities helps to reduce manual labour. Some of the approaches used for abnormal behaviour detection methods in intelligent video systems are discussed in [161].

To learn any action, the system requires training samples. But in case of abnormal actions collecting training data might not be easy. Movie clips [43], YouTube videos [109], and amateur videos [66] might be a good source of training data for both normal and abnormal actions. Due to the unconstrained or wild settings in such videos, it would not be easy to learn from such sources. To learn from movie clips, [43] applied a text based classifier to movie scripts and subtitles to determine the period where the action is occurring in the movie. The main requirement is scripts or subtitles used must be properly aligned with scenes and must account for substantial variability of action expressions in the text.

Some of the methods ([106, 126, 128]) have been developed specifically for detecting abnormal behaviours. Defining an action as abnormal is very difficult as it depends on the action context. Detecting aberration requires these methods to

¹View-invariance

²Occlusion

³Execution rate

⁴Anthropometric variations

⁵Camera motion

⁶Cluttered Background

⁷Generalizability

⁸Abnormality detection

define criteria which characterize the abnormality. For instance in [126], h-LDA is employed to learn the distribution of the spatio-temporal words and correspond to various action classes. In this case, the criteria for abnormal action is defined using the probability distribution of motion text words learned by h-LDA.

The functional space of an action is modelled in [106] by applying time warping transformations on the action trajectories. These approaches require fewer training samples and is independent of features chosen for action representation, making extension to abnormality detection easy. The only issue would be defining abnormality criterion as it would vary based on the representation used. [128] developed a method based polygonal shapes to detect abnormal actions in low resolution images. Activities are modelled using polygonal shape of the configuration of point masses and their deformation over time. Abnormality is defined by learning the mean shape and dynamics of the shape change using manually indexed location data. This method can be applied only to static shape activities, which limits its usage.

2.2.3 Classifier Training

Classifiers and training strategies employed are briefly discussed in this subsection. Classifier approaches from other research fields are also included here which could be explored for action recognition purposes. Most common and simplest classification approaches include nearest neighbour ([87, 103, 119, 120, 131, 137, 151, 154] etc) and SVM ([45, 58, 108, 120, 127]). For distance calculation, most methods employ Euclidean distance between features, which is not effective in high dimensional spaces such as action recognition. Researchers have employed different metrics for distance calculation such as weighted Euclidean distance [22], normalized Levenshtein distance [151], Chamfer distance [105], Mahalanobis distance ([6, 19, 60])

and Riemannian metric [137]. The main assumption made in these cases is that data distribution in training and test are the same. However, this might not be true in many cases [112]. For capturing all variations, these classifiers require huge amount of training data.

Several techniques focus on automatic collection of training data or employ classifier strategies that require less training. For automatic collection of training data [90] used the web. An incremental procedure was employed to remove irrelevant data from a pool of training images collected from a search engine. Transfer learning model ([112,162,163]) allows the method to learn attributes and features in one dataset and apply them to another target dataset. From training image dataset, these models learn conceptual clues using SVM or AdaBoost, which will be applied to target action dataset. Active learning paradigm [164] requires a small amount of training data but is restricted in usage since wearable sensors are used to extract context information. Feature representation models generated by Incremental learning methods [67] and slow feature analysis [152] can be updated based on new training samples. Thus extensive training is not required for these approaches. Such methods do not make any distinctions between positive and negative training data.

For good performance, a classifier must be able to learn discriminative features with more weightage. Classifier [165] tries to learn from partially representative data in huge datasets that include negative data. ([55,149]) applied data mining to learn discriminative compound features that are characteristic of an action. Fuzzy rule based classifications ([97,98]) have also been shown to discriminate between features. [98] introduced McFIS classifier that can be used in an incremental manner since it automatically decides what, when and how to learn based on its available knowledge and new training sample. These methods are robust to most action recognition challenges. However, requiring the task adaptive classi-

fiers such as Hough forests [25] and adaptive vocabulary forests [166] to handle all the challenges in action recognition might not be easy. For instance, a separate Hough forest [25] needs to be created for every view angle we want to handle. These approaches learn appearance and model action based on codebooks, which suffer from quantization errors.

Based on keyboard inputs, [167] aimed to create user profiles by combining a tree data structure with fuzzy rule based classifiers. To account for behavioural changes, these classifiers adapt and evolve with time. Such a classifier structure can be used for developing a hierarchical action recognition system. Human hand motions are recognized in [102] using fuzzy empirical copula classifier with the help of finger joint angles captured via data gloves. ELM [54] is a popular classification approach due to its faster training time and good multi-class performance. Hidden nodes of an ELM classifier can be generated even without seeing the training data [168]. Because of random initialization of hidden nodes, ELM results can fluctuate. Thus, usually, ELM classification is repeated a number of times and average performance is computed as final result.

2.3 Deep Learning Approaches

In contrast to conventional framework of feature extraction and classification, deep learning focuses on end-to-end training. They operate on raw pixels directly and can classify based on them. Due to its highly discriminative representational power, they have been used for several image based applications like pose estimation [169], image classification [170, 171]. All these methods have shown very good performance in static images. Being different to traditional recognition frameworks, we look at deep learning methods available for action recognition.

For better performance, deep learning requires a lot of training data. Due to large number of image data available such as ImageNet [172], Convolutional Neural Networks (CNN) was successfully used for image based applications. For video classification purposes, it would be necessary to handle a sequence of frames and incorporate temporal dynamics observed. Also it is not easy to collect huge dataset of videos which can be used for training. For video based applications, [173] proposed a multi resolution and foveated CNN architecture to capture local spatio temporal features. To speed up the training process, the architecture consists of foveat and context streams. The context stream handles low resolution video frames where as foveat stream focuses on only middle portion of high resolution frames. To combine the information from network handling sequence of frames, 3 fusion techniques, namely, early, late and slow fusion were adopted.

Another popular network architecture, Recurrent neural networks (RNN) was adopted in [174] to develop long-term convolutional network (LRCN) model for video classification purposes. LRCN models allow compositional spatial and temporal layers and long term dependencies are learnt by incorporating nonlinearities into the network state updates allowing for complex temporal dynamics. LRCN's effectiveness was shown in video based tasks such as activity recognition, image description and video description tasks.

[175] proposed a two stream ConvNet architecture that separately handles spatial and temporal aspects of a video. CNNs are well established for static image recognition purposes. Due to this spatial convNet, which works on separate static video frames can be pre-trained on ImageNet Challenge dataset for better classification. On the other hand, the temporal convNet works on multi-frame dense optical flow. By combining output of both convNets, great improvement was shown in action recognition datasets. To improve the performance of the temporal convNet, other motion features such as motion vectors [176], trajectory pooled convolutional de-

scriptors [177] were proposed. Temporal segment networks [178] extended the conventional two stream network to incorporate long range temporal structure modelling observed in an action. A long video is divided into a fixed number of segments and sparse temporal sampling strategy is employed to select snippets from these segments. The classification loss function is optimized based on the segment consensus. They also showed warped optical flow as an input modality to handle camera motion.

Two stream architectures fail to register appearance with optical flow and how these cues evolve over time [179]. To alleviate this limitation, [179] developed an architecture to fuse spatial and temporal cues at several levels of granularity in feature abstraction, with both spatial and temporal integration. [180] introduced dynamic images, a single RGB image that encodes both spatial and temporal information of a complete video using rank pooling. Convolutional networks are trained on these dynamic images to classify into different actions.

An important consideration with deep learning is the huge requirement of training data. All deep learning methods discussed are pre-trained on ImageNet classification challenges for best results. They have also shown that when pre-training is not done, the performance reduces drastically. For non-upright actions, the image and video training data available is very limited. Due to this recognizing non-upright actions is not easy in a deep learning architecture. For the action recognition challenges discussed, the training data is important as these methods learn from them. Also, [173] showed that CNNs struggle to learn complete invariance across all possible angles, speeds of camera translation and zoom (camera motion).

2.4 Datasets

Datasets are an integral part of any pattern recognition or computer vision related tasks. To benchmark developed algorithms and allow comparison with other methods, it is necessary to have datasets so that qualitative and quantitative performance analysis can be done. One important concern about dataset is whether diverse nature of actions are captured in it and different above mentioned challenges are showcased in them. To test a method's robustness to any of the challenge it is required for the dataset to adequately portray that challenge. It is very difficult for to capture all possible challenging scenarios in a single dataset. Also, some datasets are not publicly available. This is the primary reason why many researchers create their own dataset for evaluation ([39, 40, 82, 105, 121, 128]). A detailed survey on the available datasets for performance evaluation is given in [181].

Most commonly used action recognition datasets include KTH & Weizmann human action datasets. KTH dataset [58] contains 6 action classes performed by 25 actors in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors (s4). KTH dataset is restricted because cameras are relatively stationary, and only zooming of camera is considered as camera motion. Weizmann dataset [2] contains 10 action classes by 9 actors in uniform background and fixed camera [181]. In addition, silhouettes or foreground masks are provided with the dataset. Both these datasets are relatively simple dataset focusing on simple actions like walking, running etc.

To properly capture variations in view point, datasets like IXMAS [33], CASIA motion [47], i3DPost Multi-view [1], MuHAVi [182], ViHASi [183] [184] and West Virginia University (WVU) multi-view [185] use multiple cameras viewing the action from different view points. This allows us to conduct experiments methods

with/ without a specific view and test robustness towards view angle changes. All these datasets do not consider cluttered background and capture videos in controlled indoor setting. UT-Tower dataset [186] contains low resolution action videos where person is not clearly seen and is captured from an aerial view point. The datasets discussed so far focus on only single person single action videos. Therefore they cannot be used for testing temporal action localization of a method.

For continuous actions, ChaLearn gesture dataset [187] and G3D, a gaming dataset [188] were introduced. Challenges in vision based action recognition is not the main focus of these datasets. ChaLearn provides multi-modal data including depth, audio and skeleton for recognizing only hand gestures. G3D is a gaming action dataset focuses only on a specific class of actions from gaming environment.

Datasets like Keck gesture set [189], UCF Sports dataset [3] and Hollywood2 human action (HOHA) datasets [108] provide good database for camera motion and background clutter. YouTube dataset [109] is collected from YouTube and broadcast videos that are captured under uncontrolled settings. In addition, HOHA dataset and YouTube dataset [109] depict actions in very challenging scenarios. These datasets contain videos from movies and broadcast that are shot in uncontrolled setting. They depict actions with subtle motions and in the presence of severe occlusion making it extremely difficult to recognize. UCF Sports [3], HOHA [108] and YouTube [109] are good datasets for dynamic background since they are extracted from sports, movies and web videos. They also can provide good benchmarks for multi-person actions and interactions.

One common factor among most of these datasets is the lack of non-upright human action depiction. Most datasets restrict themselves to actions being performed in upright human postures like walking, running, kicking, standup etc. There is a definite lack of dataset that focuses on action being performed in non-upright

postures. HOHA datasets [108] does contain a few non-upright posture cases but they might not be enough to test the robustness under this situation. Due to this there is a lack of research study on non-upright human posture actions.

2.5 Concluding Remarks

In this chapter, we have explored available methods based on their robustness to the various challenges faced in action recognition including view invariance, occlusion etc.

It can be observed that bottlenecks like execution rate and anthropometric variations have effectively solved if adequate number of training samples are available. Both these challenges have been solved easily by considering them as intra class variations. Methods that tackle camera motion are even fewer. Methods employing motion compensation or the multi view geometry along with standard epipolar geometry to resolve the discrepancies have shown good performance but are still limited in nature.

Attaining view-invariance in action recognition is one of the most challenging task. To train with a set of view angles and match it to a different 2D view is extremely difficult as there is loss of information when 3D real world data is converted to 2D image data. Moreover, for real world settings it is necessary to tackle other challenges also. Several methods have achieved moderate view-invariance with provided benchmark datasets.

Occlusion and background clutter have been tackled by several researchers, but there is still a lot of scope to obtain completely systems that are completely invariant to these changes. From the perspective of different challenges, it can be seen that challenges such as view-invariance, occlusion require more attention from

researchers. It is also evident that there is a lack of research in actions performed in non-upright postures. As mentioned earlier, pose variations are a result of either view-invariance, occlusion, ability to perform actions in different postures etc. Thus there is a necessity to develop a pose-invariant action recognition system that can handle all the above mentioned challenges. Apart from this, datasets need to be developed that can depict different challenges to benchmark any method's robustness to them. In this thesis, we focus on developing a pose-invariant action recognition system and extend research study into non-upright posture actions.

Chapter 3

Forward Path for Pose-Invariant Action Recognition

3.1 Introduction

Actions are visible effects of one's intentions making it a significant part of behavioural analysis. As pointed out earlier, reliable action recognition can be the first step towards understanding a person's behaviour. Several bottlenecks such as view angle, occlusion, cluttered background etc make it very difficult to recognize actions. One main concern in action recognition is the amount of pose variations observed is always high. It is a result of articulated nature of human body and other factors like view angle variation, occlusion, anthropometric variations, object interactions etc. Thus the main focus of this thesis is to come up with a generalized framework for pose-invariant action recognition and study non-upright posture actions.

In this chapter, an overview of the proposed overall novel mutually reinforcing two step framework to combine kinematic motion features and pose cues that

is pose-invariant, partially view-invariant and can handle occlusion is presented. The proposed framework is inspired from control system theories and combines motion and pose cues in a closed loop reinforcement framework. In this chapter, the first step of the proposed framework, propagation motion forward (PMF) path is explained in detail. PMF path extracts pose-invariant motion features based on two assumptions, namely, neck reference point is available and major viewing direction (front, back, left, right) is available.

Figure 3.1: Proposed pose-invariant action recognition framework. Each video passes through PMF and CPF paths. In PMF path, based on neck point and foreground, body is divided into head, torso, leg grids and kinematic features are computed and re-aligned in the human body centric space. Using these features, an initial action hypothesis is estimated. In CPF path, canonical pose hypothesis scheme uses canonical stick poses for initial estimated action to determine the pose in the frame. Using the estimated pose, two additional kinematic features are computed (re-aligned in human body centric space) and grid division is corrected. This forms the new set of features for action recognition. The loop is repeated until the recognition converges to an action

3.2 Overview of Complete Framework

When only motion features are used, it is not possible to differentiate it as either required foreground or unnecessary background motion due to moving objects and other people. However, this could be improved if the body parts could be detected to differentiate the body part's motion versus motion due to background. But detecting body parts correctly is very challenging unless some pose/ appearance information is available. On the other hand, using pose alone will fail when there is partial occlusion or when only parts of the person are visible. This is because the occlusion or partial image will cause mis-match in the pose. In addition, pose is inherently dependent on view while motion features could be made invariant to pose and view if local information is used and computed relative to the body's reference frame. This motivates us to propose a mutually reinforcing framework that exploits the strength of motion features and pose in such a way that will still achieve pose invariance and be able to withstand partial occlusion.

Figure 3.1 shows the overall framework of our proposed approach. It comprises of propagation motion forward path (PMF) that extracts pose-invariant motion features and provides initial hypothesis of possible actions present in the input video. For computing pose-invariant motion feature, the proposed framework assumes neck point and the person's viewing direction is available. Automatic detection of neck point might not be easy in all videos, but person detectors like deformable parts model [190] can be used for low resolution. Whereas for high resolution, head detectors or upper body detectors can be used to estimate a neck point. A propagation mechanism is used to divide the foreground region into 3 grids corresponding to head, torso and legs respectively. PMF path extracts kinematic motion features and encodes them in a pose-invariant manner by representing motion with respect to the body itself. These set of kinematic features computed by

PMF path is used to provide an initial hypothesis of possible action.

Subsequently, the canonical pose feedback path (CPF) uses the action recognized by PMF to identify the pose in a frame and correct the erroneous motion features. Every action is characterized by specific set of canonical poses. An offline algorithm is run on the training videos to determine the canonical stick poses for each action. Using the canonical sticks of action recognized by PMF along with a body part detector, a pose hypothesis scheme identifies the pose in the frame. Apart from correcting erroneous motion features, the obtained pose information is used to compute two new sets of pose-invariant kinematic features focusing on capturing motion of each body part and temporal dynamics of the action. These set of features are again used in the PMF to enhance the selection of correct action until loop structure comprising both paths converge to the correct action.

The available training videos serve 2 purposes, training the action classifiers based on both initial and final pose-invariant kinematic features and secondly, to extract the canonical sticks for every action.

3.3 Propagation Motion Forward Path

The propagation motion forward (PMF) path takes as input the preprocessed video whereby the foreground region is extracted. The framework uses either the already available foreground masks (eg. those provided by the benchmark datasets) or a simple average frame background subtraction to obtain the foreground. Also, methods like [123] can be used to extract the foreground by using background modelling. The PMF path comprises of three stages, namely, propagation mechanism for grid division, kinematic feature computation and conversion to human-centric space with pose-invariant and partially view-invariant encoding.

3.3.1 Propagation Mechanism

The first stage of PMF path, the propagation mechanism for grid division, focuses on dividing the foreground region into 3 grids, namely head, torso and legs respectively. This is to obtain a discriminative representation of motion of each body part during an action. Given an estimate of head size as $2x$, using the prior information from the natural proportionality of human body as shown in figure 3.2, the torso and leg sizes will be approximately $5x$ and $6x$ in size in the normal upright posture. Hence, we can estimate the grids for head, torso and leg respectively. The pre-processed foreground and available neck point serve as an initial pose estimate. The body orientation can be roughly estimated as the direction between neck and center of mass. Using the body orientation and neck point the head grid's location and size can be estimated. Our propagation mechanism uses this approximate head size and propagates in the direction of body orientation to obtain an initial set of grids. This mechanism is able to identify grids correctly even for non-upright postures as long as the segmentation is correct.

Figure 3.2: Proportionality of body parts used in Propagation Mechanism. Sample Images from Weizmann [2] and UCF Sports [3]

3.3.2 Kinematic Feature Computation

The second stage involves computation of kinematic features. A new set of kinematic features is proposed, namely, Divergence, Curl, Projection and Rotation respectively which is motivated from [191] used for emotion recognition. These features are described by the following equations

$$Div(p) = \frac{\partial OF_x(p)}{\partial x} + \frac{\partial OF_y(p)}{\partial y} \quad (3.1a)$$

$$Curl(p) = \frac{\partial OF_y(p)}{\partial x} - \frac{\partial OF_x(p)}{\partial y} \quad (3.1b)$$

$$Proj(p) = \vec{OF}_p \cdot \hat{P}_{neck} \quad (3.1c)$$

$$Rot(p) = \hat{P}_{neck} \times \vec{OF}_p \quad (3.1d)$$

where Div measures the expansion and $Curl$ measures the dynamics of circular motion at foreground pixel p in a local neighbourhood using OF_x and OF_y , the x and y components of optical flow at foreground pixel p respectively. $Proj(p)$ and $Rot(p)$ are two new kinematic features that measures the scalar product and vector product of OF_p (optical flow at p) with respect to a unit vector at a stable reference point. The neck point ($\hat{P}_{neck}$) is used as the reference point in our current framework since it is stable and can be viewed easily from all directions. Figure 3.3 shows a few example images with neck markings. The neck point location is assumed to be available or an approximate location can be obtained using deformable parts [190].

