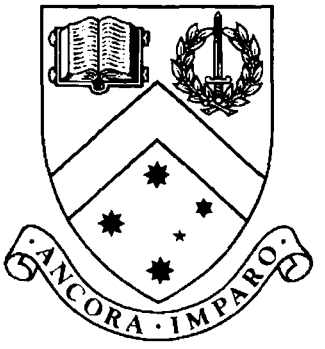
Visual Cues for View-invariant Human Action Recognition

by

Anwaar-ul-Haq, M.S. Computer System Engineering



Thesis

Submitted by Anwaar-ul-Haq

for fulfillment of the Requirements for the Degree of

Doctor of Philosophy

Gippsland School of Information Technology
Monash University

November, 2012

Copyright Notices

Notice 1

With extreme love to my creator, My LORD

Contents

List of Tables viii

List of Figures ix

List of Publications xiv

Abstract xv

Acknowledgments xvii

1 Introduction 1
- 1.1 Human Action Recognition: An Overview 4
  - 1.1.1 Action Representation 5
  - 1.1.2 Action Classification 6
  - 1.1.3 Application Areas 7
  - 1.1.4 Action Recognition Research Datasets 8
- 1.2 Challenges and Objectives 10
- 1.3 Motivations 10
- 1.4 Contributions 11
  - 1.4.1 Temporal Order Invariance for view-invariant action recognition (Chapter 3) 12
  - 1.4.2 Tracking-free and Training-free solutions for View-invariant Action Recognition (Chapter 4) 13
  - 1.4.3 Fast Frequency-domain View-invariant Action Recognition (Chapter 5) 14
  - 1.4.4 Contextual Action Recognition in Nighttime Videos (Chapter 6) . . 15
  - 1.4.5 Contextual Enhancement of Nighttime Videos (Chapter 6) 15
- 1.5 Organization of the Thesis 17
- 1.6 Conclusions 18

2 Visual Recognition of Human Actions: A Survey 19
- 2.1 Feature based Approaches 20
- 2.2 Geometry based Approaches 22
- 2.3 Template based Approaches 25
- 2.4 Context based Approaches 26
- 2.5 Conclusions 30

3 Action Analysis using Space-time Features 31
- 3.1 Spatio-temporal Ordered Packets (STOPs) 34
  - 3.1.1 Enclosed Features - Spatio Temporal Cuboids 34
  - 3.1.2 Envelope Features - Maximally Stable Volumes 36
- 3.2 Bag-of-Volumetric Packets 36

3.2.1 Volumetric Packet Matching . . . 36
3.2.2 Action Matching . . . 38
3.3 Experimental Results and Discussion . . . 38
3.3.1 Dataset and Experimental Setup . . . 38
3.3.2 Action Recognition Performance . . . 40
3.3.3 Robustness against noise and occlusions . . . 41
3.3.4 Computation Time . . . 42
3.4 Seeking Temporal Order Invariance for View-invariant Action Recognition . 43
3.5 The Proposed Approach . . . 45
3.5.1 Spatio-temporal Feature Fusion using Principal Component Analysis 46
3.5.2 Multiple View Feature Fusion Tables . . . 46
3.5.3 Action Classification . . . 47
3.6 Temporal Order Invariance: Experimentation . . . 48
3.6.1 Multi-view Action Datasets . . . 48
3.6.2 Experimental Setup . . . 48
3.6.3 Performance Comparison . . . 49
3.6.4 Importance of Geometrical Order Consistency . . . 51
3.6.5 Impact of Important parameters . . . 51
3.6.6 Limitations and Average Computation Time . . . 52
3.7 Conclusions . . . 52

4 Action Analysis using Epipolar Geometry . . . 53
4.1 Calculating Action Matching Score using Static Fundamental Matrix . . . 56
4.1.1 Action Representation . . . 56
4.1.2 Establishing Fundamental Matrix . . . 57
4.1.3 Derivation of Action Matching Score . . . 58
4.2 Calculating Action Matching Score using Multi-body Fundamental Matrix . 59
4.2.1 Establishing Two-body Fundamental Matrix . . . 59
4.2.2 Derivation of Action Matching Score . . . 60
4.3 Seeking Spatio-temporally Consistent Flow Correspondences . . . 61
4.3.1 Multi-frame Spatio-temporally Consistent Optical Flow : Four Frame Case . . . 61
4.3.2 Spatio-temporally Consistent Optical Flow based on Multi-frame Matching . . . 62
4.4 Robustness to Anthropometric Variations, Occlusion and Noise . . . 62
4.4.1 Dealing Anthropometric Variations . . . 64
4.4.2 Dealing Noise . . . 64
4.4.3 Dealing Occlusion . . . 64
4.4.4 Dealing Temporal Synchronization . . . 65
4.5 Experimental Results . . . 65
4.5.1 Datasets and experimental Set-up . . . 65
4.5.2 Action matching and retrieval . . . 67
4.5.3 Action recognition . . . 69
4.5.4 Robustness to noise and occlusion . . . 70
4.5.5 Comparison to other approaches . . . 72
4.5.6 Discussion and limitations . . . 72
4.6 Conclusions . . . 73

5 Action Analysis using 3D Frequency-Domain Filtering 74
5.1 The Action ST-DCCF filter 76
5.1.1 Action Classification 78
5.2 Action Representation 79
5.2.1 ST-DCCF for Vector Value data 79
5.3 VIEW-DCCF : View-invariant Space-time distance classifier correlation filtering 79
5.3.1 Filter Theory 80
5.3.2 View-invariant Action Classification 81
5.4 Experimental Results and Discussion 83
5.4.1 Dataset and Experimental Setup 83
5.4.2 Performance Comparison 85
5.4.3 Impact of Important parameters 86
5.5 VIEW DCCF Experimentation 87
5.5.1 Multi-view Action Datasets 87
5.5.2 Performance Comparison 88
5.5.3 Action Retrieval 89
5.5.4 Impact of Important parameters 90
5.5.5 Computational Time 91
5.6 Conclusions 91