Figure 3.3: Neck points marked for sample dataset images

3.3.3 Human-body Centric Space

A key difference is that to achieve pose-invariance, we compute the features with respect to the body itself. In [191], they used unweighted (only sign) and weighted (magnitude and sign) histograms of each kinematic feature separately accumulated over the entire video assuming that the face is frontal or near frontal when the facial expression is acquired. This is not applicable for our case as we have to represent the actions or the motion of body parts in 3D as well as accounting for different views. For this purpose, the third stage in the PMF path is used to transform the features into the human-body centric space. This space can be thought of as a cylinder with origin at the neck and characterized by three dimensions with respect to the body, namely, up-down, left-right and front-back as shown in figure 3.4. Given a frame, such as figure 3.5(a), only two of these dimensions can be determined using the body orientation estimated by the propagation mechanism and the available viewing direction of the person (up-down & left-right). Therefore, Divergence and Projection will lie in one of these dimensions whereas Curl and Rotation will lie in the front-back dimension. For any frame, computed kinematic features will lie in either one of the dimensions of human-body centric space.

To obtain a more accurate representation of body part motion, initial set of grids

Figure 3.4: Human-body centric space cylinder

are divided into smaller cells as shown in figure 3.5(a). A 3x2 cell configuration is used for each grid. Figure 3.5 shows subjects performing actions in different view and body postures. While dividing into cells based on body orientation and viewing direction, we account for view angle and posture variation by labelling the cells as shown in figure 3.5. The kinematic features in each labelled cell are re-oriented based on alignment and accumulated over the entire video. This labelling of cells helps us to compare the features in a more view-invariant manner.

The kinematic features are grouped into two classes, one comprising of Divergence and Curl that measure the local dynamics and the other comprises of Projection and Rotation, which are measured with respect to the stable neck point. Our action representation contains 2 components

- a. Weighted and Unweighted histograms of divergence and curl for each human-body centric space dimension of each cell in each grid.
- b. Weighted and Unweighted histograms of projection and rotation measured

Figure 3.5: Division of grids into smaller cells & Labelling of cells based on the change in view, body orientation, Only labelling of Grid 2 cells are shown

with respect to the neck reference point for each human-body centric space dimension of each cell in each grid.

L2-normalization is applied on the kinematic features for each labelled cell over the video length before computing the above mentioned histograms to account for inconsistencies that might arise due to speed variations. To represent one cell, both weighted and unweighted histograms are used. Unweighted histogram focuses on sign of the motion, thus is represented by 3 values corresponding to either positive, negative or neutral. Weighted histograms focus on magnitude and sign, thus is represented by 2 values corresponding to positive and negative. There are in total 3 dimensions for one cell, up-down, front-back, left-right. Thus each cell is represented by 9 values in unweighted and 6 values in weighted histograms. Each grid has 6 cells making a total of 18 cells for entire foreground region. Therefore entire foreground region is represented by $(18 \times 9) = 162$ features in unweighted and $(18 \times 6) = 108$ features in weighted histogram features. Thus a total of $162 + 108 = 270$ features is used for one class of above mentioned kinematic features. The complete action representation in PMF path comprises of $270 \times 2 = 540$ features. For initial action recognition, ELM [168] classifier is used and trained with these histograms as features during the training phase. ELM is chosen due

Table 3.1: Pose-invariant Motion Feature Extraction in PMF path

Algorithm
<ol style="list-style-type: none"> 1. Propagation Mechanism <ol style="list-style-type: none"> (a) Determine body orientation using neck and foreground pixels. The body orientation can be roughly estimated as the direction between neck and center of mass obtained from foreground pixels. (b) Divide into the initial grids by using the approximate head size and propagates in the direction of body's orientation. 2. Compute the kinematic features for every foreground pixel. 3. Determine the direction in which the kinematic features are aligned with respect to the human-body centric space. 4. Divide the grids into smaller cells and label them according to body orientation and viewing direction. 5. Accumulate the kinematic features for each labelled cell over the entire video and compute the following <ol style="list-style-type: none"> (a) Histograms for local dynamics in each direction. (b) Histograms for dynamics with respect to neck in each direction. 6. The combination of both histograms for all cells forms the proposed action representation.

to its fast training time and good multi-class performance [54]. Table 3.1 provides the step by step description of the pose-invariant motion feature extraction.

3.4 Experimental Results and Discussion

In this section, experimental results for the proposed framework with only PMF path is presented. The experiments have been conducted on 4 benchmark datasets, Weizmann [2], KTH [58], UCF sports dataset [3], Hollywood dataset [43] and HMDB51 dataset [192]. The aim of these experiments is to find out how good is the proposed PMF path. From the results key insights and limitations can be identified which can be used to improve the whole framework. For ELM classification, two parameters are to be fixed, namely, number of hidden layer neurons and activation

Table 3.2: Performance of PMF path in different datasets

Dataset	PMF path Performance (%)
Weizmann	92.47
KTH (16 + 9)	87
KTH (LOV)	90
UCF Sports	76.31
Hollywood	54.12
HMDB51 (General Body movements category)	63.03

function. For all experiments, sigmoid activation function is used and number of hidden layer neurons are varied from 200 – 40000 and the best performance obtained for each dataset is tabulated in 3.2.

3.4.1 Weizmann Dataset

[2] introduced Weizmann dataset, which is one of the most popular dataset in action recognition. It is a very simple dataset with simple background and relatively static cameras. The dataset contains 9 people performing 10 actions (bend, jack, jump, pjump, run, side, skip, walk, wave1, wave2) and evaluated in leave one person out strategy. The foreground masks are provided with the dataset itself. The neck point is manually marked for each frame. The major viewing direction is also assumed to be available. From table, with only PMF path a performance of 92.47% is observed. The main confusion is between *skip-jump*, *jump-walk*. Analysing *skip* and *jump* videos, it can be seen that both are quite similar actions except for the variation in leg position. Since the propagation mechanism focuses mainly on motion and does not differentiate between posture of legs within the grids making recognition difficult. Similarly, *walk* and *jump* actions could also be confused as leg posture is not identified. The initial head size determined in PMF path is not always correct. For instance, consider the figure 3.6, due to presence of hands above the head, the computed head size is too big that leads to wrong torso and

Figure 3.6: Incorrect head size determined by propagation mechanism

leg grids. A few error cases were seen in *wave2* action due to wrong head size estimate.

3.4.2 KTH Dataset

KTH dataset [58] comprises of a total of 600 videos showing 6 actions (boxing, handclapping, handwaving, jogging, running, walking) performed by 25 actors in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3), and indoors (s4). KTH dataset is shot in relatively simple backgrounds with more or less stationary cameras except for zooming in and out. Due to the large amount of videos in this dataset, manual marking of neck would be difficult. Hence, [190] implementation of deformable parts model is used to detect the person in the frame and use the obtained parts to identify the neck. Average background subtraction is applied within the person's bounding box to obtain the foreground.

KTH dataset is tested in two settings, 16 train + 9 test suggested split in [58] and leave one person out (LOV) setting. Results for both settings are tabulated in Table 3.2. The six actions can be classified into 2 classes, hand actions (box-

ing, handclapping, handwaving) and leg actions (jogging, running, walking). The framework with only PMF path does not much confusion between hand and leg actions. Considering the hand actions, boxing is performed in side view whereas handclapping and handwaving are done in a frontal view. From obtained results, boxing is identified 100% correctly and only confusion is between handclapping and handwaving. This indicates that by aligning the features on human-body centric space and encoding them based on the view is useful. It is noted that the waving action leads to incorrect head size estimate leading to wrong grids.

Considering the leg actions, jogging, running and walking are very similar actions due to which erroneous detection are observed. There are two main reasons for such erroneous detection. No person is detected in several frames of video dataset by [190]. Due to this, motion of the person could not captured well to represent these action. Moreover to characterise and discriminate similar actions it would be necessary to have all the frames properly detected. The second drawback is the average background subtraction method. In many cases depending upon the background, the leg region of the foreground is not at all detected, which means leg motions cannot be learned properly. In addition, the presence of shadows also disrupt the neck detection and foreground detection. Due to the large number of actors, pose variations and execution rate variation exhibited is high. By identifying the pose, the recognition performance and pose-invariance of the framework can be improved.

3.4.3 UCF Sports Dataset

UCF Sports dataset [3] comprises of a total of 152 videos with 10 sports based action classes (Diving, Golf-Swing, Kicking, Lifting, Riding-Horse, Running, Skate-Boarding, Walk, Swing-Bench, Swing-SideAngle). The number of videos for each

Figure 3.7: Confusion matrix for UCF Sports dataset with only PMF path

action class varies from 6 to 22. This is a challenging dataset due to cluttered background, moving cameras, view changes etc. Several reasons such as view angle change, anthropometric variations contribute to pose variations in this dataset. For this dataset, a person bounding box is marked and average background subtraction is applied to obtain the foreground. For testing, a leave one out approach is adopted as per the other methods in the literature. The proposed method shows an accuracy of 76.4% with only PMF path and confusion matrix is shown in figure 3.7.

Even though actions such as *Diving*, *Swing-Bench*, *Swing-SideAngle*, *Lifting*, *Walk* are recognized correctly other actions show a lot of confusion. It can be noticed that *Golf-Swing*, *Kicking*, *SkateBoarding* are the worst performing actions. Upon analysing the cases, *Diving*, *Swing-Bench*, *Swing-SideAngle*, *Lifting* are done indoor where as *Golf-Swing*, *Kicking*, *SkateBoarding* are outdoor games which means background clutter will be more. As expected the foreground detected in many cases where not clear enough to show only the person. Also the videos are with moving cameras as they are shot with the intent of focusing on the player or the object in the scene. Another reason could be the amount of pose variations exhibited for each action. With no knowledge about the posture of the person, it would be difficult to learn the actions.

3.4.4 Hollywood Dataset

Hollywood dataset is a publicly available dataset that depicts 8 action classes (SitUp, SitDown, StandUp, Kiss, HandShake, HugPerson, Answer Phone, Get Out Car) collected from 32 movies. This is a very challenging dataset due to several reasons such as occlusion, view angle changes, subtle movements to depict actions, cluttered background. Similar to UCF sports, neck point and person bounding box is manually marked (used for average background subtraction). Since the method cannot handle actions involving 2 persons, only one person is marked for Kiss, HandShake and HugPerson actions. Test and train splits (*train_clean* and *test_clean*) are already provided with the dataset.

Handling Occlusion

For the proposed propagation mechanism to properly identify the grids, body orientation needs to be identified. PMF path uses the neck point and foreground region to determine the orientation. Under the presence of occlusion, it is very difficult to identify the orientation based on the foreground region. Hence the body orientation is assumed to be available to us. The dataset exhibits varying amount of occlusion. For instance, videos showing *StandUp* action contain only head or head-torso or complete body, which means the comparing motion features would be very difficult since the proposed action representation is based on grid based feature extraction. To handle this, the number of required grids for each action is defined, for example, Kiss action requires only grid 1. The set of required grids for each action is shown in table 3.3. When features are extracted for training videos, the required grids are only computed, the rest of them are padded as zeros. For actions where all 3 grids are required, only the available grids are computed, if a grid is occluded it is just considered as zero. For test video, the features are

Table 3.3: Required Grids for actions in Hollywood dataset

Action Class	Required Grids
SitUp	All 3 grids
StandUp	All 3 grids
SitDown	All 3 grids
GetOutCar	All 3 grids
AnswerPhone	Grid 1 & 2
HugPerson	Grid 1 & 2
HandShake	Grid 1 & 2
Kiss	Grid 1

computed for the available or unoccluded grids and directly passed to the classifier.

The performance in this dataset is good and comparable with many other state of the art methods. This dataset is extremely challenging due to reasons such as pose variations, occlusion, camera motion, cluttered background, view angle variations etc. A few cases show non-upright postures in the actions. The adjustments made to handle occlusion are effective. Upon analysis of errors cases it was observed that lack of posture knowledge, improper head size estimation and foreground detection were main causes for the recognition errors. Head size estimation is even more difficult in this dataset as foreground detection, occlusion and pose variations exhibited are too high. From the experiments, it can be seen that by knowing more about the posture of the person in the frame, pose-variations can be handled better as body parts can be differentiated. This would make the proposed action representation more discriminative and pose-invariant.

3.4.5 HMDB51 dataset

HMDB51 [192] is a large human motion database with around 6800 videoclips with 51 action classes. The actions can be broadly classified under five categories, namely, general facial motion, facial action with object manipulations, general body movements, body movements with objects and interactions. Each action

class has a minimum of atleast 100 videos with variations in view, background, postures, presence of occlusion etc. In this dataset, apart from motion, appearance and context also play an important role in action recognition. Thus, deep learning approaches [175, 178, 179] capable of capturing both spatial and temporal cues are more suitable for these datasets. The proposed action recognition framework is mainly suited for body movements and not suited for facial motion, facial motion with objects and interactions categories. For the ‘body movements with objects’ category, the proposed framework is not suited as it cannot capture object contexts or differentiate between objects and person. The proposed action recognition framework might not be able to achieve very good performance in cases where appearance and context dominate. The proposed framework mainly focuses on achieving pose-invariant action recognition for whole body movements including non-upright postures. A subset of 19 actions belonging to general body movements category are chosen and the proposed PMF path is applied to it. Similar to UCF sports and Hollywood, the neck point and person bounding box is manually marked (used for average background subtraction) for the videos. Since multiple persons cannot be handled by the framework, only one person is marked if there are more than one. Person’s viewing direction is provided as a part of meta label in each video.

To test on HMDB51 dataset, 3 training/testing splits are provided. PMF path is applied on the videos of selected action classes for feature extraction. To handle occlusion, section 3.4.4 is closely followed. Only 3 actions *wave*, *turn* and *clap* use grid 1 and 2 and the rest use all 3 grids for recognition. The PMF path shows an recognition accuracy of 63.03%. The chosen set of actions show a good mix of upright and non-upright actions and the framework is able to handle these changes and show good performance which is comparable to other methods. Also this dataset includes clips with pose variations, occlusion, camera motion, cluttered

background, view angle variations making recognition very difficult. Among the chosen 19 actions, only *dive*, *fall_floor*, *jump* and *wave* are confused in all 3 splits. *Dive* and *jump* are confused with each other or with *climbing* since they share similar kind of motion patterns. *fall_floor* involves whole body moving down due to which it is confused with *sit*. *wave* is wrongly detected as *turn* in many cases because of occlusion, improper head detection and lack of posture knowledge to differentiate hand motion from whole body motion. Many wrong detections can be attributed to the above mentioned reasons and this can be reduced by estimating person's pose.

3.4.6 Discussion

From table 3.2, the performance of the proposed PMF path is not so high. Thus a comparison with state of the art methods is not done in this chapter. One of the observed shortcoming was the improper foreground detection due to background clutter, occlusion, improper lighting condition etc. As a result, action representation was not good since important motion features were lost.

A common cause for erroneous detection was improper estimate of head size, resulting in improper grid division. Due to this, motion features are accumulated in to the wrong grid resulting in erroneous histogram features. Also confusion was observed between similar actions like running, jogging etc. This is observed because of two reasons, namely, no temporal characteristics are taken into account while histogram is used and no information on the posture of the person is available. If the person's pose can be estimated it would be really beneficial as the location of each body part can be used to identify the grids and make action representation more discriminative.

3.5 Non-Upright Action Dataset (NUAD)

In this section, a new non-upright action recognition dataset is introduced. Currently traditional benchmark datasets available to the research community focus mainly on actions being performed in a upright standing posture. But the articulated nature of the body allows it to be in different postures. To develop a generalized pose-invariant action recognition dataset and explore the research study on non-upright postures, such a dataset is required. Therefore, a set of 8 actions have been chosen with a mix of standing, sitting and lying down poses included. The dataset contains 8 actions, namely, bending, climber, double-hand sitting wave, knee bender, pushup, single-hand sitting wave, lying down wave and squat. All these actions involve some kind of non-upright or sitting posture which is different from other datasets. Some of the example actions are shown in figure 3.8. The dataset consists 35 actors performing these 8 actions and captured from 3 views (front, left, right). Due to the non-upright posture, actions such as lying down wave, pushup, climber will be severely occluded in front view, therefore, these actions are captured from higher altitude (close to top view) such that most body parts are visible. To capture enough pose variations, action videos of 35 people are captured from 3 views. The dataset is captured in mostly uniform, simple backgrounds. The neck point for each frame is manually marked. The foreground marking has also been done manually. 11 body parts (Head, torso, neck, Right(R)& Left(L) arms, R& L legs, R& L hands, R& L feet) are marked separately for every frame in the form of bounding boxes (neck is just a point) and stored as text files. If a body part is occluded, it's bounding box coordinates are '1 1 1 1'.

This dataset can be used to test the ability to handle non-upright postures and pose variations observed during an action. It also allows us to perform view

Figure 3.8: Sample Images from NUAD dataset**Table 3.4:** Performance of PMF path in NUAD

Dataset	PMF path Performance (%)
NUAD	90.1

invariance testing using multi-view data. The results of leave one person out setting performed on NUAD is tabulated in 3.4. The PMF path can achieve a performance of 90.1% even in presence of non-upright postures indicating that it can handle such changes. As with previous datasets, confusion is observed between similar actions. For instance, there are 3 types of waving actions, *bend-squat*, *climber-pushup*. It is very difficult to differentiate these actions without knowledge of each body part. For instance, *bend-squat* appear quite similar and especially in front view. Also evident was the improper head size estimate during some of the actions as shown earlier in figure 3.6.

3.6 Concluding Remarks

In this chapter, an overview of novel pose-invariant action recognition framework combining motion and pose cues in a mutually reinforcing manner was proposed. This chapter elaborates propagation motion forward path and important contributions made in it. PMF path uses a propagation mechanism based on the natural proportionality of body parts to identify an initial set of grids corresponding to head, torso and legs respectively. A new set of kinematic features are borrowed from emotion recognition domain and converted to human-body centric space to handle 3D motion and variation in view. The computation of features with respect to a stable neck point and conversion to human-body centric space allows for pose-invariant action recognition. The initial set of features extracted from PMF path is used for action recognition. To focus on non-upright posture actions, a new dataset with 8 actions is introduced. Experiments have been conducted on the newly introduced dataset as well as on several other publicly available datasets such as Weizmann, KTH, UCF Sports, Hollywood and HMDB51. Results indicate that the proposed method is well suited to non-upright actions. But the method's accuracy is not close to the state of the art due to some shortcomings like improper foreground detection, improper head size estimate, confusion between similar actions. These shortcomings can be addressed by knowing the pose of the person in the frame. By knowing the posture of the person, the grids can be divided accurately and location of each body can be used to make the proposed action representation more discriminative. The pose information can also be used to compute temporal dynamics of an action, which could be useful to remove confusion between similar actions.

Chapter 4

Likelihood Score Computations for Body Part Detection

4.1 Introduction

Lack of knowledge about the person's pose and location of body parts is affecting the PMF path of the proposed framework. Detection of body parts in the frame can help to estimate the pose. By knowing the pose, the grids can be divided correctly and action representation can be improved by using each identified body part. For this purpose, a body part detection or pose estimation method is required. Most of the available methods require the person to be in upright postures as they impose connectivity or shape constraints to identify the body parts. Also detection of body parts in low resolution videos poses a challenge to many available detectors. But for the proposed action recognition framework, body part detection has to work in non-upright human postures as well as low resolution videos. Therefore, a novel idea to compute likelihood scores for each of the initial classified body parts based on Bayes' theorem using classifier's output value (different from

the predicted class label) is proposed. A body part detection framework with 3 stages is proposed, namely, initial segmentation, likelihood score computation for each segment, a voting strategy that uses inter-frame detected body part segments to remove erroneous detections and detect body parts in the current frame. This will be explained in detail in this chapter together with accompanying experiments to validate the proposed approach.