6 Action Analysis using Contextual Associations 92
6.1 Context Enhancement 94
6.1.1 Through Video Fusion 94
6.1.2 Through Dynamic Contents Transfer 95
6.2 Contextual Action Recognition 95
6.2.1 Action Silhouette Processing, Information Fusion and matching . . . 96
6.2.2 Context Processing, Information Fusion and matching 98
6.2.3 Contextual Action Matching Score 99
6.3 Contextual enhancement using Video Fusion 99
6.4 System Architecture 101
6.5 The Proposed Video Fusion and Colorization Approach 102
6.5.1 Fusion and colorization in RGB color space 102
6.6 Objective Quality Evaluation 105
6.6.1 Color Similarity Measure (CSM) 106
6.6.2 Color Fusion Quality Index ($CFOI$) 106
6.6.3 Case1: Color image fusion with original color sensors 106
6.6.4 Case2: color image fusion without color sensors 109
6.7 Experimental Results and Discussion 110
6.7.1 Contextual Action recognition 110
6.7.2 Dataset and Experimental Setup 110
6.7.3 Action recognition 110
6.7.4 Automatic Contextual Action Annotation 111
6.7.5 Contextual Enhancement Using Color Transfer: Experimentation . . 112
6.7.6 Illustrations for Visual Inspection 112
6.7.7 Qualitative and Quantitative Comparison 113
6.7.8 Selection of Source Color Image 114
6.7.9 The significance of SCENT 114
6.7.10 An interesting Application: Contextual Action Recognition at Nighttime 115
6.7.11 CFOI: Experimentation 117
6.8 Conclusions 121

7 Conclusion and Future Work 122
7.1 A Recap of our Research Problem 122
7.2 The Significance and Impact of our Research Methodology 122
7.3 A Re-cap of our proposed Approaches 123
7.4 A Brief Review of our Proposed Approaches 123
7.5 Future Research Directions 126
7.6 Concluding Remarks 127

8 Appendix 128

List of Tables

3.1 The general structure of Feature Fusion Table T_k, $1 \leq k \leq K$. Each of these matrices have $i = 1, \ldots, v$ rows (viewpoints) and $j = 1, \ldots, m$ columns (number of features). 47
3.2 Performance comparison with the existing techniques. Average recognition is the average performance for all five cameras. 51
3.3 The effect of different values of γ on average recognition accuracy comparison for IXMAS dataset. 52

4.1 Performance comparison with the existing techniques. Average recognition is the average performance for all five cameras. 72

5.1 Average recognition accuracy comparison for KTH dataset with other state of the art techniques. 86
5.2 The effect of different values of β on average recognition accuracy comparison for KTH dataset. 86
5.3 Performance comparison with the existing techniques. Average recognition is the average performance for all five cameras. 88
5.4 The effect of different values of α on average recognition accuracy comparison for WVU and IXMAS datasets respectively. 90

6.1 Breakdown of CPU Time 115

8.1 List of Abbreviations for Chapter 3 130
8.2 List of Abbreviations for Chapter 4 130
8.3 List of Abbreviations for Chapter 5 131
8.4 List of Abbreviations for Chapter 6 131

List of Figures

1.1 Organizational Chart showing the flow of research work in this thesis. . . . 17

3.1 An illustration of same action (Kicking) from seven different cameras with different viewpoints by same actor. It shows how strongly viewpoints variations effect the description of an action. 32

3.2 Left to Right: Action volumes of v-cycling from You Tube data set, the respective maximally stable volume (MSV) and two different views of STOP features which encapsulate $spatio - temporal$ cuboids inside $MSVs$. . . . 33

3.3 The spatio-temporal ordering constraint and Indexing for volumetric packets. (Above) Consistent relative ordering between matched spatio-temporal features (Below) Inconsistent ordering between spatio-temporal features. . 34

3.4 Inverted file Index: The structure shows how geometrical order information is stored in this data structure. 35

3.5 Top: Action volumes of jack, wave2 and run from Weizmann dataset and their respective STOP features. Below: Action volumes of boxing, waving and clapping from KTH dataset and their STOP features. (Note that horizontal axis is X, vertical axis is Y and temporal axis is T. and lack of clarity is due to scale of volumes in low resolution.) 37

3.6 (Three Left Columns) Volumetric Packets. First column: Video sequences from Hollywood data set; Middle column: Maximally stable volume corresponding to the sequence; Third column: Volumetric packet with local features encapsulated within the volume.) 40

3.7 (Top Left)Confusion Matrices for Weizmann dataset without using geometric constraint (accuracy 94.44%). (Top Right) Confusion Matrices for KTH data set without using geometric constraint (accuracy 92.16%). (Below) Confusion matrices for YouTube dataset with mean accuracy of 65.6% without geometric constraint. 41

3.8 ((Top Left) Confusion Matrices for Weizmann dataset with geometric constraint (accuracy 100%). (Top Right) Confusion Matrices for KTH data set with geometric constraint (accuracy 95.3%).(Below) Confusion matrices for YouTube dataset with mean accuracy of 65.6%. 42

3.9 Average class precision for 8 actions in Hollywood data set with mean average precision as 36.06% and 40.4% for without and with geometric constraint, respectively. X-axis shows actions 1-8, which include, AnswerPhone, getoutofCar, HandShake, HugPerson, Kiss, SitDown, SitdUp and StandUp actions and Y-axis is the precision for recognition. 43

3.10 Temporal Flow within an Activity: Pick-up, Stand-up and Throw actions from three different cameras with different viewpoints. (Sample frames from IXMAS dataset, only three views are shown). It shows that temporal order of actions remains same with the whole activity irrespective of viewpoint variations. 44

3.11 The Proposed Framework: (above) Training is performed for all available viewpoints for getting fusion tables for each action class by repeating described steps for each action class in the dataset (sample instance of scratch-head is shown), and (below) Testing sequence for unknown query action video from an arbitrary viewpoint. 45
3.12 Confusion matrix for WVU dataset which shows average recognition accuracy of all viewpoints. 49
3.13 The Confusion Matrix for IXMAS Action dataset (with geometric consistency) with recognition accuracy (83.51%). 49
3.14 Recognition performance of IXMAS dataset from five different cameras with different viewpoints with geometric consistency ON. 50
3.15 Actions instances of IXMAS dataset from five different cameras with different viewpoints with geometric consistency OFF. 50
3.16 Action matching score for individual Action (Throwing) in WVU dataset. Fusion table from (T1—T11) are trained for seven different views and stand for respective action classes, nodding head, clapping,waving 1 hand,waving 2 hands, punching, jogging, jumping-jack, kicking, picking, throwing and bowling. It shows that score is maximum for the respective query class, a throwing action for different view. 51

4.1 Traditional trajectory based action representations which show landmark detection on actor body and tracking of landmark points. (a) walking action tracking, (b) complex trajectories for an activity and (c) trajectories within an action volume. 54
4.2 Flow diagram of general framework of our approach, AVITAR (achieving view-invariant tracking-free action recognition). It shows how two video sequences are processed to calculate action matching score is calculated from a series of steps. First, based on option (AVITAR1—AVITAR2), silhouettes are extracted or multi-frame feature matches are calculated and corresponding optical flow is used for calculation of flow correspondences then score is calculated based on rank of corresponding observation matrix. 55
4.3 Static epipolar geometry with two fixed cameras. A denotes 3D point (point on actor body), A_r and A_l denote projections on right and left image planes. As cameras are static, fundamental matrix should be satisfied between respective frames if sufficient correct correspondences are available. 57
4.4 Two views of two independent objects in each image, one static (belonging to background) and other dynamic (belonging to actor) and two-body epipolar geometry is explored in this scenario. 59
4.5 Spatio-temporally Consistent Optical Flow Multi-frame Setup: It utilizes four frames, two consecutive frames from each stereo video sequence. It is calculated between two stereo frames but temporal consistency constraint is used obtain only temporally consistent flow values. Fig. shows the original setup but as AVITAR 1 uses intensity based silhouettes, it utilizes silhouettes. 63
4.6 Effect of anthropometric variations. Epipolar geometry is fitted between points on two subjects with physical differences in structure but with similar pose. In this experiment, we click points on body left subject and respective epipolar line is automatically drawn on right subject passing through respective body point. It validates posture constraint. 65
4.7 The video matching results against the query action sequence for WVU multi-view dataset 66

4.8 Video Retrieval results for walking action of Alba action against different view long video sequence which contains 1200 frames. 67
4.9 Confusion matrix for IXMAS dataset against AVITAR1 and AVITAR2 . . . 68
4.10 Confusion matrix for WVU dataset against AVITAR1 and AVITAR2 69
4.11 Effect of occlusion on recognition accuracy of two datasets used: IXMAS and WVU using AVITAR 1 and AVITAR 2 approaches 70
4.12 Effect of noise on recognition accuracy of two datasets used: IXMAS and WVU using AVITAR 1 and AVITAR 2 approaches 71
4.13 Performance for five views of IXMAS dataset against static fundamental matrix based metric. 71
4.14 Performance for five views of IXMAS dataset against two-body fundamental matrix based metric. 72

5.1 Two representative action classes, (Lifting, Walking) show strong intra-class similarity and inter-class discrimination which should be encapsulated by a discriminative filter. 75
5.2 The schematic diagram of Action ST-DCCF filtering showing Transformation H which increases inter-class distance while simultaneously making each class more compact. It shows that after the transformation, distance d1 is the smallest making test action closest to walking class. 77
5.3 The simplest case of 2-class ST-DCCF filter. (Above (a),(b)) sample action volumes of wave and bend action (Bottom) A synthesized ST-DCCF transformation for two classes. (c) is visually ambiguous due to encapsulation of many training samples. 78
5.4 The schematic diagram of VIEW-DCCF filtering for single viewpoint showing Transformation H which increases inter-class distance while simultaneously making each class more compact. It shows that after the transformation, distance d1 is the smallest making test action closest to walking class. Similar transformation are required for each view cluster. 80
5.5 The detected action classes in KTH action dataset which include 6 action categories. This dataset is quite well known as a benchmark. 84
5.6 The detected action classes in UCF-Sports Action dataset which contain collection of broadcast sports action videos of 10 action classes including running,walking, diving, kicking, high-bar,lifting, skating, swinging, horse riding and golf actions. 84
5.7 The Confusion Matrix for KTH Action dataset with recognition accuracy (93.16%) for actions, 1-boxing, 2-clapping, 3-waving, 4-jogging, 5-walking and 6-running 85
5.8 The Confusion Matrix for UCF Sports Action Dataset (accuracy 77.66%) for actions, 1-diving, 2-golf, 3-kick, 4-lifting, 5-riding, 6-running, 7-skating, 8-swinging, 9-walking and 10-pole-vaulting. 86
5.9 An illustration of Different actions of IXMAS dataset by same actor. These actions include 1-check-watch, 2- cross-arms, 3- scratch-head, 4-sit-down, 5-get-up, 6-turn-around, 7-walk, 8-wave, 9-punch, 10-kick and 11-pick-up. . 87
5.10 The Confusion matrix for WVU dataset which shows average recognition accuracy of all viewpoints (89.8%). 88
5.11 The Confusion Matrix for IXMAS Action dataset with recognition accuracy (82.9%). 89
5.12 Recognition performance of IXMAS dataset from five different cameras with different viewpoints 89

5.13 Action matching score for individual Action (Punch) in WVU dataset. View-DCCFs are trained for seven different views, each containing respective action classes, nodding head, clapping,waving 1 hand,waving 2 hands, punching, jogging, jumping-jack, kicking, picking, throwing and bowling. It shows that score is maximum for the respective query class, a punch action for different view not used in training phase. 90

6.1 A nighttime scenario of *waving* action captured by low light visible and infra-red sensors which presents visual information of complementary nature and lack certain visual information on individual basis. 93

6.2 An illustration of video fusion: (a) An infra-red video stream (b) A registered video stream from low light visible and (c) A fused video sequence from (a) and (b). 94

6.3 Color Transfer based Video Fusion: (a) An infra-red video stream (b)Registered video stream in low light visible spectrum (c) a source color image for color transfer and (d) a color fused video stream which contains structural fusion from (a) and (b) and color transfer from (c). 95

6.4 Dynamic Content Transfer: (a) A static scene video sequence (a stack of static scene frames), (b) An infra-red video stream with dynamic scene of actor performing an action and (c) Video sequence (a) after transferring motion contents from (b). 96

6.5 The flow diagram of our contextual action recognition system for context enhanced multi-sensor videos which includes background subtraction, information fusion, action and context similarity estimation. However, it shows offline training for only one action class and every action class needs to be trained separately in a similar manner. 96

6.6 A fused action silhouette volume for wave1 action class based on OR-based-fusion rule applied at every frame instance from all training examples related wave1 action class. 97

6.7 An example of SIFT context image. SIFT descriptors are computed on a regular dense grid (for each pixel in an image). This visualization is obtained by mapping the first three principal components of each descriptor into the principal components of the RGB color space 98

6.9 Flow chart of the color morphing based video fusion and colorization system: SCENT (system for color exploitation at nighttime). 101