4.2 Related work

Body part detection and pose estimation can be considered as the same since pose estimation is also able to detect body parts. Researchers have made significant progresses in both these fields. After the success of HOG features [193] in person detection, several methods [26, 194, 195] have used it separately or in combination with Color histogram, LBP [196] or Shape features [197]. Features such as edgelets [198], silhouettes and blobs [199], Haar-wavelets [200, 201] and optical flow [202, 203] etc have also been used extensively for this purpose.

There does not exist a generalized model for human body posture with fixed number of body parts. One early approach [201] used Haar Wavelet features to represent body posture with just four parts, head, leg, left and right arm. This is a straightforward approach developed using example based object detection with SVM classifiers. This method is limited as they assume upright posture of the person based on which constraints are imposed to identify body parts. [204] assumed a 26 part human body model and used a tree-structured deformable part model to represent human pose. They developed an occlusion resistant method by imposing constraints such as symmetric nature of limbs. [202, 205] considered a 10 part human body model. [205] estimates the best suited pose by imposing a global shape on the body part assembly. The constraints imposed on the method

limits the usage of these methods to different postures. Our proposed detector does not impose any constraints and instead is based on Bayesian inference model that estimates a likelihood value for each of the identified body parts.

Pictorial structures [206] models body posture using different parts by taking into account the articulated nature of human body. For posture estimation, [207] combined pictorial structures with densely sampled shape context descriptors, which model body part appearances. Pictorial structures represent body parts as rectangular segments which is not true. Thus to allow deformations in body parts, [208] generalized pictorial structures using deformable shape model. The shape model for each body part includes shape information in a low-dimensional space and also captures pairwise potential between parts. Shape variations for each part are learned independently using PCA and coupled with pairwise potentials. The method employs a non-parametric belief propagation algorithm to estimate pose. In this method, the deformable human body model is learned from 2D projections of a realistic 3D human body model (SCAPE). Pictorial structures are very good in body part detection but require comprehensive and diverse training data to handle all possible posture variations.

To detect body parts, [76, 79] proposed poselets that uses both 2D image appearance and 3D joint configuration to train an SVM classifier. Therefore, these poselet detectors are quite robust to view angle changes and varied appearances. Poselets were originally introduced for image based action recognition. For videos, [77] came up with two types of poselets based on HOG templates and BoW features comprising of SIFT, HoF and HoMB. [209] used poselets as conditioning variables of pictorial structures model to estimate human pose. For training poselets, an extensive manual annotation is required, thus limiting its usage.

In recent years, deep learning has gained significant importance in machine learn-

ing and computer vision community. Deep networks can extract complex features with greater representation power than shallow networks owing to the presence of multiple hidden layers [210]. Thus many researchers have used deep networks for image related applications [169, 170, 211]. Recently, [211] combined CNN with active template regression for human parsing task to semantically label images into different body parts and fashion accessory items. Pose estimation has been posed as a regression problem towards body joints in [169]. All of these methods consider CNN and train the networks with large amount of training data. For instance, [170] uses 1.3 million high resolution images and [211] uses around 7,700 images for training. The main drawback with deep learning is the amount of training data required to get an accurate model.

4.3 Body part classifiers

The objective of a classifier is to learn the intra and inter-class variations accurately. In this case, learning intra-class variations of body parts is extremely challenging as physical appearance of person, color and clothes worn by them vary considerably [201]. Learning variations in appearance of body parts under different postures is also difficult. Other than these variations, handling occlusion and view angle changes is also very challenging. In this work, the human body is divided into 4 constituent body parts, namely, head, torso, arms and legs. The model assumes that the body is made up of one head, one torso, two arms and two legs.

4.3.1 Collection of Training Images

For proper classifier training, it is necessary the training images encapsulate all possible intra-class variations. For this purpose, large amount of training images have to be collected. In poselets based methods [76, 77, 79, 209], classifiers are trained with manually collected and annotated images. Collecting images manually is a strenuous task, thus a python web crawler was developed to gather images from web. The crawler internally used Google AJAX APIs to collect images based on search queries provided. Given a search query, Google AJAX APIs allows access only to the first 64 image results. Therefore to collect a training image dataset that encompass all possible intra-class variations, a number of search queries for each body part class is used. For instance, to collect images for ‘arm’ class, different search queries such as ‘muscular arm’, ‘tanned arm’, ‘bicep’ etc have been used. Apart from the body part categories, a separate ‘None’ classifier with non-body part images have also been collected. Some of the obtained images might be erroneous ones such as cartoons or drawings of body parts, skeletal structures etc., which need to be removed manually. Figure 4.1 shows sample images downloaded for training each body part.

4.3.2 Feature Extraction and Classifier training

The next step is to extract discriminative features and train a classifier that can detect body parts in a reliable manner. To identify discriminative features, experiments have been conducted with three different features, namely, LBP (robust to global illumination), LGP (robust to local intensity variations) and HOG (robust to local pose variations) [68] and the best one is chosen based on the results. To compare HOG features, all images have to be of the same size. Following [193], all training images are resized to 64 x 128 size before HOG extraction. A total of

Figure 4.1: Sample training images collected using web crawler for arm, head, leg and torso

3016 images of each body part and 4029 images of 'None' category were used.

For this experiment, only ELM classifier is used due to its fast training time. The number of hidden neurons needs to be tuned for an ELM classifier. A 4-fold cross validation experiment was conducted on every body part category. The results obtained are tabulated in Table 4.1. For each category and feature combination, the best results are obtained using different number of hidden neurons. HOG features provides the best results for head, torso and leg categories. The performance in 'arm' category is still comparable with the best results obtained. Hence in this work HOG features are chosen for training both ELM and deep learning classifiers since they give best results.

The next step is to train the classifier (ELM/Deep Learning) with HOG features from training images. Deep learning classifiers require lots of training data for high performance. Deep learning can be built based on CNN, RNN, RBM etc.

Table 4.1: 4-Fold Cross validation results using LBP, LGP and HOG features with ELM

Category	LBP	LGP	HOG
head	70.81	70.87	86.18
torso	63.73	64.19	74.44
arm	70.80	69.17	68.94
leg	62.92	61.88	72.21

Table 4.2: Accuracy (in %) Comparison of 4-Fold Cross validation results between ELM and DBN

Category	ELM	DBN
head	91.91	91.51
torso	96.03	93.10
arm	95.16	94.70
leg	95.29	95.68

For this work, Deep Belief Networks (DBN) are used, which have been previously employed for image classification [212] and object recognition [213]. Deep learning's applicability to low resolution images could be studied. Both ELM and DBN body part classifiers were trained using a total of 4000 body part images. For DBN implementation, [214] has been followed with only change in first or input layer. This implementation is generalized to handle input images of any size and even one dimensional data (does not have to be square images). Thus it can be used directly for body part detection only by changing all training images to the above mentioned fixed size. The 4-fold cross validation performance of both DBN and ELM are tabulated in Table 4.2. It is expected that deep learning should perform better but both DBN and ELM have comparable accuracy. This is attributed to the fact that deep learning requires more training data.

Figure 4.2: Proposed Body part detection framework

4.4 Body Part Detection Framework

The proposed body part detection framework is shown in figure 4.2. The framework comprises three main stages, initial segmentation, likelihood score computation for identified segments and inter-frame voting strategy to make a final body part detection result. Segmentation and likelihood score computation are done in single frames and inter-frame voting strategy uses a neighbourhood set of frames to detect the final body parts in a frame. Foreground is assumed to be available or can be extracted easily using an average background subtraction.

4.4.1 Foreground Segmentation

Foreground regions correspond to the whole body of the person. To identify body parts, the first task is to divide the available foreground into smaller segments for initial classification. In this work, two segmentation approaches have been explored, namely, color based mean shift segmentation and newly proposed motion based segmentation algorithm. From each of the obtained segments, HOG features are extracted and classified as either a body part or ‘None’ by the classifiers (either

ELM or DBN). Figure 4.3 shows a sample frame, averaged background subtracted frame and mean shift segments generated.

Figure 4.3: (a) Sample video frame (b) Averaged background subtraction (c) Mean shift segments generated

Proposed Motion Segmentation

Human body has an articulated structure which allows the body parts to move independently. Therefore, motion observed in a scene might be a good approach to segment body parts from obtained foreground. A motion based segmentation algorithm is proposed for each frame. First, optical flow (OF) is extracted using [215]. In low resolution videos, motion features could be noisy. Thus, our proposal is to cluster the motion vectors into basic kinematic features that are more robust. Kinematic features capture motion by combining and aggregating OF in a neighbourhood, thus reducing the effect of noise. For this purpose, Divergence and Curl features are computed for each foreground pixel using equation 3.1.

Motion orientation for each pixel is also computed using x and y components of OF. The orientations are quantized into 8 bins. Each pixel is characterised by the kinematic features and the quantized motion orientation. Using these as features, an RMS error equation is proposed to determine if two foreground pixels belong

Table 4.3: Identifying Consistent Motion Segments Algorithm

Algorithm
<ol style="list-style-type: none"> 1. Start with an initial seed point i. 2. Identify the error value E_{ij} between i and all other foreground pixels j. 3. All pixels with error value less than a predefined threshold are considered to be in the same segment. 4. Repeat steps (1) - (3) for all other pixels which have not been assigned a segment until all pixels are assigned

to the same segment. Sometimes the foreground subject might move only a single body part resulting in erroneous segmentation results. To solve this, we include the distance between the 2 pixels also as a feature. The proposed RMS error equation defining whether 2 pixels belong to same segment is given by

$$E_{ij} = \sqrt{\frac{d_{ij}}{N} \sum_{k=1}^4 (C_i^k - C_j^k)^2} \quad (4.1)$$

where $k = 1, 2, 3, 4$ implies C is Divergence, Curl, motion orientation and distance between pixels respectively. i and j are the two foreground pixels being compared. d_{ij} is the distance between the 2 pixels i and j . N is the total number of foreground pixels. $\frac{d_{ij}}{N}$ is the weighting factor. The proposed segmentation algorithm is summarized in table 4.3. Sample results obtained are shown in figure 4.4.

Figure 4.4: Consistent Motion Segments identified in 2 frames (a) Waving action (b) Walking action, black is the background region, rest denote different identified segments

4.4.2 Likelihood score computation

Classification of body parts based on only HOG features is very difficult as they capture only shape features. As a result, segments can be wrongly identified or can have more than one label. To remove such errors, a likelihood score based on Bayesian inference model is proposed that computes probability that a segment was correctly classified. For predicting body part labels, both ELM and DBN classifiers compute an output for each class. Based on these values the final class is predicted. The final class predicted is simply a class label (0 and 1 in case of binary) whereas the output value is a number indicating certainty of prediction. The output value of the classifier for predicted class of each segment is used as an event in this model to compute likelihood scores. The Bayes' theorem states

$$P(A|B) = \frac{(P(B|A) * P(A))}{P(B)} \quad (4.2)$$

where A and B are two events. $P(A|B)$ and $P(B|A)$ represent the conditional probability of occurrence of event A, given event B has already occurred and vice versa. $P(A)$ and $P(B)$ are the individual probabilities for occurrence of event A and B respectively. The terms on the right hand side of the equation are prior probabilities and can be computed based on available data. The proposed likelihood score can be expressed as the following equation

$$P(x|f2, f1) = \frac{(P(f2|f1, x) * P(f1|x) * P(x))}{(P(f2|f1) * P(f1))} \quad (4.3)$$

where x is the event that a segment is classified as a body part (arm, head, leg,

torso), f_2 is the event that a segment is classified as 'None' category as well and f_1 denotes the range in which the classifier output value falls for that segment. To compute the likelihood score in the equation, prior probabilities in the right hand side of the equation need to be calculated.

4.4.3 Computation of Prior probabilities

The prior probabilities for ELM and DBN are computed using the classification results on the available training images. These images are passed through both body part and 'None' classifiers and results are collected. For body part classifiers (ELM & DBN), the output value is also collected. To compute the prior probabilities, the range of classifier output values is divided into several bins using non-linear step size. For instance, in case of ELM, most output values lie in the range of $[0-0.2]$, therefore, a smaller bin size of 0.02 is used in this range and a bigger bin size of 0.2 for values more than 0.2. A similar binning strategy is employed for DBN as well. Using this binning strategy and the classifier output values for each training image, prior probabilities $P(f_2|f_1, x)$ and $P(f_1|x)$ are computed. To compute $P(f_2|f_1)$ and $P(f_1)$, body part and 'None' classifiers are run on both positive and negative training images. The classifier output values obtained are once again divided using the same binning strategy. The mentioned prior probabilities are estimated from the data collected from training images and applied to the equation 4.3 to compute likelihood score for each segment.

In the proposed framework, human body model assumed has 6 body parts comprising of a head, torso, two arms and two legs. Hence, $P(x)$ can be computed

using the equation

$$P(x) = \frac{t}{N} \quad (4.4)$$

where t is 2 for arms, legs and 1 for head and torso and N is the total number of segments in the frame (from mean-shift or motion segmentation).

Using the prior probabilities and $P(x)$, likelihood score can be calculated for each segment generated for a given video frame. Figure 4.5 shows a sample frame where the likelihood for each detected arm with ELM in the frame is shown. Even though there are many arms detected in the frame, the highest likelihood score is obtained at the actual arm location.

Figure 4.5: Arms detected in a frame with their respective likelihood scores (ELM is used)

4.4.4 Inter-frame Voting Strategy

During initial testing, two issues could be seen in the likelihood score computation process. Firstly, not always the correct body part location gets the higher

likelihood and secondly, in some frames, one or more body parts are never identified. First issue is mainly a result of using only classifier's output value as events for likelihood score computation instead of using clues from the video frame such as appearance, motion etc. To reduce the effect of these issues, an inter-frame voting strategy is used. The notion behind this approach is that body parts can be correctly detected in neighbouring frames. When a body part is wrongly detected or completely missed, the proposed strategy uses the body part detections in neighbouring frames and aggregate their respective likelihood scores to identify the final body part locations.

The following steps explain the proposed inter-frame voting strategy:

- Step 1: For frame (F_i), a frame neighbourhood comprising of 5 frames ($F_{i-2} - F_{i+2}$) is defined. The number of frames in this neighbourhood is not a critical parameter as it affects only the speed of action and it also depends on the frame rate.
- Step 2: For any body part detected in these frames, likelihood scores are computed.
- Step 3: In each frame, top 3 arm, head, leg and torso locations are collected based on the likelihood scores. These will be used for majority voting.
- Step 4: Each body part likelihood scores are normalized separately. In each frame, the likelihood scores for each body part (arm, head, leg, torso) are normalized based on the maximum likelihood score obtained for that body part, say m . Each of the likelihood scores for that body part (arm, head, leg, torso) are divided by m and normalized to the range $[0 - 1]$.
- Step 5: During voting, each arm segment is compared to all other collected arm segments. If the intersection between the locations of 2 arm segments is greater than 70%, both are considered to be the same body part segment

and vote is given to both of them. The higher likelihood value among both the segments is used as the vote in the incrementing process.

Step 6: At the end of voting, the segment with most votes is chosen as the arm.

Step 7: Repeat steps 5 and 6 for head, leg and torso.

Upon completion of voting, one head, one torso ,two arms and two legs with maximum votes are chosen as the final body part detected for each frame.

4.5 Experimental Results and Discussion

Most available human pose estimation datasets are of high resolution [216,217] or available for only images [217,218]. Here our intention is to test proposed body part detection framework under different pose variations and also in low resolution videos. For this purpose, action videos from NUAD dataset, which is shot in 192x108 resolution and exhibits a lot of pose variations (upright & non-upright). Also the person in the frame will be very small compared to the total video size. Out of the available 35 subjects, only 5 are used for this testing. From a total of 120 videos, the first 10 frames are chosen giving a total of 1200 frames to test the proposed body part detection framework. The manually marked foreground regions are used. In order to perform the comparison, four different variations of the body part detection framework are run on the frames, 1) ELM and mean shift (ELM Mean), 2) DBN and mean shift (DBN Mean), 3) ELM and proposed motion segmentation (ELM Motion), 4) DBN and proposed motion segmentation (DBN Motion)

The body part detection performance is measured as amount of correct head, torso, arm, 2 arms, leg, 2 legs detections for all the above mentioned 4 frameworks. The results are tabulated in table 4.4. Using the available ground truth markings

Table 4.4: Performance (in%) comparison of the four different variations mentioned, ELM Mean, DBN Mean, ELM Motion and DBN Motion

Correct De- tections	ELM Mean	DBN Mean	ELM Mo- tion	DBN Mo- tion
head	20.83	23.08	52.58	20.2
torso	45.67	40.83	59.83	51.41
arm	44.17	45.92	65.17	78.17
2 arms	24.33	29.83	41.75	53
leg	52.17	56.92	67.85	77.33
2 legs	33.50	42.17	58.25	67.08

for each body part's location and detected locations in the frame, the percentage of body part correctly detected and missed can be computed. A detected body part location is considered to be correct, if more than 70% of the body part is correctly detected and less than 70% of the body part is missed.

From table 4.4, some interesting observations can be made. DBN Motion give best results in limb detections and ELM Motion gives best results for head and torso detection. As there are no connectivity constraints imposed, the proposed method is able to detect non-upright postures. The inclusion of the proposed motion based segmentation scheme increases the detection performance considerably. Our idea of segmenting separate body parts using motion can help to improve the accuracy of body part detection with trained classifiers. But this scheme also suffers from over or under segmentation errors as shown in the figure 4.4. More than one body part might be present in a segment (head and arm in figure 4.4 (a)) or a single body part might be separated into several segments (torso in figure 4.4 (b)) making detection difficult.

When compared to ELM, DBN classifiers shows better performance in detection because of it's better representation power. Figure 4.6 shows body part detection results obtained in the video frames. It can be observed that results obtained using DBN are better in the first two rows when compared to ELM than the

next two rows. From the table 4.4, we observe that leg and arm detection are better than ELM. Head and torso are at the centre of human body and so when segmented these body parts are either over or under segmented making classification difficult. Also in low resolution, the body parts might not appear clearly resulting in confusing appearance of features. This could also result in erroneous or misdetection of body parts. For accurate classification, deep learning requires huge training data but our DBN classifiers are trained with only 4000 positive and negative images each. If DBN is trained with a larger dataset, it would improve the detection performance.

Depending upon the action being performed, some of the body parts might be partially occluded by other parts (head gets occluded by hands during waving action in side view) or appearance of body parts changes due to their motion (head is not visible when bend down in front view). This is observed in both ELM and DBN as shown in figure 4.7.