6.10 RGB color channel Fusion: (a) infrared input, (b) low light visible input ,(c) pseudo-fused color output 103

6.12 Illustration of global scene information encapsulated by respective GIST features. 105

6.13 Color Image Fusion with Color Sensors 107

6.14 Color Image Fusion by Color Transfer 109

6.15 The Confusion Matrix without contextual cues (84.87%) for actions, 1-walking, 2-wave1, 3-wave2, 4-stand-up, 5-sit-down, 6-hands-up, 7-clapping and 8-pick-up 111

6.16 The Confusion Matrix with contextual cues (91.75%) for actions, 1-walking, 2-wave1, 3-wave2, 4-stand-up, 5-sit-down, 6-hands-up, 7-clapping and 8-pick-up 111

6.17 An illustration of automatic contextual action annotation of multi-sensor video data in which action and its contextual scene is rightly recognized. . . 112

6.18 Four frames of video sequence and results (Scene A). (Above) grayscale frames from infra-red video sequence, (Middle) grayscale frames from low light visible video sequence and (Below) fused and colorized frames as result of SCENT 113
6.19 Four frames of video sequence and results (Scene B). (Above) grayscale frames from infra-red video sequence, (Middle) grayscale frames from low light visible video sequence and (Below) fused and colorized frames as result of SCENT. 113
6.21 CSM comparison of our proposed system SCENT with other competitive techniques.The values of objective quality measure are in the range [0-1]. The large value are indication of better quality 114
6.22 CFOI comparison of our proposed system SCENT with other competitive techniques. The values of objective quality measure are in the range [0-1]. The large value are indication of better quality 115
6.23 Color source Image Selection showing query image , Precision-recall curve and selected images from color image collection by (c) proposed contextual association (d) structural association [146]. False positives are shown in red bounding boxes 116
6.26 (a) 3D MSV for moving actor, (b) spatio-temporal cuboids encapsulated within 3D MSV 116
6.24 Snapshot of GUI (graphical user interface) developed for SCENT 117
6.25 Scene Description for Contextual Action Recognition at Nighttime , (a) IR frame, (b) low light image, (c)color source (d) colorized frame 118
6.27 Color Image Fusion results with Color Distortions 118
6.28 CFOI comparison for color distortions 119
6.29 Original images (a,b), color fused image (c) and blurred images (d,e,f) . . . 119
6.30 CFOI comparison for edge burring effect 120
6.31 Original images (a,b), color fused with averaging (c), Laplacian (d) and wavelet (e) 120
6.32 CFOI comparisons for image visual information integration 121

List of Publications

1. Anwaar-ul-haq, I. Gondal and M. Murshed, *On Temporal Order Invariance for View-invariant Action Matching*, IEEE Transactions on Circuits and Systems for Video Technology, Vol (22), DOI 10.1109.TCSVT.2012.2203213, 2012, ERA RANK A, Impact Factor 2.55.

2. Anwaar-ul-haq, I. Gondal and M. Murshed, *On Dynamic Scene Geometry for View-invariant Action Matching*, *In Proc.* CVPR, Colorado Springs, USA. 2011, ERA RANK A.

3. Anwaar-ul-haq, I. Gondal and M. Murshed, *Contextual Action Recognition in Night-time Video Sequences*, *In Proc.* DICTA, Noosa Resort, AUS, 2011.

4. Anwaar-ul-haq, I. Gondal and M. Murshed, *Action Recognition using Spatio-temporal Distance Classifier Correlation Filters*, *In Proc.* DICTA, Noosa Resort, AUS, 2011, ERA RANK B.

5. Anwaar-ul-haq, I. Gondal and M. Murshed, *Automated Multi-sensor Color Video Fusion for Night-time Video Surveillance*, *In Proc.* IEEE ISCC, Riccione, Italy, 2010, ERA RANK B.

6. Anwaar-ul-haq, I. Gondal and M. Murshed, *A Novel Color Image Fusion QoS measure for Multi-sensor Night Vision Applications*, *In Proc.* IEEE ISCC, Riccione, Italy, 2010, ERA RANK B.

7. Anwaar-ul-haq, I. Gondal and M. Murshed, *SCARF: Semi-automatic Colorization and Reliable Image Fusion*, *In Proc.* DICTA, Sydney, 2010, ERA RANK B.

8. Anwaar-ul-haq, I. Gondal and M. Murshed, *SCENT: System for Color Exploitation at Nighttime*, *In Proc.* ARCHER, Melbourne, 2010 (CSIRO OVERALL BEST PAPER AWARD).

9. Anwaar-ul-haq, I. Gondal and M. Murshed, *AVITAR: Achieving View-invariant Tracking-free Action Recognition*, *(In review)* IEEE Transaction on image Processing, 2012.

10. Anwaar-ul-haq, I. Gondal and M. Murshed, *VIEWDCCF: A Spatio-temporal Frequency-domain Filter for Matching Actions Across Viewpoint Variations*, *(to appear)* IEEE Transactions on Circuits and Systems for Video Technology, 2012.

Abstract

Human action is a visually complex phenomenon. Visual representation, analysis and recognition of human actions has become a key focus of research in computer vision, artificial intelligence, robotics and other related scientific disciplines. Various applications of automated action recognition include but not limited to intelligent health care monitoring, smart-homes, content based video search, animation and entertainment, human-computer interaction and intelligent video surveillance. The main focus of all these application areas surrounds a fundamental question: Given a human subject doing something in the field of sensory input, what is the person doing? If machine is able to correctly answer this question, it can greatly benefit computer vision system development and practical usage.

However, machine recognition of human action is a daunting task due to complex motion dynamics, anthropometric variations, occlusion and high dependency over camera viewpoint. In this thesis, we exploit the importance of rich visual cues from human actions and utilize them to propose valuable solutions to human action recognition. The important problem of view-invariance under viewpoint variations is taken as a case study. We collect and explore these visual cues from geometrical relationships, spatio-temporal patterns and features, frequency domain signal analysis, contextual associations of actions and derive action representations for machine recognition.

Actions are known as spatio-temporal patterns and temporal order plays an important role in their interpretations. We, therefore, explore invariance property of temporal order of actions during action execution and utilize it for devising a new view-invariant action recognition approach. We apply order constraint and feature fusion on local spatio-temporal features. These features are representation of choice for action recognition due to their computational simplicity, robustness to occlusion and minor view-point changes. We introduce STOPs (spatio-temporal ordered packets) that combine discriminative characteristics of multiple features for better recognition performance. In addition, we introduce spatio-temporal ordering constraint that removes discrepancy of orderless formation of bag-of-feature framework for action recognition.

Furthermore, to deal with limitations of feature based approaches, we explore multiple view geometry which has alleviated various complex problems in computer vision. We thoroughly study applications of static and multi-body flow fundamental matrix in context of relating across-view information. We introduce spatio-temporally consistent dense optical flow to avoid explicit manual human body landmark point detection and explicit point correspondences. We employ rank constraint to derive novel tracking and training-free action similarity measures across viewpoint variations.

Next, we investigate that despite the considerable success of geometrical techniques, computational complexity due to dense optical flow calculations plays a hindering role. Therefore, we study and track frequency domain analysis of action sequences. It leads toward the derivation of spatio-temporal correlation filters that use frequency domain filtering to give fast and efficient solutions to action recognition. However, these filters are originally view-dependent solutions. To achieve this objective, view clustering is explored that extends frequency domain techniques to achieve view-invariance.

Contextual information is another important cue for interpreting human actions especially when actions exhibit interactive relationships with their context. These contextual clues become even more crucial when videos are captured in unfavorable conditions like extreme low light nighttime scenarios. We, therefore, take case study of night vision and present contextual action recognition at nighttime. We discover that context enhancement is imperative in such challenging multi-sensor environment to achieve reliable action recognition which leads us to develop novel context enhancement techniques for night vision using multi-sensor image fusion.

Extensive experimentation on well-known action datasets is performed and results are compared with the existing action recognition approaches in literature. The research findings in this thesis greatly encourage the exploitation of spatio-temporal visual cues for deriving novel action recognition approaches and increasing their performance.

Acknowledgments

All majesty be to my Lord who made me belonging to Him and to the Apostle who gave me realization to whom I belong.

Every accomplishment is eventually based on acknowledgements. Emotional upheavals during this lifespan have remained so strongly connected to my research work that they imparted to me forbearance against challenges, commitment to the objective and vision for foresight. The fountain of this research work gushed forth with huge motivational encouragement and support of my beloved wife (late Saima Anwaar), channeled through grievous valley of her awful disease and untimely death and finally came to an end by giving me the vision of a redeemed life. A beautiful floret withered away without full blossom and when I needed it the most. This time was a cultivating landscape on which not every color was bright. However, it didn't make me colorblind but acquainted me with the artistry of colors. It has opened up a path of self-discovery, resolute anticipation of actions and transparent vision of objectivity. Thanks to Saima for her every contribution during her short stay of three years with me leaving everlasting impressions on my entire life.

I have always been a devotee of visionary and progressive thinking of my dear father who deserves all the credit for everything in my life. His disciplined approach caused me to face life and his love caused me to paint life. However, this bright coherent light substitutes merely a single band of broader and diversely colorful love spectrum of my dearest mother whose contribution is beyond acknowledgements. To complete this beautiful picture, both of my sisters (Noreen and Mubeen) are standing on my each side resembling my arms and hands, giving names to love, care and utmost support.

However, in the darkness of ignorance, I needed beckons of light and all my teachers from the one who taught me the first letter to the ones who supervised my PhD are my beckons of light. I am obliged to all of them in similitude of an eye to the visible light. Their contribution in my world is beyond words. They all are my heroes. I am specially thankful to my PhD supervisory committee members Dr. Iqbal Gondal, Dr. Manzur Murshed and Dr. Mubarak Shah for their valuable support, encouragement and supervision but I wish to highlight the immense support I received from my supervisors Dr. Iqbal Gondal, Director ICT and Dr. Manzur Murshed, Head of School, GSIT, Monash University. All credit goes to Dr. Iqbal Gondal who successfully steered me through my PhD through his academic expertise and especial kindness, in the most turbulent time of my life. I am greatly indebted to Dr. Iqbal Gondal and his family for their emotional support and encouragement.

The outstanding support of my friends, professional colleagues and relatives always remained a lifeline for me. I am specially grateful to Dr. Ashraf Kazi (BUSECO, Monash University) who greatly encouraged me in my troubled times. I am thankful to my colleagues and friends Ammar Haider, Farrukh Sheikh, Shahid Hussain, Muhammad Amar,

Farzaneh Armaghani, Xueliang Hua (Nevo), Wang Li, Zhouyu Fu (Joey) and Saad Ali (Sarnoff Corporation), USA. I feel great pleasure to thank my friends and house mates, Bianca Leach, Mellisa Mace, Justine Carroll and John Mace for a lovely and friendly time at home.

Finally, I would like to express my gratitude towards Prof. Ed Byrne, Vice Chancellor Monash University, Prof. Helen Bartlett, Pro-vice Chancellor and especially Prof. Philip Taylor, Director Research and Graduate Studies, Monash University, who proved pillars of support for me. Their personal efforts and encouragement helped me to continue my research in changed circumstances.

The love of my life, my little Asil is the continuation of my Lord's blessings on me. May Lord bestow on him His ultimate knowledge, wisdom and gnosis.

Anwaar-ul-Haq

Monash University
November 2012

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Anwaar-ul-Haq
November 22, 2012

Chapter 1

Introduction

> All human actions have one or more of these seven causes: chance, nature, compulsions, habit, reason, passion and desire.
>
> ~Aristotle (384 BC - 322 BC)

Actions are the true manifestation of human qualities, as human desires, hopes, beliefs and intentions eventually result in their actions. In his landmark treatise, Ludwig Von Mises [1] points to this purposive characteristic of human action in these words, "Human action is a purposeful behavior. It is the ego's meaningful response to stimuli and a persons conscious adjustment to the state of the universe that determines his life". The ability to perceive, interpret, understand and predict human actions is vital to the understanding of the nature of human life and it is the very subject of this thesis.

Since Aristotle's Nicomachean Ethics [2], action analysis has attracted strong attention of many philosophers. Hegel, Max Weber, Ludwig von Mises, August Cieszkowski, John Martin Fischer and Donald Davidson greatly contributed towards the philosophy of action [7]. The primary concerns of the philosophy of action are to analyze the nature of actions, individuating actions, explaining the relationship between actions and their effects, explaining how an action is related to the beliefs and desires, the role of the nature of free will and mental or physical states that cause the actions. The concerns of the action theorists overlap with those doing work in other areas of the philosophy of mind and metaphysics, moral philosophy, the philosophy of religion, logic, epistemology, legal philosophy, and with the recent growing interest in social action theory, social and political philosophy. Action theory is one of those unique areas in philosophy with a boundary that is difficult to fix.