4.6 Concluding Remarks

In this chapter, to attain knowledge about the pose of the person, a body part detection framework has been introduced. Just using HOG features and classifiers (ELM/DBN), a body part likelihood score has been proposed based on Bayes' theorem for conditional probability. In the detection framework, an inter-frame voting strategy was used to track the body parts even when separate body parts are not detected for that frame. Since no shape or geometrical constraints are imposed, this framework can detect body parts in non-upright postures. Experiments show that motion segmentation performs better than mean-shift segmentation. Also, ELM performs better in head and torso detection whereas DBN performs better in limb detection.

Chapter 5

Feedback Path for Pose-Invariant Action Recognition

5.1 Introduction

With only the PMF path, the proposed action recognition framework is able to extract pose-invariant kinematic motion features that is partially view-invariant and generalized to handle non-upright human posture actions. From the experiments, several insights and shortcomings of proposed PMF path were identified. Main reasons for reduction in performance were found as improper estimation of head size for grid division, confusion between similar actions and lack of information on posture and body parts. With knowledge on posture of the person, action representation can be made more discriminative.

For identifying body parts in non-upright postures, body part detection framework with no shape or connectivity constraints was introduced in Chapter 4. Even though proposed body part likelihood score is able to identify body parts correctly, it does not detect the complete posture correctly in all frames. In this chapter,

Canonical pose feedback (CPF) path of the proposed framework is introduced. The intention of this feedback path is to identify the posture of the person using the initial action hypothesis of PMF path and body part likelihood score. CPF path aims to use the identified posture to realign the grids and correct the motion features in these grids. Also a new set of body part referenced kinematic motion features and temporal stick features are introduced to make action representation more discriminative.

5.2 Canonical Pose Feedback path

PMF path computes features initially based on an approximate head size that governs the location of each grid which is the main component for pose-invariant feature extraction. CPF path aims to correct these grids by using the pose cue observed in the frame. With knowledge on posture, motion of each body part can be captured to obtain an accurate representation of an action. To capture the dynamics of an action, temporal stick features are introduced with the help of the pose identified for each frame. In the following subsections, canonical stick pose generation and representation, pose hypothesis computation algorithm and how motion-pose cues reinforce each other are explained.

5.2.1 Canonical Stick Pose Generation and Representation

Using the features from PMF path an initial action is recognized, which will be the input for the CPF path. Any action is characterized by a series of poses, termed here as canonical poses. Given the video of an action, the probability of observing these canonical poses is very high. Therefore canonical poses for any action can be used to identify the posture in the frame. To extract these poses, an one-time

Figure 5.1: Canonical stick pose marked on the obtained canonical pose image

offline algorithm is introduced in this subsection. This algorithm is applied to available training videos for each action to extract the canonical poses for that action. Table 5.1 gives step by step description of this algorithm. The proposed method is based on non-negative matrix factorization applied to data collected for each action. Some of the canonical poses obtained for the corresponding actions are shown in the figure 5.2. We observe that the poses obtained for every action mark important moments in an action and capture each body part's location properly. Depending upon the training frames used, the proposed algorithm generates poses that are combination of poses from different views.

The canonical poses obtained are converted to stick man poses by manually marking them as shown in figure 5.1. To allow scale-invariant comparison with any frame, a new normalized stick pose representation is introduced. Referring to figure 3.2, the farthest point from neck in the body is $11x$ (Torso + leg size), where x is dependent upon head size. Once the body part markings are complete, head size can be determined from which $11x$ can be estimated. With neck point as origin and $11x$ as maximum, all sticks are normalized. Since these sticks can be scaled according to any estimated head size, comparison can be made with any frame.

Table 5.1: Canonical Pose Generation for a Given Action

Algorithm
Data Generation for each action <ol style="list-style-type: none"> 1. Each training frame F_i cropped to contain only foreground and convert it to grayscale. 2. Compute the gradient image of the grayscale image and threshold it with mean gradient value. 3. Resize the threshold image to a fixed dimension. 4. Apply Non-negative matrix factorization on the data and collect the first N canonical poses. <p>The data for each action comprises all training frames F_i passing through the first 3 steps and in the step 4 canonical poses are generated.</p>

5.2.2 Canonical Stick Pose Hypothesis

The proposed canonical stick pose representation comprises of 6 body parts, head, torso (represented using 2 points), left and right arm and leg (represented using 3 points) as normalized coordinates. The occluded body parts are also marked in the canonical sticks. In this sub-section, canonical stick pose hypothesis scheme to identify the most probable canonical stick pose present in the current frame based on a few observations and constraints is explained.

Pose can be defined as how different body parts are arranged with respect to each other. Pose hypothesis scheme intends to check if body parts present in canonical stick pose are located at the same region in the frame. The body part detector discussed in chapter 4 that provides a likelihood score l_i that the image segment is body part i as a basis for this pose hypothesis computation. In our implementation, ELM based body part likelihood score is selected as they offer better head and torso detection. DBN based approach still needs more training data, which is difficult to collect, for better body part detection. This will be a part of our future work.

Figure 5.2: Canonical Sticks extracted for different actions in different datasets

Using extracted kinematic motion features, an initial segmentation of foreground is done (using section 4.4.1). The body part detector is applied to each segment to identify if they belong to any body part. Assuming that M initial segments were generated using motion based segmentation, the pose hypothesis for each pose PH_N is a summation of all the likelihood for each body part in the stick pose subjected to distance and orientation constraints is given by the equation 5.1

$$PH_N = \sum_{i=1}^6 \sum_{j=1}^M L_{(i,j)} \quad (5.1)$$

$$\text{where } L_{(i,j)} = \begin{cases} l_{(i,j)} & \text{if } d_j \leq \tau_d \wedge (\theta_j - \theta_{i,N}) \leq \tau_\theta \\ 0 & \text{otherwise} \end{cases}$$

where M is the number of initial segments generated using motion based segmentation and there are N canonical sticks for an action. CPF path begins with the approximate head size. In order to impose the distance constraints, each segment's corners are converted to normalized stick coordinate based on neck and approximate head size. The smallest distance among the four corners of the segment (d_j) is chosen and a distance threshold ($\tau_d = 0.05$) is used for this constraint. For body part orientation constraint, edge image of the frame is divided into smaller windows. Eigenvectors are computed for edge pixels in each window using covariance matrix. The body part orientation is estimated using Eigenvectors. If the difference between body part orientation (θ_j) determined in image segment j and stick pose ($\theta_{i,N}$, denotes i th body part in stick pose N) are within a certain threshold ($\tau_\theta = 10$), then likelihood score $l_{i,j}$ is included in the pose hypothesis. If a segment j is classified as body part i with likelihood score $l_{i,j}$ and satisfies the distance and orientation constraints, it accumulates with likelihood for each body part in stick pose $L_{i,j}$.

Based on this pose hypothesis scheme 5.1, likelihood value PH_N for each pose N being in the frame is computed. Top 3 poses with the highest hypothesis are chosen and a mean pose is computed. The pose hypothesis is calculated in normalized stick representation for only the visible parts making it robust to occlusions and partial canonical sticks extracted. Figure 5.7 shows some sample frames with pose identified in presence of occlusion and partial images.

The mean pose obtained is used to estimate the accurate head size and body orientation. For this purpose, a pixel wise segmentation is performed, where each pixel is segmented as one of the 6 body parts based on which body part's stick they are close to in the stick pose. The correct body orientation is estimated using obtained torso region and neck point. The body orientation and estimated head region is used to get a correct estimate of head size. The pose hypothesis scheme is

repeated with updated head size to obtain the correct pose for each frame until the difference between the initial and computed head size is less than a predetermined threshold. Table 5.2 summarizes pose hypothesis scheme to identify pose for a given video frame.

5.2.3 Computation of Body Part Referenced Kinematic Features

The obtained canonical pose cue is used to improve kinematic features in two ways. Firstly, from the obtained canonical pose, grids are realigned such that they correspond to head, torso and leg regions respectively. Due to this re-aligning, contribution of motion features in the different cells of the grids would vary. Secondly, Body part referenced projection ($BodyProj(p_i)$) and rotation ($BodyRot(p_i)$) features for a pixel p classified as body part i are computed using the following equation

$$BodyProj(p_i) = O\vec{F}_{p_i} \cdot \hat{P}_c \quad (5.2)$$

$$BodyRot(p_i) = \hat{P}_c \times O\vec{F}_{p_i} \quad (5.3)$$

where for each body part i , the unit vector (P_c) is obtained from the center pixel of the region obtained using pixel-wise segmentation for that body part. This feature provides more information on motion of each body part with respect to the center of the body part thus improving action representation. The weighted and un-weighted histograms for body part referenced features are also computed for each human-body centric space dimension of each cell in each grid.

Table 5.2: Canonical Pose Hypothesis for a Given Frame

Algorithm
<ol style="list-style-type: none"> 1. Start with likelihood score L_i for each part i in stick pose as 0, $L_i = 0$. 2. Using kinematic motion features, obtain an initial segmentation of the foreground region. 3. Pass each of these segments through the ELM body part classifier from Chapter 4, to know if the segment is a body part i or not. 4. If segment m is detected as body part i associated likelihood score $l_{i,m}$ is obtained. 5. If segment m satisfies distance and orientation constraints, the likelihood score L_i for body part i in stick pose is accumulated by $l_{i,m}$ (Distance constraints are imposed in normalized stick coordinates). 6. Repeat steps 3 - 5 for every segment and obtain the final L_i for the canonical stick pose. 7. The pose hypothesis PH_n for canonical stick pose n is summation of all L_i. 8. Repeat steps 1 - 7 for every canonical pose n and compute pose hypothesis PH_n. 9. Choose the top 3 poses with highest pose hypothesis and compute the mean pose. 10. Perform a pixel-wise segmentation into one of the body parts based on the distance from each body part's stick in the mean pose. 11. Compute the body orientation using obtained torso region and neck point. 12. Compute the head size using obtained head region and body orientation. 13. Repeat steps 5- 12 if computed head size and initial approximate head size vary by more than a threshold.

5.3 Temporal Stick Features

In addition to the neck point, canonical stick pose comprises of 16 points (2 each for head and torso, 3 each for each limb). To capture dynamics of an action, temporal stick features that capture how each stick joint and body part is moving during an action is introduced.

Once the canonical stick pose for each frame is computed, the whole video volume is divided into T temporal segments. Using the stick poses obtained and the available neck points for the frames in a segment an average stick pose and neck point is computed for the segment. The average stick pose and neck between consecutive segments is used to compute the motion. Projection and rotation features can now be computed for each point in the stick pose with respect to the neck point in every temporal segment. The idea behind temporal stick features computation is illustrated in figure 5.3.

These features are encoded in a view-invariant manner in the same way as mentioned in subsection 3.3. The weighted and un-weighted histograms for each stick pose joint and body part (by combining stick pose joints) are computed. These histograms are combined with previously computed histograms for local dynamics (div & curl), referenced motion (Proj & Rot) and body part referenced motion (BodyProj & BodyRot) to form final action representation. Table 5.3 summarizes proposed scheme to extract temporal stick features.

PMF and CPF complete a loop-like structure that is repeated iteratively to identify the best possible action hypothesis. BodyProj, BodyRot and temporal stick features are added to more discriminative representation of each body part's motion and temporal dynamics. Due to addition of these discriminative features, the loop iteratively reinforces action hypothesis. As the canonical poses can be shared between 2 actions, for instance, running and jogging poses. Thus when an ac-

Figure 5.3: Computation of Temporal Stick Features from video

tion is recognized wrongly, the pose obtained in canonical pose hypothesis is more likely to be that one of these shared poses. The pose identified by canonical pose hypothesis is either a shared canonical pose (with actual action) or an erroneous pose. Error in each iteration of the loop is computed based on this notion and inherently captured in CPF. The identified poses determine how well the grids are realigned and body part motion are represented, which is reflected in the action hypothesis. Thus, the proposed iterative framework uses action hypothesis as stopping criterion. When action hypothesis does not vary between 2 consecutive iterations, the framework has converged to final action result.

Table 5.3: Temporal Dynamics of Stick Pose Joint and Body part

Algorithm
<p>Input: Canonical stick pose for every frame in the video.</p> <ol style="list-style-type: none"> 1. Divide the video volume into T temporal segments. 2. For every temporal segment t, <ol style="list-style-type: none"> (a) Using the canonical stick pose and neck point for every frame in the segment, compute an average stick pose and neck point. (b) Compute the motion of each stick pose joint and neck between consecutive segments. (c) Projection and rotation features are computed for each stick pose joint with respect to the neck in each temporal segment. (d) View-invariant encoding of the features using the human-body centric space. 3. Weighted and unweighted histograms of the features for each stick pose joint is computed. 4. Weighted and unweighted histograms of the features for each body part (considering more than one stick pose joint) is computed.

5.4 Experimental Results and Discussion

Experimental results were provided in section for initial framework with only PMF path. From initial results, it was observed that initial framework can be improved by adding posture information. Therefore, CPF path was included to obtain posture of the person and to complete the overall reinforcement framework. To test pose-invariance and improvement obtained after including CPF path, experiments are conducted on 6 datasets including Weizmann [2], KTH [58], UCF sports [3], Hollywood [43], newly introduced NUAD and MSR Action dataset [219]. With MSR action, the intention is to test if canonical sticks from another (KTH) dataset can be directly used. For each dataset, foreground is assumed to be available or can be extracted easily. When foreground masks are not available, average background subtraction method is applied to either available or manually marked person bounding box. The neck and major viewing direction of the person is as-

sumed to be available. For temporal stick features, video segments are divided into $T = 5, 10$ temporal segments and both are included in the final action representation. $N = 100$ canonical stick poses are extracted for each action and are manually marked as per section 5.2.1. For classification, ELM classifier is employed. Two parameters need to be fixed, namely, number of hidden layer neurons and activation function. For all experiments, sigmoid activation function is used and number of hidden layer neurons are varied from 200 – 40000 and the best performance obtained is reported in the following subsections.

5.4.1 Weizmann Dataset

In Weizmann dataset, framework is tested in a leave one person out validation approach. The performance of our framework with only PMF path, with CPF path included and comparison with other state of art methods are provided in table 5.4. Upon including the CPF path that corrects the grids and includes motion of each body part, the proposed framework's performance increases to 100% which is comparable with the state of the art. The proposed method is primarily based on kinematic features for action recognition that uses pose as a cue to quantify motion of each body part during an action. From table 5.4, proposed method after inclusion of CPF path is comparable with state of the art methods that use both shape and motion features [9, 74]. Confusion between similar actions like *skip-jump*, *jump-walk* is not seen due to correct grid division and information on motion of each body part.

5.4.2 KTH Dataset

KTH dataset is tested in two settings, 16 train + 9 test suggested split in [58] and leave one person out (LOV) setting. Results for both settings and compar-

Table 5.4: Performance Comparison for Weizmann dataset, Proposed method uses foreground masks given with dataset

Method	Performance (%)
Proposed Framework (only PMF)	92.47
Proposed Framework (PMF + CPF)	100
[9]	100
[74]	100
[85]	99
[64]	98.9
[220]	95.75
[6]	93

ison with the state of art methods are tabulated in Table 5.5. The six actions can be classified into 2 classes, hand actions (boxing, handclapping, handwaving) and leg actions (jogging, running, walking). After inclusion of CPF path, the performance of our method has increased in both settings. Confusion that existed between *handclapping-handwaving* and leg actions have reduced considerably showing inclusion of pose information is useful.

Under the first setting, our proposed method is comparable with the other approaches like deep learning [221] and subspace analysis [222]. It might be noted that deep learning and subspace analysis would provide more discriminative power to their features. In the second setting, several other methods have superior performance compared to our proposed framework, which can be attributed to subspace analysis [223,224], system based modelling approaches [110,223], combining pose/shape with motion features [9,51], usage of contextual features [55,114]. The proposed method is based on motion only feature and is comparable with methods that use only motion [220,225]. Still confusion is mainly seen between running and jogging, which is very difficult to distinguish with motion feature alone. Other reasons for action recognition failures are improper foreground detection and failure to detect a person using [190] in many frames of jogging, walking and running videos.

Table 5.5: Performance Comparison for KTH dataset

Method	Performance (%)	
	16 + 9 [58]	LOV
Proposed Framework (only PMF)	87	90
Proposed Framework (PMF + CPF)	90	93.32
[223]	-	99
[224]	-	98.3
[114]	-	96.16
[9]	-	95.77
[51]	-	95.5
[110]	-	95.17
[55]	-	94.5
[220]	-	87.7
[50]	89.34	93.18
[221]	94.39	-
[222]	93.6	-
[225]	93.3	-
[43]	91.8	-

5.4.3 UCF Sports Dataset

In UCF Sports dataset [3], the person bounding box and neck point are manually marked. Average background subtraction is applied within the bounding box to obtain foreground. UCF sports is a very difficult dataset due to huge amount of pose variations, view angle changes and moving cameras exhibited in the broadcast videos. Leave one out setting is used for testing these methods.

When CPF path is included, the performance increase of 11% is observed in the proposed framework. The results of our proposed method and the state of art are tabulated in Table 5.6. Figure 5.5(a) shows confusion matrix obtained using complete framework and it shows that actions *diving*, *lifting*, *walk*, *Swing-Bench* and *Swing-SideAngle* are always detected correctly. It is to be noted that main confusion is observed mainly between *skateboarding-walk*, *running-kicking*. The main reason for this confusion is that leg movement in most skateboarding videos are similar to walk and kicking also involves an initial running before the subject

Table 5.6: Performance Comparison for UCF Sports dataset, Proposed method uses manually marked bounding box

Method	Performance (%)
Proposed Framework (only PMF)	76.31
Proposed Framework (PMF + CPF)	87.4
[223]	96.6
[226]	90
[138]	88.7
[227]	88
[9]	88
[85]	88
[110]	87.3
[228]	86.6
[50]	81.5
[136]	75

kicks.

From table 5.6, the reinforcement framework has comparable performance with the state of the art methods in the range of 86 - 88%. [223] shows the highest performance of 96.6%, which can be attributed to the ARMA modelling used to model the actions within the Grassmanian manifold. When ARMA modelling is not used the accuracy is 88.7%, which is still comparable with our method.

The proposed CPF path is dependent on the stick poses extracted for each action. These stick poses determine how well the pose in the frame can be identified by our method. The framework's pose-invariance is attributed to two reasons, namely, pose-invariant motion features and the variation in the canonical poses extracted from the training videos. To test the pose-invariant nature of the framework, an experiment was conducted with the canonical sticks extracted. For each action, training videos are separated by their view angles and canonical sticks are extracted for each view and manually marked. The proposed complete framework is run with lesser number of canonical sticks to test the pose-invariance. Only front facing sticks are used for Golf-Swing, Kicking, Riding-Horse, Running,

Figure 5.4: Poses identified using all canonical sticks (Top) and using lesser canonical sticks (Bottom), *Head - Green, Torso - Yellow, Arms - Blue & Red, Legs - Cyan & Magenta*

SkateBoarding and Swing-Bench actions. Left facing sticks are used for Diving, Swing-SideAngle as only side view videos are available. Only half of the extracted sticks are used for Walk and Lifting since they are done in frontal view only. A decrease of 4% is observed in the performance of proposed method. Figure 5.5(b) shows confusion matrix with lesser canonical sticks that shows lot of confusion between actions as available poses are not enough to handle pose variations. Figure 5.4 also shows a comparison of poses detected using all available and lesser canonical sticks. This indicates that our proposed method can handle pose variations if enough canonical sticks are available for each action. From the figure 5.4, some cases are observed where pose is wrongly identified with lesser canonical sticks, but grid division is still correct.