According to the action theory (philosophy) [3], actions are defined as *the behaviors* caused by *the agents*(actors) in particular circumstances. These are processes that are caused by willful human bodily movements of more or complex nature. More rigorously, the actions represent bodily movements that are believed to be driven by the intentions. For instance, throwing a ball is an example of action; it involves an intention, a goal, and a bodily movement guided by the agent. On the other hand, feeling a headache is not considered as an action because it is something which happens to a person, not something done by someone. In other words, an agent doesn't intend to get pain or engage in bodily movement during an experience of headache that excludes it from being considered as an action.

Action Analysis in various Scientific Disciplines:

In psychology and sociology, human action is investigated in the name of behavior study and referred to the action and mannerism made by the human in conjunction with its environment. The behaviorist school of thought [4] maintains that behaviors can be described scientifically without recourse either to internal physiological events or to hypothetical constructs such as the mind. Modern-day behaviorism, known as "behavior analysis," is a thriving field. In sociology, behavior in general is considered as having no meanings, being not directed at other people, and thus is the most basic human action. To many theorists, the locus of interest lies in actors, actions, and interactions between actors. Collectively, they try to answer, "What is a human action ?".

The American sociologist Talcott Parsons [5] created a model of human action which stressed that the most basic interesting event to recognize is goal-directed action. It was further refined by his student Robert K. Merton. In this model [6], human actions are made up of: (i) The actor or agent performing an action (including their intentions, schemas, knowledge, motives, and identity); (ii) The goal, or a future state of affairs that is desired (which may be human communicative action or be an object-oriented action; and be either a creative goal or reaction to a dilemma); (iii) The situation in which action is located, including both: the conditions of action that include the normative background, the obstacles in the way of achieving the goal, and the human ecology of the setting, the means of action and the actual consequences of the action (which may be foreseeable or unforeseeable, and either intended or unintended).

Cognitive neuroscience studies action analysis and recognition functionalities by investigating the nervous system. An important contribution by neuroscience is the recent discovery [8] about mirror neurons which are believed to be fired when when an actor acts and when the actor observes the same action performed by another. The neurophysiologists placed electrodes in the ventral premotor cortex of the macaque monkey to study neurons specialized for the control of hand and mouth actions; for example, taking hold of an object and manipulating it. During each experiment, they recorded response from a single neuron in the monkey's brain while the monkey was allowed to reach for pieces of food, so the researchers could measure the neuron's response to certain movements. They found that some of the neurons they recorded would respond when the monkey saw a person pick up a piece of food as well as when the monkey picked up the food. Further experiments confirmed that about 10 percent of neurons in the monkey inferior frontal and inferior parietal cortex have 'mirror' properties and give similar responses to performed hand actions and observed actions. Brain imaging experiments using functional magnetic resonance imaging (fMRI) have shown that the human inferior frontal cortex and superior parietal lobe is active when the person performs an action and also when the person sees another individual performing an action. It has been suggested that these brain regions contain mirror neurons, and they have been defined as the human mirror neuron system. Human infant data using eye-tracking measures, suggests that the mirror neuron system develops before 12 months of age, and that this system may help human infants understand other people's actions.

The undisputed fact is that actions are essence of human existence and substance of great importance. Therefore, action analysis is a subject of primary importance, worth of scientific investigation and exploration. It is open for scientific enquiry with no defined boundaries. In other words, the analysis of human action is not restricted to some specific area of science, it is the subject of study in various scientific disciplines like neuroscience, cognitive science, agronomic, economics, psychology, praxeology and (computer vision) artificial intelligence (AI). This thesis investigates human actions and their recognition in the context of visual analysis, modeling, recognition and understanding by machine vision.

These aspects of action analysis research come under the scientific discipline of computer vision.

Computer Vision and Action Understanding:

Computer Vision is a branch of artificial intelligence (AI) that is intended to develop visual perception techniques for computers that are indispensable part of our modern life. It is the science and technology of the machines that see, and this capability is possible through analysis and interpretation of single image or sequence of images to determine if they contain some object, feature or activity of interest. In other words, computer vision is the the theory of artificial systems that take information from images with the objective of attaining awareness or understanding of sensory information.

The history of computer vision is bit accidental as it was mistakenly considered as simple Artificial Intelligence (AI) problem. A fine historic example is the MIT *copy demo* problem [9] that was given as assignment to students. The idea was to write a computer vision program to analyzes an image of a scene containing several stacked blocks, recover the structure of the blocks, and generate a code for a robot to build an exact copy of the block structure. None of the students was able to solve this problem. Later on researchers realized that it was actually a high level vision problem and technology had not yet solved low level vision problems.

Vision is one of the principal senses of the human being. The human vision acts as a lower bound on our ambitions with regard to computational image analysis. Giving the future robots, a similar vision to the human is a big challenge. It was known that the human brain processes visual information in semantic space mainly, that is, extracting the semantically meaningful features such as line-segments, boundaries, shape and so on. But by recent information processing techniques, these kinds of features cannot be detected by computers robustly so that in computer vision it's still difficult to process visual information as humans do. Computers have to process visual information in data space formed by the robustly detectable but less meaningful features such as colors, textures etc. Therefore, the processing methodology in computers is quite different from that in human beings. The trouble is that pixels have no meaning for humans. One must create from them other entities that capture properties of a picture that are meaningful to people and that is not an easy task. This difference between human perception of pictures and pixel statistics is called the semantic gap. However, extensive research on computer vision approaches in last three decades has greatly contributed towards its success and it has gradually made the transition away from understanding single images to analyzing video sequences, or video understanding.

The closely related research areas to computer vision are image processing, machine vision and pattern analysis. There is a significant overlap in the range of techniques and applications they cover. This implies that the basic techniques that are used and developed in these fields are more or less identical, but with subtle differences. Success in one field supports the other. Similarly, when we distinguish each of the fields from the others, the following characterizations appear relevant:

Image processing and image analysis tend to focus on 2D images, how to transform one image to another, e.g., by pixel-wise operations such as contrast enhancement, local operations such as edge extraction or noise removal, or geometrical transformations such as rotating the image. This characterization implies that image processing/analysis neither require assumptions nor produce interpretations about the image content. Whereas computer vision includes 3D analysis from 2D images. This analyzes the 3D scene projected onto one or several images, e.g., how to reconstruct structure or other information about the 3D scene from one or several images. Computer vision often relies on more or

less complex assumptions about the scene depicted in an image. Machine vision is the process of applying a range of technologies and methods to provide imaging-based automatic inspection, process control and robot guidance in industrial applications.

Machine vision tends to focus on applications, mainly in manufacturing, e.g., vision based autonomous robots and systems for vision based inspection or measurement. This implies that image sensor technologies and control theory often are integrated with the processing of image data to control a robot and that real-time processing is emphasized by means of efficient implementations in hardware and software. It also implies that the external conditions such as lighting can be and are often more controlled in machine vision than they are in general computer vision, which can enable the use of different algorithms.

Pattern recognition is a field which uses various methods to extract information from signals in general, mainly based on statistical approaches. A significant part of this field is devoted to applying these methods to image data and computer vision deals visual patterns and their interpretations.

Video understanding addresses the understanding of video sequences, e.g., recognition of activities or events inside a video. The main difference between a single image and a video (a sequence of images) is motion. Therefore, the major transition in the classic paradigm has been from the recognition of static objects in the scene to motion-based recognition of actions and events. The most interesting subject of majority of the videos are about human, therefore, the analysis of human motion has become a main focus of video understanding research.

In this thesis, we investigate the subject of human action understanding as a research problem of computer vision and video understanding. We search important visual characteristics that can help in proposing computer vision algorithms for human action recognition. We explore rich visual cues from geometrical relationships, spatio-temporal patterns and features, frequency domain signal analysis and contextual associations of actions and actors to derive action representations for machine recognition.

1.1 Human Action Recognition: An Overview

Human action recognition is a problem in computer vision intended to label video sequences with action labels. This may be regarded as a classification problem due to extraction of corresponding discriminative static or motion features, building a model or representation and labeling it into different action classes.

This recognition can be performed at various levels of abstraction. Different taxonomies have been proposed in this regard. A simple hierarchy is based on action primitive/element, action and activity [10]. An action primitive is an atomic movement that can be described at the limb level. An action consists of action primitives and describes a, possibly cyclic, whole-body movement. Finally, activities contain a number of subsequent actions, and give an interpretation of the movement that is being performed. For example, left leg forward is an action primitive, whereas running is an action. Jumping hurdles is an activity that contains starting, jumping and running actions. These are low level, higher level and mid-level vision problems respectively. Once actions are recognized using some representation of action primitives, activity recognition can be performed based on sequence of actions.

One common approach [11] is to divide the problem into two phases: (i) action representations and (ii) action classification. Both phases have their own significance and challenges based on the data and application in hand. Here we introduce some well-known achievements in each domain.

1.1.1 Action Representation

Action representation include extraction of important discriminative features from video sequences or modeling of representative characteristics of actions suitable for use in action classification. In other words, these representations must be sufficiently rich to allow for robust classification of the actions. Ideally, they should generalize over small variations in actor appearance, background, viewpoint and action execution. Time is an important parameter. Some of the action representations explicitly take into account the temporal dimension, others extract static features for each frame in the sequence individually and deal it at classification stage. We further divide these representations into two categories:

- global representations
- local representations

(i) Global Representations: Global representations work in a top-down fashion [12]: first, an actor is localized using background subtraction or tracking. Then, the region of interest is encoded as a whole, which results in corresponding descriptor. These representations are powerful since they encode much of the relevant information. However, they rely on accurate localization, background subtraction or tracking. In addition, these representation are more sensitive to viewpoint, noise and occlusions. In case of good control of these factors, global representations achieve considerable performance.

Common global representations are derived from silhouettes, edges or optical flow. They are sensitive to noise, partial occlusions and variations in viewpoint. To partly overcome these issues, grid-based approaches have been proposed that spatially divide the observation into cells, each of which encodes part of the observation locally. Multiple frames over time can be stacked, to form a three-dimensional space-time volume, where time is the third dimension.

For instance, the silhouette of a person in the image can be obtained by using background subtraction. Generally, silhouettes contain some noise due to imperfect extraction. Also, they are somewhat sensitive to different viewpoints, and implicitly encode the anthropometry of the person. They encode a great deal of visual information. When the silhouette is obtained, there are many different ways to encode either the silhouette area or the contour. Two popular representations include silhouettes from a single view and aggregate differences between subsequent frames of an action sequence [13]. This results in a binary motion energy image (MEI) that indicates where motion occurs. Another representation is motion history image (MHI) that is constructed where pixel intensities are a recency function of the silhouette motion.

Instead of (silhouette) shape, motion information can be utilized. Motion within the region of interest (ROI) can be described with optical flow, the pixel-wise oriented difference between subsequent frames [102]. Flow information can be used when background subtraction is difficult to perform. However, dynamic backgrounds can introduce noise in the motion descriptor. Similarly, camera movement results in observed motion, which can be compensated by tracking the actor. By dividing the ROI into a fixed spatial or temporal grid, small variations due to noise, partial occlusions and changes in viewpoint can be partly overcome. Each cell in the grid describes the image observation locally, and the matching function is modified accordingly from global to local. These grid-based representations resemble local representations, but require a global representation of the region of interest (ROI).

A 3D spatio-temporal volume (STV) is formed by stacking frames over a given sequence. Accurate localization, alignment and possibly background subtraction are required. Motion history volume [15] and 3D maximally stable volume (MSV) [16] are examples of 3D spatio-temporal volumes.

(ii) Local Representations: Local representations work in a bottom-up fashion [12]: first, spatio-temporal interest points are detected, and local patches are calculated around these points. Finally, the patches are combined into a final representation in form of descriptors. They represent the observation as a group of local independent patches. Patches are sampled either densely or at space-time interest points. Local representations are comparatively assumption free, less sensitive to noise and partial occlusion, and do not strictly require background subtraction or tracking. However, these representations depend on the extraction of a considerable amount of relevant interest points.

For instance, space-time interest points [17] are the locations in space and time where sudden changes of movement occur in the video. It is assumed that these locations are more informative for the recognition of human action. Space-time interest points are those points where the local neighborhood has a significant variation in both the spatial and the temporal domain. The scale of the neighborhood is automatically selected for space and time individually. Usually, points that undergo a translational motion in time will not result in the generation of spacetime interest points. One example of space-time interest points is 3D cornet detector. An improved example is 3D cuboid features [18] that use Gabor filtering on the spatial and temporal dimensions. The number of interest points is adjusted by changing the spatial and temporal size of the neighborhood in which local minima are selected.