It is also to be noted that several videos in this dataset like *Golf-Swing-Back-002*, *Run-Side-009*, *Riding-Horse-009*, *Skateboarding-Front-007* are single instances of these actions performed in back, front, front and back views respectively in the dataset. The method is still able to detect these actions correctly even when the training videos do not contain the same view. Walk action is depicted only in

frontal direction. Due to this, when actions are wrongly detected as *Walk* by PMF path, the pose identified for each frame by CPF path was not correct as only frontal sticks were available. Upon including side view *Walk* sticks extracted from Weizmann and KTH dataset, the framework was able to identify the actions correctly instead of confusing with *Walk*. This re-iterates proposed framework's pose-invariant nature.

Figure 5.5: UCF Sports dataset Confusion Matrix (a) Complete Framework (PMF + CPF) with all canonical sticks (b) Complete Framework with lesser canonical sticks

The proposed framework consists of four classes of motion features projected in the pose-invariant space to recognize action. The four classes include (a) divergence + curl, (b) projection + rotation, (c) body part referenced projection + body part referenced rotation and (d) temporal stick features to quantify observed motion with respect to local pixel neighbourhood, fixed neck reference point, body part center and temporal evolution of stick poses respectively. During ELM classification, each class of feature was removed and different combination of features was used to recognize the action. The performance obtained with different combination of features is tabulated in table 5.7. It is clear from the results that the performance drops when even one of the features is removed. The maximum decrease is seen when temporal stick features are removed since they capture implicit dynamics of actions such as *kicking*, *walking*, *skateboarding*. When the other

Table 5.7: Effect of feature in pose-invariant action recognition using UCF sports dataset (a) div + curl, (b) proj + rot, (c) BodyProj + BodyRot, (d) temporal stick features

Feature Combination	Performance (%)
(a) + (b) + (c) + (d)	87.4
(a) + (b) + (c)	77.62
(a) + (b) + (d)	83
(a) + (c) + (d)	82.67
(b) + (c) + (d)	81.04

features are removed other actions such as *Swing-SideAngle*, *Diving*, *Run* etc are confused. This indicates all four features combined together form a discriminative action representation even in the presence of pose variations.

5.4.4 Hollywood Dataset

Hollywood dataset is a publicly available dataset that depicts 8 actions (SitUp, SitDown, StandUp, Kiss, HandShake, HugPerson, Answer Phone, Get Out Car) collected from 32 movies. This is a very challenging dataset due to occlusion, view angle changes, subtle movements to depict actions, cluttered background. For handling occlusions, section 3.4.4 is followed for PMF path.

To tackle occlusion, slight adjustments are made to CPF path. When pose hypothesis is generated for each pose, limits are imposed on the sticks. For this purpose all calculations are done in normalized stick coordinates. From the extent of foreground region and the neck point, maximum and minimum limits for x and y direction can be determined. All sticks are subjected to these limits and pose hypothesis is computed only for the region within the limits. Figure 5.7 shows poses identified in sample images of Hollywood dataset when only partial body is observed that shows our method is able to account for occlusions.

The features are extracted for the required grids only as mentioned in table 3.3

Figure 5.6: No motion & neck visible in Hollywood dataset

of section 3.4.4. From table 5.8, it is clearly seen that our proposed method is comparable with the state of the art methods indicating pose-invariant nature of method and ability to handle occlusions. Upon investigation of error cases, some video clips depict actions in a subtle manner or motion of the body part is not at all visible, for instance, motion of hands not seen in HandShake action. In some videos the neck of the person is not visible to extract kinematic features resulting in erroneous detection. See figure 5.6 for some example error cases.

As with UCF Sports, the canonical sticks extracted determine our method's performance. During initial testing, it was observed that due to the presence of occlusion, variation in lighting, cluttered background, the sticks obtained were not satisfactory to account for different poses observed in an action. Also the number of action videos in each view differs from 0 to 27. This resulted in confusion between actions. Therefore, sticks were included for certain actions like SitDown, StandUp, SitUp from other datasets. For other actions, suitable sticks could not be found from other datasets. Upon inclusion of these sticks, increased performance was observed with our framework as shown in table 5.8. This reiterates our method's pose-invariant nature and ability to handle occlusion. It also shows that canonical stick poses used are generalized (can be used from other datasets) and must be able to handle all possible pose variations.

Table 5.8: Performance Comparison for Hollywood dataset, Proposed method uses manually marked bounding box

Method	Performance (%)
Proposed Framework (only PMF)	54.12
Proposed Framework (PMF + CPF (With stick poses from other dataset))	56.87
Proposed Framework (PMF + CPF (Without stick poses from other dataset))	51.89
[229]	59.9
[55]	53.3
[43]	38.2

Table 5.9: Performance Comparison on HMDB51 dataset, proposed method run on subset of actions, others are run on whole dataset

Method	Performance (%)
Proposed Framework (only PMF)	63.03
Proposed Framework (PMF + CPF)	66.29
[178]	69.4
[175]	59.4
[179]	69.2
[180]	65.2
[177]	65.9
[230]	63.3

5.4.5 HMDB51 dataset

HMDB51 dataset includes 51 action classes under five broad action categories. The proposed action recognition framework is mainly suited for body movements and not suited for facial motion, facial motion with objects and interactions categories. For the ‘body movements with objects’ category, the proposed framework is not suited as it cannot capture object contexts or differentiate between objects and person. A subset of 19 actions belonging to general body movements category are chosen and the proposed framework is applied to it. For handling occlusions, section 3.4.4 and section 5.4.4 are followed. Only actions *clap*, *turn* and *wave* require grid 1 and 2, all other actions use all 3 grids for recognition. For actions common with other datasets, canonical stick poses from those datasets are added.

Figure 5.7: Poses identified in Hollywood dataset when part of body is not visible or occluded, *Head - Green, Torso - Yellow, Arms - Blue & Red*

The 3 training/testing splits provided are used for measuring the accuracy of the proposed method. The performance of available state of the art and proposed method are tabulated in table 5.9. The proposed method is run only on a subset of actions and so cannot be directly compared to state of the art. For a proper and fair comparison, both the proposed framework and state of the art [178] must be compared in the general body movements action category alone. This is because proposed method is not suitable for other actions where appearance and object context are also important. Thus using the recognition scores provided by [178],

performance was calculated for each action category in each training/testing split. The results of the proposed method and [178] on general body movements category is given in table 5.10. Spatial and temporal ConvNets focus on appearance and motion cues respectively in [178]. Following [178], score weights of 1 and 1.5 are used for fusing the spatial and temporal ConvNets prediction scores. The proposed framework based on only motion has a comparable performance to state of the art with both ConvNets. Since both appearance and motion cues are used in [178], they have better performance. When we use only temporal ConvNet (motion) for prediction, the proposed framework has better performance in all 3 splits. Figure 5.8 shows some of the poses obtained in HMDB51 and shows the proposed framework is capable handling pose variations, occlusions and view point variation.

Upon analysis, it was noticed that major confusion was *dive* as *jump* or *climb* and *fall_floor* as *sit* actions. The videos show that dive involves jump towards ground or ball or with rope attached to the person. For proper recognition, it would be necessary to capture appearance or context information like objects involved, ground or scene estimation clues. A similar reason can be attributed for confusion between *fall_floor* and *sit* as one involves falling to ground and other involves a chair. Similar to Hollywood dataset, neck point is not visible in many videos or important moving body parts are occluded which makes recognition difficult.

Most available deep learning approaches [175, 177–179, 230] employ pre-training with available standard ImageNet database to better learn appearance and context. When pre-training is skipped, the performance of these methods reduces. Since non-upright action or pose database available are quite limited and pre-training might be more difficult for these cases. Two stream architectures fail to register appearance with optical flow and how these cues evolve over time [179]. In the proposed method, both pose and corresponding body part motion and how

Figure 5.8: Poses identified in HMDB51 using available canonical sticks, *Head - Green, Torso - Yellow, Arms - Blue & Red, Legs - Cyan & Magenta*

they move with respect to each other can be obtained. The proposed motion features will be an ideal candidate for temporal ConvNet but needs to be improved to include context or object motion. This will allow us to recognize actions in all categories of HMDB51.

5.4.6 NUAD

NUAD is a newly proposed non-upright action dataset introduced in chapter 3 with 8 actions done by 35 subjects in 3 views. The proposed complete framework

Table 5.10: Performance Comparison for HMDB51 dataset for General Body Movements category, Proposed method uses manually marked bounding box

Method	Split 1(%)	Split 2(%)	Split 3(%)	Average(%)
Proposed Framework (only PMF)	62.61	62.15	64.34	63.03
Proposed Framework (PMF + CPF)	65.04	66.19	67.63	66.29
[178] (spatial + temporal)	69.12	68.42	68.25	68.6
[178] (temporal)	64.91	64.56	67.19	65.55

Table 5.11: Performance of Proposed framework in NUAD, Proposed method uses ground truth person location

Dataset	Performance (%)
NUAD	91.4

is tested on this dataset in a leave one person out approach. From table 5.11, proposed framework shows performance increase of 1.3%. Non-upright postures observed in dataset are well handled by the framework. The confusion observed between similar actions in PMF path has reduced considerably with inclusion of CPF. For some of the error cases such as *bend-squat*, *climber-pushup*, when body parts are occluded, pose obtained might be correct but motion of occluded part is not available. Due to this actions cannot be recognized correctly.

5.4.7 MSR Action Dataset

MSR Action Dataset [219] is mainly action detection dataset consisting 16 videos with 63 different instances of 3 actions (boxing, handclapping, handwaving). The objective of the dataset is action localization/detection. Using this dataset, the ability to use cross dataset canonical sticks is tested. The groundtruth provided with the dataset is used to extract features for each instance of an action and canonical sticks extracted for the 3 actions in KTH dataset are used.

Table 5.12: Action Recognition Performance in MSR Action dataset, Proposed method uses manually marked bounding box

Method	Performance (%)
Proposed Framework (only PMF)	90.2
Proposed Framework (PMF + CPF)	92.46

Table 5.12 shows performance of the framework. Accuracy of 92.46% indicates that canonical sticks from KTH dataset can match very well here. This is an added advantage of proposed normalized stick pose representation. As with hand actions of KTH dataset, only handclapping and handwaving are confused with each other since they are done in frontal direction. Also some reasons for erroneous action recognition could be improper foreground detection, movement from other persons, body part detection errors etc.

5.4.8 Discussion

Effect of Parameters

In the proposed PMF and CPF paths, few parameters have been assumed such as number of grids, cell configuration or number of cells, distance and orientation thresholds imposed in the canonical pose hypothesis component. For testing the effect of these parameters experiments have been conducted using Weizmann dataset because it is a very simple dataset without artefacts such as occlusion, dynamic background. The number of grids is always assumed to be 3 as they would correspond to head, torso and legs respectively. Thus this parameter is always assumed to be 3.

For cell configuration within each grid 3×2 configuration is used. The notion behind dividing grids into cells is to get finer details in the action representation. By dividing the grids into 2 halves, the whole body can be divided into either

Table 5.13: Experiments for various cell configurations in Weizmann dataset

Cell Configuration	Performance (%)	
	Proposed Framework (only PMF)	Proposed Framework (PMF + CPF)
3 x 2	92.47	100
4 x 2	100	100
5 x 2	97.78	100

up-down, left-right or forward-backward halves based on observed view and orientation of the person. In short, it will correspond to the medial axis of the body (between neck and hip). The only other parameter remaining is how many cells should be in each half of the grid. Experiments were conducted on Weizmann dataset with three cell configurations 3×2 , 4×2 , 5×2 to study the effect of this parameter. The results obtained with only PMF and PMF + CPF paths are tabulated in table 5.13. It can be seen that for all cell configurations 100% performance is achieved when both PMF and CPF paths are included. The main variation exists in performance of PMF path where 4×2 cell configuration performs best. The important concern when dividing into the cells is that the grids should not be over-divided, which will result in loss important information. Considering the grid sizes were as low as 7×15 , 3×2 cell configuration allows us to not over divide the grids and also obtain the same performance when both PMF and CPF paths are included.

In canonical pose hypothesis equation 5.1 of CPF path, a distance and orientation threshold ($\tau_d = 0.05$ and $\tau_\theta = 10$) are imposed to identify the best possible pose in the image. These thresholds are imposed to check if the body part's distance from neck and its orientation are similar to a canonical stick pose. If these thresholds are very high wrong canonical stick poses will be chosen. As this parameter only affects proposed framework with both paths, an experiment was conducted with 3 values each for $\tau_d = 0.05, 0.075, 0.1$ and $\tau_\theta = 10, 15, 20$. The results obtained

Table 5.14: Effect of varying τ_d and τ_θ in Canonical pose hypothesis using Weizmann dataset

τ_d	τ_θ	Performance of Proposed Framework (PMF + CPF)(%)
0.05	10	100
0.05	15	96.92
0.05	20	97.78
0.075	10	95.81
0.075	15	95.81
0.075	20	94.7
0.1	10	96.92
0.1	15	94.7
0.1	20	96.67

are tabulated in table 5.14 which shows that $\tau_d = 0.05$ and $\tau_\theta = 10$ give the best performance among all of them. The reason for reduction in performance could be wrong stick pose could have been identified for some cases which can cause erroneous output. The results also indicate that $\tau_d = 0.05$ shows best performance with any τ_θ . Similarly, $\tau_\theta = 10$ shows best performance for any τ_d .

Figure 5.9: Pose identified in CPF path with (a) Groundtruth Neck (b) Tolerance Range Neck (c) Outside Tolerance Range neck

Component Evaluation

The proposed motion feature's pose-invariance is mainly attributed to quantifying of body part's motion with respect to the body itself, that is, all motion features are measured with respect to a reference point (neck) and also converted to a human-body centric space. This allows better representation of an action under pose variations. Thus for successful feature extraction, neck and major viewing direction are required and assumed to be available. As seen in some video clips of Hollywood dataset, the neck is not at all visible resulting in erroneous feature detection. The neck point serves as the origin for canonical pose identification and representation. If a wrong neck point is chosen, the pose selected might be completely wrong or scaled improperly resulting in error. Also automatic detection or manual marking of neck reference point could be approximate or erroneous. Thus for proper functioning, the proposed framework needs to be robust to handle erroneous neck detection or marking.

To test the robustness of the proposed method to error in neck reference marking and its range of tolerance, an experiment was conducted with Weizmann, UCF sports and Hollywood datasets. Weizmann dataset is chosen as it is simple and foreground masks are available. Therefore any error in neck marking would directly affect the action recognition performance directly. To simulate erroneous neck markings, the point where neck reference could be located is varied in normalized stick coordinate system as shown in figure 5.10. The proposed framework was run with 3 types of neck points, namely, groundtruth neck, any point within tolerance range but outside groundtruth and outside tolerance range. From table 5.4, performance with groundtruth neck position can be seen as 100%. For case 2 neck markings, performance of framework decreases by a maximum of 3%. When performance of performance with only PMF path for both case 1 and case 2 neck

markings are compared the reduction is 0.6%. For case 3 neck markings, the decrease in performance is very high due to improper pose identification or grid division. Figure 5.9 shows poses obtained when groundtruth neck, tolerance range neck and outside tolerance range neck are used. It can be observed that pose identified for case 2 markings is still correct where as pose for case 3 is very poor. Due to this grid division and capturing motion of body parts are affected resulting in lesser action recognition performance. The tolerance range for neck marking was identified in normalized stick coordinate system to be between mid point of head (nose) to mid point of torso (chest).

Experiments were conducted with UCF sports and Hollywood dataset to obtain tolerance range for neck markings in the presence of pose variations and occlusions. For UCF sports, the proposed framework was run with groundtruth neck and any point within tolerance range but outside groundtruth (between nose and midpoint of chest). In UCF sports, the recognition accuracy when neck point within tolerance range is used dropped only by 5% compared to groundtruth neck. The performance drop in Weizmann and UCF sports is comparable and the proposed framework's robustness to neck marking even in presence of pose variations. This indicates that our neck reference assumption is not very stringent and obtaining a neck reference within identified tolerance range using automatic detection method would be realistic. Experiments were conducted in Hollywood dataset with 3 types of neck points, namely, groundtruth neck, point between groundtruth and nose, point between groundtruth and midpoint of chest. The identified tolerance range was divided into two because many videos in this dataset exhibit severe occlusion with only head being visible. The observed performance was different when compared to previous datasets. The performance in both cases (only PMF and PMF +CPF) increased by nearly 6% for case 2 markings with 60.17% for only PMF path and 63.98% when both paths are used. For case 3 markings, the

behaviour was different with performance reducing. The performance with only PMF path was 53.35% (a reduction of less than 1% compared to groundtruth). But performance when both paths are used reduced significantly by more than 14% to 42.18%. This reduction is mainly due to the severe occlusion present in this dataset. As mentioned in section 5.4.4, to handle the problem of occlusion, limits are imposed on the canonical stick poses to be compared with the frames. These limits are based on the neck point and size of head (for normalization). When neck is placed between groundtruth and mid point of chest, size of head increases significantly which will impose erroneous limits on the canonical sticks. Applying wrong limits on videos with severe occlusion can cause errors in pose estimation and due to this division of grids and motion in the visible grids are varied significantly. It could also indicate that the tolerance range for datasets with severe occlusion is much lesser than the easier datasets.

Viewing direction of the person is assumed to be available in the framework. But in real life scenarios such information might not be available. From a given video frame, few cues might be used to estimate the viewing direction of the person. For automatic view direction estimation for the proposed framework an experiment was conducted with Weizmann dataset. To estimate view, three main cues are considered namely, head pose, major motion occurring direction in the body and body contour shape. Head pose can be best indicator of viewing direction in cases of occlusion. But in low resolution videos (like Weizmann) where facial features cannot be captured it would be difficult to estimate head pose. Most available methods [231] for head pose estimation in low resolution employ the ratio of skin and non-skin pixels and is available only for upright postures. For this experiment based on available neck and estimated orientation state head grid is determined. The skin and hair pixels in the head are identified based on the notion that skin is brighter than hair. Based on the skin/hair ratio and orientation state a view

can be determined as either back, front, left and right. This is a very simple and coarse way to determine head pose and can be replaced by any other method. Deep learning based head pose estimation can be used for datasets such as UCF sports, Hollywood with better resolution.

Body shape can be good indicator of view direction in low resolution. But this might fail in cases of occlusion as complete body is not visible. For this experiment body contour shape feature [232] is extracted from training videos and an ELM classifier is trained. The classifier gives a likelihood value for each view and is normalized to $[0,1]$. Using the observed 2D motion, direction in major motion is occurring can be estimated as one of the 5 classes (up, down, left, right and no motion). For estimating the final view, a weighted scheme is used. In this scheme, body shape feature is given more weightage as it is a better indicator of view in low resolution and no occlusion. For a general framework this scheme has to be changed according to amount of occlusion, resolution etc.

The proposed view estimation method is able to estimate the correct view for 70.07% of the frames. The confusion still exists between *front-back* and *left-right*. The action recognition framework was run with the estimated view and the results are tabulated in table 5.15. With the estimated view, the proposed framework can still perform well. The performance drops only 1.1% in the PMF path whereas it drops by 4.5% in case of PMF + CPF path. Some of the errors from PMF path were persistent in the feedback path as well making it difficult to attain a performance closer to baseline. The cues used by the proposed view estimation algorithm are primitive in nature and very susceptible to errors. The proposed action recognition mainly fails when frontal view is wrongly classified as either left or right and vice-versa. For instance *side* action videos of Weizmann dataset contain people performing in front or back view but the major motion and body shape can be quite similar to left or right views. Even with erroneous view esti-

Table 5.15: Performance of proposed action recognition framework with proposed view estimation algorithm in Weizmann dataset

View Estimation	Performance of Proposed Framework (Only PMF)(%)	Performance of Proposed Framework (PMF + CPF)(%)
Baseline with assumed view	92.47	100
View estimated using proposed scheme	91.37	95.56

Figure 5.10: Groundtruth & Tolerance Range for Neck reference position

mation, the action recognition performance drops only by 4.5%. For better view estimation, head pose estimation must be done correctly with available head or face detectors [233–235] and deep learning [236, 237].

As shown in the experiments, CPF path is mainly dependent on canonical sticks extracted from training videos. The proposed framework's ability to handle pose variations is directly related to the canonical sticks available. But it is not always possible to have enough training videos that exhibit pose-variations. Since the availability of these stick poses directly determines the working of the framework, the generated stick poses are dependent upon pose variations exhibited in

the available training videos. For proper functioning of our framework, all pose variations must be encompassed by the canonical stick poses.

Figure 5.11: Erroneous stick pose identified by CPF path

Body part detector in Chapter 4 and canonical pose hypothesis scheme forms the basis of pose estimation in the proposed framework. Since the pose hypothesis scheme is dependent on the body part likelihood score, erroneous body part detection can lead to wrong pose identification. Figure 5.11 shows some examples of wrong pose identified for a frame. In unconstrained settings, body part detection is difficult due to several factors such as cluttered background, structural similarity in leg-arm, occlusion, huge intra-class variations etc. The aim of the pose hypothesis scheme is to estimate the poses correctly even when the body part detector is affected above mentioned factors. To check the effect of body part detection errors, a small subset of videos in Weizmann dataset were chosen (only *daria* videos). The body part detector was passed on all the motion consistent segments. The accuracy of each body part detection is measured using true positive, true negative, false positive and false negative. As with chapter 4, a detected body part location is considered to be correct, if more than 70% of the body part is correctly detected and less than 70% of the body part is missed. Canonical pose hypothesis scheme uses these body part detectors to estimate pose in each frame. The accuracy of the pose estimated is calculated in terms of each body

Table 5.16: Pose estimation error due to body part detection and canonical pose hypothesis in subset of Weizmann dataset

Body Part	Correct detections(%)	
	Body part detectors	After Canonical pose hypothesis
Head	28.65	97.39
Torso	56.34	94.42
Arm	57.87	85.69
Leg	64.64	90.71

part's location and orientation. The location constraint is same as the body part detector constraint mentioned above. The orientation of each body part has to be less than 10 degrees compared to groundtruth orientation. The correct pose estimation accuracy from body part detectors and canonical pose hypothesis scheme is tabulated in table 5.16. It was observed that pose estimation error using body part detectors is high due to large number of false positives. The availability of neck point and initial body part detections helps the pose hypothesis scheme to correctly identify the pose in the frame. From the table 5.16, it is clear that the pose hypothesis scheme reduces the pose estimation errors. For proper pose estimation, it is also necessary to have enough canonical stick poses for every action. If the available stick poses are lesser, pose estimation errors will increase. In the current setup, the action representation is dependent on poses and grid division in every frame. As seen in figure 5.11, grid division in most cases are still correct and computing histogram over the entire video reduces the errors caused by identifying the wrong stick pose. The results indicate that pose estimation errors are reduced considerably by pose hypothesis scheme and action recognition is also not affected by it. In unconstrained settings, improving body part detectors that can also handle non-upright postures and identifying proper pose for each frame will enhance performance of the proposed framework.

Another main reason for erroneous action recognition is foreground detection. In UCF sports, Hollywood and HMDB51 datasets, average background subtraction

does not detect foreground correctly. The manually marked bounding box is only used to reduce the space where average background subtraction is applied. Due to variations in lighting conditions and cluttered background, proper foreground detection is not possible resulting in erroneous features.

5.5 Concluding Remarks

In this chapter, to obtain knowledge of posture and improve performance of proposed PMF path, novel canonical pose feedback (CPF) path was proposed. The CPF path forms the second component of the proposed closed loop framework for action recognition. Using action recognized with the pose-invariant motion features from PMF path, a novel pose hypothesis scheme is proposed to find the best possible pose for a frame. The proposed scheme imposes distance and orientation constraints on a body part detector output to choose the most probable stick pose in the frame (stick poses for initial action recognized by PMF path is used). The identified pose is used to correct the motion features by realigning the grids made by PMF path and a new set of body part reference features (capture motion of each body part) and temporal stick features (capture temporal dynamics of action).

Thus, PMF and CPF paths combine both motion and pose cues in a mutually reinforcing manner. The proposed action recognition framework is similar to a closed loop framework in control system theories. The forward and feedback paths handle motion and pose cues respectively. Action determined by motion cues forms the input to the CPF path and pose determined by it is used to correct the motion features. The framework shows state of art results on available benchmark and NUAD datasets. Results on NUAD show framework's ability to handle non-upright postures in an action. To show the proposed framework's pose-invariance,

experiments were conducted with canonical sticks extracted from training videos. Results show that pose-invariance is attributed to two reasons, motion features extracted with respect to the body space and availability of canonical stick poses that cover possible pose variations. Results indicate that the proposed framework is pose-invariant, partially view-invariant and can handle occlusions. To study the effect of error in neck markings, an experiment was conducted with Weizmann dataset. The results indicate that an approximate neck marking is only required and a tolerance range for automatic neck detection or manual marking was identified.

Chapter 6

Conclusion and Recommendations for Future Work

6.1 Conclusion

Several challenges such as view angle changes, occlusion, anthropometric variations and pose variations make action recognition from videos difficult. Also there is a lack of research into non-upright posture actions. In this thesis, a pose-invariant action recognition framework is presented that can handle partial occlusions and is partially view-invariant. The proposed framework has novelty in pose-invariant motion feature extraction, posture estimation or identification and body part detection. Experiments have been conducted on both simple and challenging benchmark datasets to evaluate the effectiveness of the proposed framework. Results obtained validate the claim of our proposed method's robustness in challenging situations such as pose-variations, occlusion, view-angle change and camera motion.

A new dataset focusing on non-upright postures was introduced and results show the proposed method is general enough to handle these pose variations as well. Real-time recognition, proper foreground detection, availability and estimation of neck point, extraction of canonical stick poses to cover all possible pose-variations encountered are additional factors that affect the success of the framework.

In this final chapter, novel contributions made in this thesis, followed by some possible future improvements are summarized. Firstly contributions of this research work are summarized as follows:

1. **Novel action recognition framework combining motion and pose cues observed in a video -**

The proposed mutually reinforcing action recognition framework consists a propagation motion forward (PMF) path and canonical pose feedback(CPF) path to combine observed motion and pose cues. The initial action hypothesis from PMF path determined the canonical sticks to be used by CPF path. The pose identified by CPF is used to improve the motion based action representation. Both the paths combine to form a loop structure that is repeated until the action recognition result converges.

2. **Propagation Motion Forward (PMF) Path -**

To capture motion of different body part and handle pose variations, a novel propagation mechanism is proposed that utilises the natural proportionality of body parts. Kinematic motion features (*Div, Curl, Proj, Rot*) are computed for the foreground region and encoded in a pose-invariant manner into a human-body centric space. By computing features with respect to human body itself, our framework is able to handle all possible pose-variations.

3. Canonical Pose Feedback (CPF) Path -

CPF forms feedback or error path of the proposed framework that utilizes the output of forward path (action class) as feedback input. The main contributions in this component are summarized as

- a. An offline one-time algorithm to extract canonical stick poses for each action and a normalized representation for these poses.
- b. A pose hypothesis generation scheme to compare extracted canonical sticks of the action recognized by PMF with the video frame to identify the most likely canonical stick pose in it.
- c. The identified pose is used to realign the initial grids and include body part referenced kinematic features (motion of each body part) and temporal stick features (temporal dynamics) to the final action representation. Using these features, action hypothesis is refined until the action converges to yield the recognized action.

4. Likelihood score computation for Body part detection -

To allow body part detection in non-upright postures, a body part likelihood score based on Bayesian inference is proposed. The likelihood score is formulated with classifier output values as events and forms the basis of our CPF path. To test the robustness of this score, it is employed in a body part detection scheme and tested on visual data. A new dataset focusing on non-upright human posture action is introduced to test both action recognition framework and body part detection algorithm. The dataset comprises of 8 non-upright actions performed by 35 subjects and captured in 3 views.

6.2 Recommendations for Future Work

Although performance of proposed framework is good, several limitations could be identified. This would form the focus of future works.

6.2.1 Automatic Estimation of Viewing Direction and Reference Points (Neck)

An important aspect of the proposed framework is the assumption that neck and viewing points are available. For computation of kinematic features, the motion vector at neck point is essential. Also both human-body centric space and stick pose representation are constructed with neck as the origin. In the experiments mentioned above, neck point is manually marked or parts detected using person detection module [190] are used. Also [190] method works only on upright postures and fails to detect a person in many frames of KTH dataset. Our experiments have shown that neck reference assumption is not very stringent and obtaining a neck reference within experimentally identified tolerance range using automatic detection method would be realistic. Developing an automatic neck detection and tracking such that it is always within the tolerance range would be an important part of our future work.

Neck point is chosen as reference since it stable and can be seen easily from all views. But as shown in Hollywood dataset, many video clips do not show neck point at all. For these cases, it would be necessary to use different reference points. Our work is based on extracting motion features with respect to the body itself. For extending the proposed work, action can be represented as motion of body parts with respect to different reference points. But all chosen reference points might not be available in every video. To tackle this limitation, a different

matching criterion needs to be formulated such that features can be compared based on only available reference points. Also comparison of two videos when different set of reference points are available would also be important part of our future work. It would also be necessary to experimentally validate the chosen set of reference points and identify their tolerance range.

Person's viewing direction is necessary for conversion of kinematic features to human-body centric space. It is currently assumed to be available. Factors such as head pose or view, shape of body contour etc can be used to automatically estimate view of the person. A simple view estimation scheme was developed and tested in Weizmann dataset. Experiments conducted have shown that proposed action recognition framework is able to perform well even when view is wrongly estimated. A better approach for automatic view estimation should account for occlusion, resolution. In future, we plan to develop a multi-modal probabilistic approach that combines these factors to reliably estimate view direction of person.

6.2.2 Canonical Stick Pose Extraction and Availability

The set of canonical stick poses extracted for each action determine how well the proposed framework handle pose variations. To extract these canonical poses, available training videos are used by the proposed algorithm. This requires training videos must encompass all possible pose variations. But availability of limited training videos might disrupt functioning of the framework. For this purpose, it would be necessary to generate canonical stick poses that encompass all possible pose variations. For instance, only side view canonical sticks can be extracted from side view training videos. How to generate frontal canonical stick poses if frontal training videos are not available? As part of our future work, it would be necessary to devise a module that generates canonical poses not available from

those which are available. Another solution for this would be to convert available 2D sticks to 3D sticks. If such conversion is possible stick poses can be generated for all possible view and pose variations.

6.2.3 Stick Post Processing - Handling Body Part Detection Errors

Body part detection is an integral aspect of proposed framework. Detecting body parts are susceptible to errors due to huge intra-class variations, clothing, lighting variations, structural similarities between leg and arms. Due to this several body part detection errors were observed during experiments. One important reason for pose hypothesis generation scheme to identify the wrong pose for a frame is these detection errors. To tackle this it would be necessary to develop a post processing module that alters the stick pose identified. The neck point serves as a pivot point for these stick poses, which means currently any scaling or rotation is with respect to this point only. For proper pose identification, independent scaling and rotation of parts must be possible. Several image evidences such as color, appearance and motion features can be used to correct the canonical pose.

6.2.4 Deep learning approaches

In classification and machine learning domain, deep learning has gained significant research momentum in recent times. In the proposed action recognition framework, ELM is preferred due to its faster training time. Even though deep learning has been used for several image related work and not to handle temporal domain of videos. Currently such researches are being done to use deep learning for video related analysis. Some of the components in our framework are image based appli-

cations like foreground detection, body part detection etc which can be replaced by deep learning based approaches that in turn can improve the performance. For body part detection, we have experimented with DBN based deep learning approach. Even though DBN provided good performance for limb detection, the results for head and torso detection were still poor. This could be because deep learning requires a lot more training data. The proposed pose-invariant motion features are a good candidate for two stream deep learning techniques as they would provide information on which body part is moving and how they are moving with respect to other parts. Thus it can register appearance with motion allowing for better representational power. Improving the proposed framework with deep learning approaches will be important part of our future work.

6.2.5 Action Similarity Measures

For behavioural analysis, action is an important visible clue. To analyse and interpret behaviours, it would be necessary to do more than just recognize actions. One possible extension of the proposed work is to measure similarity between actions observed in a video. Such a similarity measure can be used to compare actions seen in videos. For example in physiotherapy exercises are integral part of patient recovery. These exercises need to be supervised by a therapist which is not always possible. In these cases we can compare a reference exercise video and patient's real time exercise video and measure similarity between actions. For each body part, similarity measure can be defined. Such modularity can be used to correct exercise actions of a patient. Apart from these, such measures can be used for healthcare, virtual coaching and virtual reality applications etc.

Author's Publications

Journal Papers:

1. Manoj Ramanathan, Wei-Yun Yau and Eam Khwang Teoh, ' Mutually Reinforcing Motion-Pose Framework for Pose Invariant Action Recognition', Submitted to Signal Processing: Image Communication, Springer, February 2017.
2. Manoj Ramanathan, Wei-Yun Yau and Eam Khwang Teoh, 'Human action recognition with video data: Research and evaluation challenges', Vol.44, Issue no. 5, IEEE Transactions on Human-Machine Systems, pg. 650 -663, October 2014.

Conference Papers:

1. Manoj Ramanathan, Wei-Yun Yau and Eam Khwang Teoh, ' Improving Human Body Part Detection using Deep Learning and Motion Consistency', IEEE Intl. Conf. on Control, Automation, Robotics and Vision (ICARCV), pg. 1 - 5, November 2016.
2. Manoj Ramanathan, Wei-Yun Yau and Eam Khwang Teoh, ' Human Posture Detection using H-ELM Body Part and Whole Person Detectors for Human-Robot Interaction', ACM Intl. Conf. on Human-Agent Interaction (HAI), pg. 239 - 242, October 2016.
3. Manoj Ramanathan, Wei-Yun Yau and Eam Khwang Teoh, 'Human Body Part Detection Using Likelihood Score Computations', IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM), pg. 160 - 166, December 2014, held as part of IEEE Symposia Series on

Computational Intelligence (IEEE SSCI) 2014.

Bibliography

- [1] Nikolaos Gkalelis, Hansung Kim, Adrian Hilton, Nikos Nikolaidis, and Ioannis Pitas. The i3DPost multi-view and 3D human action/interaction database. In *Conf. for Visual Media Production*, pages 159 – 168, November 2009.
- [2] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Action as space-time shapes. In *IEEE Intl. Conf. on Computer Vision*, volume 2, pages 1395–1402, October 2005.
- [3] Mikel D. Rodriguez, Javed Ahmed, and Mubarak Shah. Action MACH: A spatio-temporal maximum average correlation height filter for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2008.
- [4] Takanori Shibata. Therapeutic Seal robot biofeedback medical device: qualitative and quantitative evaluations of robot therapy in dementia care. *Proceedings of the IEEE, Invited Paper*, 100(8):2527–2538, August 2012.
- [5] Kimitoshi Yamazaki, Ryohei Ueda, Shunichi Nozawa, Mitsuharu Kojima, Kei Okada, Kiyoshi Matsumoto, Masaru Ishikawa, Isao Shimoyama, and Masayuki Inaba. Home-assistant robot for an aging society. *Proceedings of the IEEE, Invited Paper*, 100(8):2429–2441, August 2012.
- [6] Yanan Lu, Yakang Li, Yang Chen, Fang Ding, Xiaofeng Wang, Jicheng Hu, and Songtao Ding. A Human action recognition method based on Tchebichef moment invariants and temporal templates. In *Intl. Conf. on Intelligent Human-Machine Systems and Cybernetics*, pages 76–79, August 2012.
- [7] Fengjun Lv and Ramakant Nevatia. Single view human action recognition using key pose matching and Viterbi path searching. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.

-
- [8] Subhransu Maji, Lubomir Bourdev, and Jitendra Malik. Action recognition from a distributed representation of pose and appearance. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3177 – 3184, June 2011.
 - [9] Zhuolin Jiang, Zhe Lin, and Larry S. Davis. Recognizing human actions by learning and matching shape-motion prototype trees. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(3):533 – 547, March 2012.
 - [10] Daniel Weinland, Mustafa Ozuysal, and Pascal Fua. Making action recognition robust to occlusions and viewpoint changes. In *European Conf. On Computer Vision*, volume 6, pages 635 – 648, September 2010.
 - [11] Angela Yao, Juergen Gall, and Luc Van Gool. Coupled action recognition and pose estimation from multiple views. *International Journal of Computer Vision*, 100(1):16 – 37, October 2012.
 - [12] Philip Kelly, Aoife Healy, Kieran Moran, and Noel E. O’Connor. A virtual coaching environment for improving golf swing technique. In *ACM Multimedia Workshop on Surreal Media and Virtual Cloning*, pages 51 – 56, October 2010.
 - [13] Takeo Kanade and Martial Hebert. First-person vision. *Proc. of the IEEE, Invited Paper*, 100(8):2442–2453, August 2012.
 - [14] L. Palafox and H.Hashimoto. Human action recognition using wavelet signal analysis as an input in 4W1H. In *IEEE Intl. Conf. on Industrial Informatics*, pages 679 – 684, July 2010.
 - [15] Ronald Poppe. A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976 – 990, June 2010.
 - [16] Yibo Li and Yuehan Kuai. Action recognition based on spatio-temporal interest points. In *Intl. Conf. on BioMedical Engineering and Informatics*, pages 181 – 185, October 2012.
 - [17] M.S. Ryoo and J.K. Aggarwal. Human activity analysis: A Review. *ACM Computing Surveys, Article 16*, 43(3):16:1 – 16:43, April 2011.
 - [18] Xiaofei Ji and Honghai Liu. Advances in view-invariant human motion analysis - A Review. *IEEE Trans. on Systems, Man, and Cybernetic – Part C: Applications and Reviews*, 40(1):13 – 24, January 2012.

-
- [19] Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(3):257 – 267, March 2001.
 - [20] Trevor Darell and Alex Pentland. Space-time gestures. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 335 – 340, 1993.
 - [21] Gregory Rogez, J.J. Guerrero, and Carlos Orrite. View-invariant human feature extraction for video-surveillance applications. In *IEEE Conf. on Advanced Video and signal based Surveillance*, pages 324 – 329, 2007.
 - [22] Hossein Ragheb, Sergio Velastin, Paolo Remagnino, and Tim Ellis. Human action recognition using robust power spectrum features. In *IEEE Conf. of Image Processing*, pages 753–756, October 2008.
 - [23] Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas. Neural representation and learning for multi-view human action recognition. In *IEEE World Congress on Computational Intelligence*, pages 1–6, June 2012.
 - [24] Kaiqi Huang, Dacheng Tao, Yuan Yuan, Xuelong Li, and Tieniu Tan. View-independent behavior analysis. *IEEE Trans. on Systems, Man and Cybernetics - Part B: Cybernetics*, 39(4):1028 – 1035, August 2009.
 - [25] Juergen Gall, Angela Yao, Nima Razavi, Luc Van Gool, and Victor Lempitsky. Hough forests for object detection, tracking and action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(11):2188 – 2202, November 2011.
 - [26] Bhaskar Chakraborty, Ognjen Rudovic, and Jordi Gonzalez. View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts. In *IEEE Intl. Conf. on Automatic Face & Gesture Recognition*, pages 1 – 6, September 2008.
 - [27] Vasu Parameswaran and Rama Chellappa. Quasi-invariants for human action representation and recognition. In *Intl. Conf. on Pattern Recognition*, volume 1, pages 307–310, August 2002.
 - [28] D.M. Gavrilla and L.S. Davis. Towards 3-D model-based tracking and recognition of human movement. *Intl. Workshop on Face and Gesture Recognition*, pages 272 – 277, 1995.

-
- [29] Yaser Yacoob and Michael J. Black. Parameterized modeling and recognition of activities. In *IEEE Intl. Conf. on Computer Vision*, pages 120 – 127, 1998.
 - [30] Patrick Peursum, Geoff West, and Svetha Venkatesh. Combining image regions and human activity for indirect object recognition in indoor wide-angle views. In *IEEE Intl. Conf. on Computer Vision*, volume 1, pages 82–89, October 2005.
 - [31] Abhijit S. Ogale, Alap Karapurkar, and Yiannis Aloimonos. View-invariant modeling and recognition of human actions using grammars. In *IEEE Intl. Conf. on Computer Vision*, volume 5, pages 115 – 126, 2005.
 - [32] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W^4 : Real-time surveillance of people and their activities. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 22(8):809 – 830, August 2000.
 - [33] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision Image Understanding*, 104:249 – 257, October 2006.
 - [34] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Automatic discovery of action taxonomies from multiple views. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 1639 – 1645, 2006.
 - [35] Yuping Shen and Hassan Foroosh. View-invariant action recognition using fundamental ratios. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1 – 6, 2008.
 - [36] Fang Zhang, Yunhong Wang, and Zhaoxiang Zhang. View-invariant action recognition in surveillance videos. In *Asian Conf. on Pattern Recognition*, pages 580–583, November 2011.
 - [37] Feiyue Huang and Guangyou Xu. Action recognition unrestricted by location and viewpoint variation. In *IEEE Intl. Conf. on Computer and Information Technology Workshops*, pages 433 – 438, July 2008.
 - [38] M. Nireesh Kumar and Dr.K. Madhavi. Improved discriminative model for view-invariant human action recognition. *Intl. Journal of Computer Science & Engineering Technology*, 4(9):1263 – 1270, September 2013.

-
- [39] Alper Yilmaz and Mubarak Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *IEEE Intl. Conf. on Computer Vision*, volume 1, pages 150–157, 2005b.
 - [40] Yaser Sheikh, Mumtaz Sheikh, and Mubarak Shah. Exploring the space of human action. In *IEEE Intl. Conf. on Computer Vision*, volume 1, pages 144–149, October 2005.
 - [41] Saad M. Khan and Mubarak Shah. Detecting group activities using rigidity of formation. In *ACM Intl. Conf. on Multimedia*, pages 403–406, 2005.
 - [42] Cen Rao and Mubarak Shah. View-invariance in action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages II–316 – II–322, 2001.
 - [43] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, June 2008.
 - [44] Tanveer Syeda-Mahmood and A. Vasilescu. Recognizing action events from multiple viewpoints. In *IEEE workshop on detection and recognition of events in videos*, pages 64 – 72, 2001.
 - [45] Imran N. Junejo, Emilie Dexter, Ivan Laptev, and Patrick Perez. View-independent action recognition from temporal self-similarities. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 33(1):172 – 185, January 2011.
 - [46] Anwaar-Ul-Haq, Iqbal Gondal, and Manzur Murshed. On temporal order invariance for view-invariant action recognition. *IEEE Trans. On Circuits and Systems for Video Technology*, 23(2):203 – 211, February 2013.
 - [47] Ying Wang, Kaiqi Huang, and Tieniu Tan. Human activity recognition based on $\mathfrak{R}$ transform. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1 – 8, June 2007.
 - [48] Tyzzkae Ang, Weechiat Tan, Chukiong Loo, and Waikit Wong. Wavelet MACH Filter for omnidirectional human activity recognition. *Intl. Journal of Innovative Computing, Information and Control*, 8(6):1 – 20, June 2012.

-
- [49] Richard Souvenir and Justin Babbs. Learning the viewpoint manifold for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1 – 7, June 2008. DOI: 10.1109/CVPR.2008.4587552.
 - [50] Konstantinos G. Derpanis, Mikhail Sizintsev, Kevin J. Cannons, and Richard P. Wildes. Action spotting and recognition based on spatiotemporal orientation analysis. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 35(3):527 – 540, March 2013.
 - [51] Li Liu, Ling Shao, and Peter Rockett. Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition. *Pattern Recognition 46, Elsevier*, pages 1810 – 1818, July 2013.
 - [52] Li Liu, Ling Shao, Xiantong Zhen, and Xuelong Li. Learning discriminative key poses for action recognition. *IEEE Trans. on Cybernetics, Issue: 99*, pages 1– 11, January 2013.
 - [53] Michalis Raptis and Steffano Soatto. Tracklet descriptors for action modeling and video analysis. In *11th European Conf. on Computer Vision*. Springer, September 2010.
 - [54] Rashid Minhas, Aryaz Baradarani, Sepideh Seifzadeh, and Q. M. Jonathan Wu. Human action recognition using extreme learning machine based on visual vocabularies. *Neurocomputing 73, Elsevier*, pages 1906 – 1917, March 2010.
 - [55] Andrew Gilbert, John Illingworth, and Richard Bowden. Action recognition using mined hierarchical compound features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(5):883 – 897, May 2011.
 - [56] Hao Jiang, Mark S. Drew, and Ze-Nian Li. Successive convex matching for action detection. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 1646–1653, 2006.
 - [57] Olivier Chomat and James L. Crowley. Probabilistic recognition of activity using local appearance. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 104 – 109, June 1999.
 - [58] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: A local SVM approach. In *Intl. Conf. on Pattern Recognition*, volume 3, pages 32–36, August 2004.

-
- [59] Piotr Dollar, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *Joint IEEE Intl. Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, October 2005.
- [60] Ivan Laptev and Tony Lindeberg. Space-time interest points. In *IEEE Intl. Conf. on Computer Vision*, volume 1, pages 432–439, October 2003.
- [61] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatio-temporal words. In *British Machine Vision Conf. - Preliminary version*, volume 3, pages 1249–1258, September 2006.
- [62] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatio-temporal words. *Intl. Journal on Computer Vision*, 79:299–318, September 2008.
- [63] Abhinav Gupta, Praveen Srinivasan, Jianbo Shi, and Larry S. Davis. Understanding videos, constructing plots. Learning a visually grounded storyline model from annotated videos. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2012 – 2019, 2009.
- [64] Tanaya Guha and Rabab Kreidieh Ward. Learning sparse representations for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(8):1576 – 1588, August 2012.
- [65] Xinxiao Wu, Dong Xu, Lixin Duan, Jiebo Luo, and Yunde Jia. Action recognition using multilevel features and latent structural SVM. *IEEE Trans. on Circuits and Systems for video technology*, 23(8):1422 – 1431, August 2013.
- [66] Jingen Liu, Jiebo Luo, and Mubarak Shah. Action recognition in unconstrained amateur videos. In *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 3549 – 3552, April 2009.
- [67] Kishore K Reddy, Jingen Liu, and Mubarak Shah. Incremental action recognition using feature-tree. In *IEEE Intl. Conf. on Computer Vision*, pages 1010 – 1017, September 2009.
- [68] Bongjin Jun, Inho Choi, and Daijin Kim. Local transform features and hybridization for accurate face and human detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(6):1423 – 1436, June 2013.

-
- [69] Patrick Peursum, Hung H. Bui, Svetha Venkatesh, and Geoff West. Human action segmentation via controlled use of missing data in HMMs. In *IEEE Intl. Conf. on Pattern Recognition*, volume 4, pages 440–445, August 2004.
- [70] Nam T. Nguyen, Dinh Q. Phung, Svetha Venkatesh, and Hung Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 955 – 960, 2005.
- [71] Sangho Park and J.K. Aggarwal. A hierarchical Bayesian network for event recognition of human actions and interactions. *Multimedia Systems*, 10:164 – 179, 2004.
- [72] Darnell J. Moore, Irfan A. Essa, and Monson H. Hayes. Exploiting human actions and object context for recognition tasks. In *IEEE Intl. Conf. on Computer Vision*, volume 1, pages 80 – 86, 1999.
- [73] Nuria M. Oliver, Barbara Rosario, and Alex P. Pentland. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):831 – 843, August 2000.
- [74] Shah Atiqur Rahman, S.Y.Cho, and M.K.H.Leung. Recognising human actions by analysing negative spaces. *IET Computer Vision*, 6(3):197 – 213, May 2012.
- [75] Michalis Raptis, Kamil Wnuk, and Stefano Soatto. Flexible dictionaries for action classification. In *Intl. workshop on Machine Learning for Vision-based motion analysis*, pages 1 – 12, October 2008.
- [76] Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *IEEE Intl. Conf. on Computer Vision*, pages 1365 – 1372, September 2009.
- [77] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2650 – 2657, October 2013.
- [78] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *IEEE Intl. Conf. on Computer Vision*, pages 1331 – 1338, November 2011.

-
- [79] Lubomir Bourdev, Subhransu Maji, Thomas Brox, , and Jitendra Malik. Detecting people using mutually consistent poselet activations. In *European conf. on Computer vision: Part VI*, pages 168–181. Springer-Verlag Berlin, Heidelberg 2010, September 2010.
- [80] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 17 – 24, June 2010.
- [81] BangPeng Yao and Li Fei-Fei. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(9):1691 – 1703, September 2012.
- [82] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden Markov model. In *IEEE Conf. on Computer Vision and Pattern Recognition*, page 379 – 385, 1992.
- [83] Thad Starner and Alex Pentland. Real-time American sign language recognition from video using hidden Markov model. In *Intl. Sym. on Computer Vision*, pages 265 – 270, 1995.
- [84] Pradeep Natarajan and Ramakant Nevatia. Coupled hidden semi-Markov models for activity recognition. *IEEE Workshop on Motion and Video Computing*, pages 10 – 17, 2007. DOI: 10.1109/WMVC.2007.12.
- [85] Yang Yang, Imran Saleemi, and Mubarak Shah. Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(7):1635 – 1648, July 2013.
- [86] Mohiuddin Ahmad and Seong-Whan Lee. Human action recognition using multi-view image sequences features. In *Intl. Conf. on Automatic Face and Gesture Recognition*, pages 523 – 528, April 2006.
- [87] Xiaofei Ji, Ce Wang, Yibo Li, and Qianqian Wu. Hidden Markov model-based human action recognition using mixed features. *Journal of Computational Information Systems*, 9(9):3659–3666, May 2013.
- [88] Aaron F. Bobick and Andrew D. Wilson. A state-based approach to the representation and recognition of gesture. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(12):1325–1337, December 1997.

-
- [89] Abhinav Gupta and Larry S. Davis. Objects in action: An approach for combining action understanding and object perception. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, June 2007.
- [90] Nazli Ikizler-Cinbis and Stan Scarloff. Web-based classifiers for human action recognition. *IEEE Trans. On Multimedia*, 14(4):1031 – 1045, August 2012.
- [91] Jianguo Zhang, Benjamin Yao, and Yongtian Wang. Auto learning temporal atomic actions for activity classification. *Pattern Recognition 46, Elsevier*, pages 1789 – 1798, July 2013.
- [92] Yongmian Zhang, Yifan Zhang, Eran Swears, Natalia Larios, Ziheng Wang, and Qiang Ji. Modeling temporal interaction with interval temporal Bayesian networks for complex activity recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(10):2468–2483, October 2013.
- [93] Gal Lavee, Michael Rudzsky, and Ehud Rivlin. Propagating certainty in Petri nets for activity recognition. *IEEE Trans. on Circuits and Systems for video technology*, 23(2):326 – 337, February 2013.
- [94] Guangfeng Lin, Yindi Fan, and Erhu Zhang. Human action recognition using latent-dynamic condition random fields. In *Intl. Conf. on Artificial Intelligence and Computational Intelligence*, volume 3, pages 147 – 151, November 2009.
- [95] Yang Wang and Greg Mori. Hidden part models for human action recognition: probabilistic versus max margin. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(7):1310 – 1323, July 2011.
- [96] Yale Song, Louis-Philippe Morency, and Randall Davis. Action recognition by hierarchical sequence summarization. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3562 – 3569, June 2013.
- [97] Jyh-Yeong Chang, Jian-Jye Shyu, and Chien-Wen Cho. Fuzzy rule inference based human activity recognition. In *Intl. Sym. on Part of IEEE Multi-conf. on Systems and Control Intelligent Control*, pages 211 – 215, July 2009.
- [98] K. Subramanian and S.Suresh. Human action recognition using meta-cognitive neuro-fuzzy inference system. In *Intl. Joint Conf. on Neural Networks*, pages 1 – 8, June 2012.

-
- [99] Wei Huang and Q.M. Jonathan Wu. Human action recognition based on self organizing map. In *IEEE Intl. Conf. on Acoustics Speech and Signal Processing*, pages 2130 – 2133, March 2010.
 - [100] Wei Huang and Jonathan Wu. Human action recognition using recursive self organizing map and longest common subsequence matching. In *Workshop on Applications of Computer Vision*, pages 1 – 6, December 2009.
 - [101] Honghai Liu. Exploring human hand capabilities into embedded multifingered object manipulation. *IEEE Trans. on Industrial Informatics*, 7(3):389 – 398, August 2011.
 - [102] Zhaojie Ju and Honghai Liu. A Unified Fuzzy framework for human-hand motion recognition. *IEEE Trans. on Fuzzy Systems*, 19(5):901 – 913, October 2011.
 - [103] M.S. Ryoo and J.K. Aggarwal. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In *IEEE Intl. Conf. on Computer Vision*, pages 1593–1600, October 2009b.
 - [104] Claudio S. Pinhanez and Aaron F. Bobick. Human action detection using PNF propagation of temporal constraints. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 898 – 904, 1998.
 - [105] Antonio Oikonomopoulos, Ioannis Patras, and Maja Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Trans. on Systems, Man, and Cybernetics – Part B: Cybernetics*, 36(3):710–719, June 2006.
 - [106] Ashok Veeraraghavan, Rama Chellappa, and Amit K. Roy Chowdhury. The function space of an activity. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 959 – 968, 2006.
 - [107] Shah Atiqur Rahman, Liyuan Li, and M.K.H. Leung. Human action recognition by negative space analysis. In *Intl. Conf. on Cyberworlds*, pages 354 – 359, October 2010.
 - [108] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *IEEE Conf. on Computer Vision & Pattern Recognition*, pages 2929 – 2936, June 2009.

-
- [109] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos ‘in the wild’. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1996 – 2003, June 2009.
 - [110] Haoran Wang, Chunfeng Yuan, Guan Luo, Weiming Hu, and Changyin Sun. Action recognition using linear dynamic systems. *Pattern Recognition 46, Elsevier*, pages 1710 – 1718, June 2013.
 - [111] Di Wu and Ling Shao. Silhouette analysis-based action recognition via exploiting human poses. *IEEE Trans. on Circuits and Systems for video technology*, 23(2):236 – 243, February 2013.
 - [112] Wen Bian, Dacheng Tao, and Yong Rui. Cross-Domain human action recognition. *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):298 – 307, April 2012.
 - [113] Christian Thureau and Vaclav Hlavac. Pose primitive based human action recognition in videos or still images. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1 – 8, June 2008. DOI: 10.1109/CVPR.2008.4587721.
 - [114] Piotr Bilinski and Francois Bremond. Contextual statistics of space-time ordered features for human action recognition. In *IEEE Intl. Conf. on Advanced Video and Signal-Based Surveillance*, pages 228 – 233, September 2012.
 - [115] Anh-Phuong Ta, Christian Wolf, Guillaume Lavoue, Atilla Baskurt, and Jean-Michel Jolion. Pairwise features for human action recognition. In *Intl. Conf. on Pattern Recognition*, pages 3224 – 3227, August 2010.
 - [116] Yao-Hui Qin, Hong-Liang Li, Guang-Hui Liu, and Zheng-Ning Wang. Human action recognition using PEM histogram. In *Intl. Conf. on Computational Problem-Solving*, pages 323 – 325, December 2010.
 - [117] Hae Jong Seo and Peyman Milanfar. Action recognition from one example. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(5):867 – 882, May 2011.
 - [118] Lee W. Campbell and Aaron F. Bobick. Recognition of human body motion using phase space constraints. In *IEEE Intl. Conf. on Computer Vision*, pages 624–630, June 1995.

-
- [119] Anwaar-Ul-Haq, Iqbal Gondal, and Manzur Marshed. Action recognition using spatio-temporal distance classifier correlation filter. In *Intl. Conf. on Digital Image Computing: Techniques and Applications*, pages 474–479, 2011.
- [120] Yan Ke, Rahul Sukthankar, and Martial Hebert. Spatio-temporal shape and flow correlation for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, 2007.
- [121] Lihl Zelnik-Manor and Michal Irani. Event-based analysis of video. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages II–123 – II–130, 2001.
- [122] Roman Filipovych and Eraldo Ribeiro. Determining the scale of interest regions in videos. In *IEEE Intl. Conf. on Image Processing*, pages 985–988, November 2009.
- [123] Rita Cucchiara, Costantino Grana, Massimo Piccardi, and Andrea Prati. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(10):1337 – 1342, October 2003.
- [124] Sonal Kumari and Suman K. Mitra. Human action recognition using DFT. In *Third National Conf. on Computer Vision, Pattern Recognition, Image Processing and Graphics*, pages 239 – 242, 2011.
- [125] Konrad Schindler and Luc van Gool. Action Snippets: how many frames does human action recognition require? In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1 – 8, June 2008.
- [126] Lijuan He, Donghong Wang, and Hongyu Wang. Human abnormal action identification method in different scenarios. In *Intl. Conf. on Digital Manufacturing and Automation*, pages 594–597, August 2011.
- [127] Roberto Lublinerman, Necmiye Ozay, Dimitrios Zarpalas, and Octavia Camps. Activity recognition from silhouettes using linear systems and model (in)validation techniques. In *Intl. Conf. on Pattern Recognition*, pages 347–350, 2006.
- [128] Namrata Vaswani, Amit Roy Chowdhury, and Rama Chellappa. Activity recognition using the dynamics of the configuration of interacting objects. In

-
- IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 2, pages II-633 – II-640, June 2003.
- [129] Shahzad Cheema, Abdalrahman Eweiwi, Christian Thureau, and Christian Bauckhage. Action recognition by learning discriminative key poses. In *IEEE Intl. conf. on computer vision workshops*, pages 1302 – 1309, November 2011.
- [130] Jianxin Wu, Adebola Osuntogun, Tanzeem Choudhury, Matthai Philipose, and James M. Rehg. A scalable approach to activity recognition based on object use. In *IEEE Intl. Conf. on Computer vision*, pages 1 – 8, October 2007.
- [131] Alexis A. Efros, Alexander C. Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. *IEEE Intl. Conf. on Computer Vision*, 2:726 – 733, 2003.
- [132] Eli Shechtman and Michal Irani. Space-time behavior based correlation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 405–412, June 2005.
- [133] Shu-Fai Wong, Tae-Kyun Kim, and Roberto Cipolla. Learning motion categories using both semantic and structural information. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.
- [134] Jens Rittscher, Andrew Blake, Anthony Hoogs, and Gees Stein. Mathematical modelling of animate and intentional motion. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 358(1431):475 – 490, March 2003.
- [135] Shih-Yao Lin, Chuen-Kai Shie, Shen-Chi Chen, Ming-Sui Lee, and Yi-Ping Hung. Human action recognition using action trait code. In *Intl. Conf. on Pattern Recognition*, November 2012.
- [136] Liang Wang, Yizhou Wang, Tingtang Jiang, Debin Zhao, and Wen Gao. Learning discriminative features for fast frame-based action recognition. *Pattern Recognition* 46, Elsevier, pages 1832 – 1840, July 2013.
- [137] Kai Guo, Prakash Ishwar, and Janusz Konrad. Action recognition from video using feature covariance matrices. *IEEE Trans. on Image Processing*, 22(6):2479 – 2494, June 2013.

-
- [138] Yuanbo Chen, Zhixuan Li, Xin Guo, Yanyun Zhao, and Anni Cai. A spatio-temporal interest point detector based on vorticity for action recognition. In *IEEE Intl. Conf. on Multimedia and Expo Workshops*, pages 1 – 6, July 2013.
- [139] Hao Zhang and Lynne E. Parker. 4-Dimensional local spatio-temporal features for human activity recognition. In *IEEE Intl. Conf. on Intelligent Robots and Systems*, pages 2044 – 2049, September 2011.
- [140] Quoc V. Le, Will Y. Zou, Serena Y. Yeung, and Andrew Y. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3361 – 3368, June 2011.
- [141] Karinne Ramirez-Amaro, Eun-Sol Kim, Jiseob Kim, Byoung-Tak Zhang, Micheal Beetz, and Gordon Cheng. Enhancing human action recognition through spatio-temporal feature learning and semantic rules. In *IEEE-RAS Intl. Conf. Humanoid Robots*, October 2013.
- [142] Krystian Mikolajczyk and Hirofumi Uemura. Action recognition with motion-appearance vocabulary forest. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1 – 8, June 2008. DOI:10.1109/CVPR.2008.4587628.
- [143] Simone Calderara, Rita Cucchiara, and Andrea Prati. Action signature: a novel holistic representation for action recognition. In *IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, pages 121 – 128, September 2008.
- [144] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking - learning - detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(7):1409 – 1422, July 2012.
- [145] Zhaojie Ju and Honghai Liu. Fuzzy Gaussian mixture models. *Pattern Recognition*, 45(3):1146 – 1158, March 2012.
- [146] Li Wang, Li Cheng, Tuan Hue Thi, and Jian Zhang. Human action recognition from boosted pose estimation. In *Intl. Conf. on Digital Image Computing: Techniques and Applications*, pages 308 – 313, December 2010.
- [147] Yu Zhu and Guodong Guo. A study on visible to infrared action recognition. *IEEE Signal Processing Letters*, 20(9):897 – 900, September 2013.

-
- [148] Jian-Hua Wang, Jian-Jiun Ding, Yu Chen, and Hsin-Hui Chen. Real time accelerometer-based gait recognition using adaptive windowed wavelet transforms. In *IEEE Asia Pacific Conf. on Circuits and systems*, pages 591 – 594, December 2012.
- [149] Andrew Gilbert, John Illingworth, and Richard Bowden. Scale invariant action recognition using compound features mined from dense spatiotemporal corners. In *10th European Conf. on Computer Vision: Part I*, pages 222 – 233, 2008.
- [150] Henri Bouma, Gertjan Burghouts, Leo de Penning, Patrick Hanckmann, Johan-Martijn ten Hove, Sanne Korzec, Maarten Kruithof, Sander Landsmeer, Coen van Leeuwen, Sebastiaan van den Broek, Arvid Halma, Richard den Hollander, and Klammer Schutte. Recognition and localization of relevant human behavior in videos. *Proceedings of the SPIE*, 8711(10):87110B.1 – 87110B.10, April 2013.
- [151] Ferda Ofli, Rizwan Chaudhry, Gregorij Kurillo, Ren-Åte Vidal, and Ruzena Bajcsy. Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pages 8 – 13, June 2012.
- [152] Zhang Zhang and Dacheng Tao. Slow feature analysis for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(3):436 – 450, March 2012.
- [153] Bardia Yousefi, Chu Kiong Loo, and Ali Memariani. Biological inspired human action recognition. In *IEEE Workshop on Robotic Intelligence In Informationally Structured Space*, pages 58 – 65, April 2013.
- [154] Saad Ali, Arslan Basharat, and Mubarak Shah. Chaotic invariants for human action recognition. In *IEEE Intl. Conf. on Computer Vision*, pages 1 – 8, October 2007.
- [155] Chin-An Lin, Yen-Yu Lin, Hong-Yuan Mark Liao, and Shyh-Kang Jeng. Action recognition using instance-specific and class-consistent cues. In *IEEE Intl. Conf. on Image Processing*, pages 1373 – 1376, September 2012.
- [156] Masamichi Shimosaka, Takayuki Nishimura, Yu Nejigane, Taketoshi Mori, and Tomomasa Sato. Fast online action recognition with boosted combina-

- tional motion features. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pages 5851 – 5858, October 2006.
- [157] Viet Vo and Ngoc Ly. An effective approach for human actions recognition based on optical flow and edge features. In *Intl. Conf. on Control, Automation and Information Sciences*, pages 24 – 29, November 2012.
- [158] Thomas Dean, Rich Washington, and Greg Corrado. Sparse spatiotemporal coding for activity recognition. Department of Computer Science, Brown University, CS 10-02, March 2010.
- [159] Qiuxia Wu, Zhiyong Wang, Feiqi Deng, Zheru Chi, and David Dagan Feng. Realistic human action recognition with multimodal feature selection and fusion. *IEEE Trans. on Systems, Man, and Cybernetics: Systems*, 43(4):875 – 885, July 2013.
- [160] Zan Gao, An An Liu, Hua Zhang, Guang Ping Xu, and Yan Bing Xue. Human action recognition based on sparse representation induced by L1/L2 regulations. In *Intl. Conf. on Pattern Recognition*, pages 1868 – 1871, November 2012.
- [161] Honghai Liu, Shengyong Chen, and Naoyuki Kubota. Intelligent video systems and analytics: A Survey. *IEEE Trans. on Industrial Informatics*, 9(3):1222 – 1233, August 2013.
- [162] Xian Ming Lin and Shao Zi Li. Transfer AdaBoost learning for action recognition. In *IEEE Intl. Sym. on IT in Medicine & Education*, volume 1, pages 659 – 664, August 2009.
- [163] Ana Paula B. Lopes, Elerson R. da S. Santos, Eduardo A. do Valle Jr., Jussara M. de Almeida, and Arnaldo A. de Araujo. Transfer learning for human action recognition. In *SIBGRAPI Conf. on Graphics, Patterns and Images*, pages 352 – 359, August 2011.
- [164] Rong Liu, Ting Chen, and Lu Huang. Research on human activity recognition based on active learning. In *Ninth Intl. Conf. on Machine Learning and Cybernetics*, pages 285 – 290, July 2010.
- [165] Samuel Kaski and Jaakko Peltonen. Learning from relevant tasks only. In *European conf. on Machine Learning*, pages 608 – 615. Springer-Verlag Berlin, Heidelberg, 2007.

-
- [166] Tom Yeh, John Lee, and Trevor Darrell. Adaptive vocabulary forests by dynamic indexing and category learning. In *IEEE Intl. Conf. on Computer Vision*, pages 1 –8, October 2007.
 - [167] Jose Antonio Iglesias, Plamen Angelov, Agapito Ledezma, and Araceli Sanchez. Creating evolving user behavior profile automatically. *IEEE Trans. on Knowledge and data engineering*, 24(5):854–867, May 2012.
 - [168] Guang-Bin Huang, Lei Chen, and Chee-Kheong Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Trans. on Neural Networks*, 17(4):879 – 892, July 2006.
 - [169] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1653 – 1660, June 2014.
 - [170] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
 - [171] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: An astounding baseline for recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, pages 512 – 519, June 2014.
 - [172] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248 – 255, 2009.
 - [173] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1725 – 1732, June 2014.
 - [174] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2625 – 2634, June 2015.

-
- [175] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, volume 27, pages 568 – 587, November 2014.
 - [176] Bowen Zhang, Limin Wang, Zhe Wang, Yu Qiao, and Hanli Wang. Real-time action recognition with enhanced motion vector cnns. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2718 – 2726, June 2016.
 - [177] LiMin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4305 – 4314, June 2015.
 - [178] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conf. on Computer Vision*, volume 9912, pages 20 – 36, October 2016.
 - [179] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1933 – 1941, June 2016.
 - [180] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. Dynamic image networks for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3034 – 3042, June 2016.
 - [181] Md. Atiqur Rahman Ahad, J. Tan, H. Kim, and S. Ishikawa. Action dataset – A Survey. In *SICE Annual Conf. , Waseda University*, pages 1650 – 1655, September 2011.
 - [182] Sanchit Singh, Sergio A Velastin, and Hossein Ragheb. MuHAVi: A multi-camera human action video dataset for the evaluation of action recognition methods. In *IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, pages 48–55, August 2010.
 - [183] Hossein Ragheb, Sergio Velastin, Paolo Remagnino, and Tim Ellis. VIHASI: virtual human action silhouette data for the performance evaluation of silhouette based action recognition methods. In *ACM/IEEE Intl. Conf. on Distributed Smart cameras*, pages 1 – 10, September 2008.
 - [184] Hossein Ragheb, Sergio Velastin, Paolo Remagnino, and Tim Ellis. VIHASI: virtual human action silhouette data for the performance evaluation

- of silhouette based action recognition methods. In *ACM Intl. Workshop on Vision networks for behavior analysis (ACM Int'l Conf. on Multimedia)*, Vancouver, Canada, pages 77 – 84, October 2008.
- [185] S.Ramagiri, R.Kavi, and V.Kulathumani. Real-time multi-view human action recognition using a wireless camera network. In *ACM/IEEE Intl. Conf. on Distributed Smart Cameras*, pages 1–6, August 2011.
- [186] Chia-Chih Chen, M.S.Ryoo, and J.K.Aggarwal. *UT-Tower dataset: aerial view activity classification challenge*, August 2010. Available online: http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html.
- [187] Isabelle Guyon, Vassilis Athitsos, Pat Jangyodsuk, Ben Hamner, and Hugo Jair Escalante. Chalearn gesture challenge: design and first results. In *CVPR Workshops*, pages 1–6, June 2012.
- [188] Victoria Bloom, Dimitrios Makris, and Vasileios Argyriou. G3D: A Gaming action dataset and real time action recognition evaluation framework. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, pages 7 – 12, June 2012.
- [189] Zhe Lin, Zhuolin Jiang, and Larry S. Davis. Recognizing actions by shape-motion prototype trees. In *IEEE Intl. Conf. on Computer Vision*, pages 444 – 451, September 2009.
- [190] Iasonas Kokkinos. Rapid deformable object detection using dual tree branch and bound. In *Advances in Neural Information Processing Systems 24*, pages 2681 – 2689, December 2011.
- [191] Seyedehsamaneh Shojaeilangari, Wei-Yun Yau, Karthik Nandakumar, Jun Li, and Eam Khwang Teoh. Robust representation and recognition of facial emotions using extreme sparse learning. *IEEE Trans. on Image Processing*, 24(7):2140 – 2152, July 2015.
- [192] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2556 – 2563, November 2011.
- [193] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 886 – 893, June 2005.

-
- [194] Liyuan Li, Jerry Kah Eng Hoe, Shuicheng Yan, and Xinguo Yu. ML-Fusion based multi-model human detection and tracking for robust human-robot interfaces. In *Workshop on Applications of Computer Vision*, pages 1 – 8, December 2009.
- [195] Etienne Corvee and Francois Bremond. Body parts detection for people tracking using trees of histogram of oriented gradient descriptors. In *IEEE Intl. Conf. on Advanced Video and Signal Based Surveillance*, pages 469 – 475, August 2010.
- [196] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1815 – 1821, June 2012.
- [197] Xiaofeng Tong, Lei Wang, Tao Wang, and Yimin Zhang. Body detection and tracking with hierarchical scheme in dynamic scenes. In *IEEE Intl. Conf. on Multimedia and Expo*, pages 975 – 978, July 2007.
- [198] Bo Wu and Ram Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 951 – 958, June 2006.
- [199] Aras Dargazany and Mircea Nicolescu. Human body parts tracking using torso tracking: applications to activity recognition. In *Ninth Intl. Conf. on Information Technology: New Generations*, pages 646 – 651, April 2012.
- [200] Wentao Yao and Zhidong Deng. A robust pedestrian detection approach based on shapelet feature and Haar detector ensembles. *Tsinghua Science and Technology*, 17(1):40 – 50, February 2012.
- [201] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(4):349 – 361, April 2001.
- [202] Yijuan Lu and Hao Jiang. Human movement summarization and depiction from videos. In *IEEE Intl. Conf. on Multimedia and Expo*, pages 1 – 6, July 2013.

-
- [203] Katerina Fragkiadaki, Han Hu, and Jianbo Shi. Pose from flow and flow from pose. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2059 – 2066, June 2013.
- [204] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Tracking human pose by tracking symmetric parts. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3728 – 3735, June 2013.
- [205] Hao Jiang. Human pose estimation using consistent max covering. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(9):1911 – 1918, September 2011.
- [206] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55 – 79, January 2005.
- [207] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: people detection and articulated pose estimation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1014 – 1021, June 2009.
- [208] Silvia Zuffi, Oren Freifeld, and Michael J Black. From pictorial structures to deformable structures. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3546–3553, June 2012.
- [209] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 588 – 595, June 2013.
- [210] Deep networks: Overview http://ufldl.stanford.edu/wiki/index.php/Deep_Networks:_Overview.
- [211] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luoqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. Deep human parsing with active template regression. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015.
- [212] Shusen Zhou, Qingcai Chen, and Xiaolong Wang. Discriminative deep belief networks for image classification. In *Intl. Conf. on Image Processing*, volume 10, pages 1561 – 1564, September 2010.
- [213] Dong Liang, Kaijian Weng, Can Wang, Guoyuan Liang, Haoyao Chen, and Xinyu Wu. A 3d object recognition and pose estimation system using deep

- learning method. In *Intl. Conf. on Information Science and Technology*, volume 4, pages 401 – 404, April 2014.
- [214] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Intl. Conf. on Machine Learning*, pages 609 – 616, June 2009.
- [215] Thomas Brox, Andres Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conf. on Computer Vision*, volume 3024, pages 25 – 36, May 2004.
- [216] Leonid Sigal, A.Balan, and M.J.Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *Intl. Journal of Computer Vision*, 87(1-2):4 – 27, March 2010.
- [217] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [218] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.
- [219] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2442 – 2449, June 2009.
- [220] Saad Ali and Mubarakh Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(2):288 – 303, February 2010.
- [221] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *Human Behavior Understanding, Second International Workshop, Lecture Notes in Computer Science*, volume 7065. Springer Berlin Heidelberg, November 2011.

-
- [222] Binlong Li, Mustafa Ayazoglu, Teresa Mao, Octavia I. Camps, and Mario Sznajder. Activity recognition using dynamic subspace angles. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, June 2013.
- [223] Mehrtash T. Harandi, Conrad Sanderson, Sareh Shirazi, and Brian C. Lovell. Kernel analysis on grassmann manifolds for action recognition. *Pattern Recognition Letters*, 34(15):1906 – 1913, November 2013.
- [224] Michalis Vrigkas, Vasileios Karavasilis, Christophoros Nikou, and Ioannis A. Kakadiaris. Matching mixtures of curves for human action recognition. *Computer Vision and Image Understanding, Elsevier*, 119:27 – 40, February 2014.
- [225] LiMin Wang, Yu Qiao, and Xiaoou Tang. Motionlets mid-level 3d parts for human motion recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2013.
- [226] Chunyu Wang, Yizhou Wang, and Alan L. Yuille. An approach to pose-based action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 915 – 922, June 2013.
- [227] Heng Wang, Alexander Kldser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *Intl Journal on Computer Vision*, 103(1):60 – 79, May 2013.
- [228] Lin Sun, Kui Jia, Tsung-Han Chan, Yuqiang Fang, Gang Wang, and Shuicheng Yan. Df-sfa: Deeply-learned slow feature analysis for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2014.
- [229] Kaustubh Kulkarni, Georgios Evangelidis, Jan Cech, and Radu Horaud. Continuous action recognition based on sequence alignment. *International Journal of Computer Vision*, 112(1):90 – 114, March 2015.
- [230] Wangjiang Zhu, Jie Hu, Gang Sun, Xudong Cao, and Yu Qiao. A key volume mining deep framework for action recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1991 – 1999, June 2016.
- [231] N.M. Robertson and I.D. Reid. Estimating gaze direction from low-resolution faces in video. In *Proc. 9th IEEE Eur. Conf. on Computer Vision, Graz, Austria*, volume 3952, pages 402–415, May 2006.

-
- [232] Alexandros Andre Chaaraoui, Pau Climent-Perez, and Francisco Florez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters, Elsevier*, 34(15):1799 – 1807, November 2013.
- [233] Ming Zhang, Ke Li, and Yuncai Liu. Head pose estimation from low-resolution image with hough forest. In *Chinese Conf. on Pattern Recognition*, October 2010.
- [234] David van der Pol, Raymond H. Cuijpers, and James F. Juola. Head pose estimation for real-time low-resolution video. In Mark A. Neerincx and Willem-Paul Brinkman, editors, *ECCE*. Mediamatica, Delft University of Technology, The Netherlands / ACM, April 2010.
- [235] Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *IEEE Conf. on Biometrics Compendium Computer Vision and Pattern Recognition*, volume 6, June 2012.
- [236] Jiao Bao and Mao Ye. Head pose estimation based on robust convolutional neural network. *Cybernetics and Information Technologies*, 16(6):133 – 145, January 2017.
- [237] Xiabing Liu, Wei Liang, Yumeng Wang, Shuyang Li, and Mingtao Pei. 3d head pose estimation with convolutional neural network trained on synthetic images. In *Intl. Conf. on Image Processing*, pages 1289 – 1293, September 2016.