Local descriptors restate an image or video patch in a representation that is ideally invariant to background clutter, appearance and occlusions, and possibly to rotation and scale. The spatial and temporal size of a patch is usually determined by the scale of the interest point. One solution is the calculation of the patches of normalized derivatives in space and time. Another example is local HOG (histogram of gradients) and HOF (histogram of oriented flow) descriptors [14]. Several approaches combine interest point detection and the calculation of local descriptors in a feed-forward framework with feature matching and bag-of-words model. The comparison and matching of local descriptors is not straightforward due to the their detections and high dimensionality of the corresponding descriptors. Therefore, often a codebook is generated by clustering patches and selecting either cluster centers or the closest patches as codewords. A local descriptor is described as a codeword contribution. In this way, a frame or video sequence can be represented as a bag-of-features, a histogram of codeword frequencies.

1.1.2 Action Classification

When a suitable action representation is available for an observed frame or sequence, human action recognition remains as a classification problem. An action label or distribution over labels is given for each frame or sequence. There are some classifiers that classify image representations into actions without explicitly modeling variations in time domain (direct classification) whereas other approaches do model such variations of an action (temporal state-space classification).

Direct classification approaches [19] deal all frames of an observed sequence as a single representation or perform action recognition for each frame individually. Dimensionality reduction approaches come in this category. In majority of cases, image representations are high-dimensional. It makes them computationally very expensive. In addition, these representations might contain noisy features. To deal with this problem, a more compact, robust feature representation is obtained by embedding the space of image representations onto a lower dimensional space. This embedding can be learned from the training data. PCA (Principal Component Analysis)[20] is a common linear dimensionality reduction method while KPCA (Kernel Principal Component Analysis) [20] is non-linear dimensionality reduction approach. Dimensionality reduction methods learn the embedding in an unsupervised manner and do not guarantee good discrimination between classes.

Another example is k-Nearest neighbor (kNN) classifier [20] that uses the distance between the image representation of an observed sequence and those in a training set. The most common label among the k closest training sequences is chosen as the classification. For a large training set, such comparisons can be computationally expensive. Alternatively, for each class, an action prototype can be calculated by taking the mean of all corresponding sequences. Their ability to cope with variations in spatial and temporal performance, viewpoint and image appearance depends on the training set, the type of image representation and the distance metric. KNN classification can be either performed at the frame level, or for whole sequences.

The third example is discriminative classifiers that focus on separating two or more classes, rather than modeling them. e.g. Support vector machines (SVM) [20] learn a hyperplane in feature space that is described by a weighted combination of support vectors. In a boosting framework, a final strong classifier is formed by a set of weak classifiers, each of which usually uses only a single dimension of the image representation.

The time domain consideration is used by temporal state-space models [12] that consist of states connected by edges. These edges model probabilities between states, and between states and observations. In these models, each state summarizes the action performance at a certain moment in time. An observation corresponds to the image representation at a given time. Temporal state-space models are either generative or discriminative.

A generative model is a model for randomly generating observable data, typically given some hidden parameters. It specifies a joint probability distribution over observation and label sequences. Hidden Markov Models (HMM)[21] use hidden states that correspond to different phases in the performance of an action. They model state transition probabilities and observation probabilities. To keep the modeling of the joint distribution over representation and labels tractable, two independence assumptions are introduced. First, state transitions are conditioned only on the previous state, not on the state history. This is the Markov assumption. Second, observations are conditioned only on the current state, so subsequent observations are considered independent. The independence assumptions in HMMs assume that observations in time are independent, which is often not the case.

Discriminative models [22] overcome this issue by modeling a conditional distribution over action labels given the observations. These models can take into account multiple observations on different timescales. They can be trained to discriminate between action classes rather than learning to model each class individually, as in generative models. Discriminative models are suitable for classification of related actions that could easily be confused using a generative approach. In general, discriminative graphical models require many training sequences to robustly determine all parameters.

1.1.3 Application Areas

The application and useability of action recognition is widely recognized. The important application areas for automatic human action recognition and understanding include human-computer interfaces, content based video indexing, video surveillance, and robotics. Various new application areas have been suggested recently by researchers. Here, we briefly mention few important application areas:

(i) Video Indexing and Retrieval: In recent years, internet has emerged with a great amount of multimedia content. Popular websites like YouTube, Google, Facebook provide opportunity to their users to upload and publish their own images and videos. The move towards user-generated content is motivated by a number of factors, primarily the decrease in the cost of devices like digital cameras, high quality mobile phones, high-bandwidth connections, and great popularity of online social networking web sites. The result is an overwhelming increase in the amount of multimedia content mostly in form of videos. To make multimedia data effectively available, the high-level indexing

aimed at meaningful, semantically-oriented retrieval is a critical goal. While for text, the words themselves convey quite directly its semantics, in the case of visual information, the connection between low-level encoding (i.e., pixels) and semantic meaning is far from immediate. In these scenarios, high level computer vision can play its part by introducing visual semantics. As principal video subject is human and their actions, automated action labeling can speedup video matching and retrieval.

(ii) Video Surveillance: In the context of smart video surveillance, robust action recognition constitutes an essential capability which can improve upon current manual inspection processes. Although video surveillance systems are already in use, videos recorded by these surveillance systems are usually stored in the form of recording for manual inspections later on. This post-processing behavior loses an important benefit as an active real-time warning system. Action recognition systems which are both robust and efficient will likely have a great impact on the transition of video surveillance from a forensic tools that are used after the fact to active crime prevention systems.

(iii) Human Computer Interface: Action recognition can prove an important extension to existing speech-based control systems within human-computer interfaces. Action recognition provides better detailed visual cues through action and gesture recognition as well as facial action classification. Robust methods for recognizing human motion patterns are sources of providing automatic sign-language translation between agents and signaling specific instructions in high-noise environments. An example is KidsRoom [23], an environment able to interpret and react to specific actions of a group of children in a closed space. A similar application is proposed in a system called smart classroom, where the actions performed by a teacher are recognized to allow automatic camera motion and a virtual mouse. Similarly, facial actions have been recently explored as a tool to enhance HCI (human computer interface) to analyze the affective behavior of psychiatric patients [24].

(iv) Analysis of sports videos: The analysis of sport videos is another useful application of automated action recognition. An example is the video summarization in which the classification of video segments between play and break intervals is suggested to summarize the video, by taking out the breaks. Soccer games are also analyzed, in which text and the players trajectories are used to build a system aimed at helping coaches in tactical analysis. In a similar system, six actions of a cricket umpire are analyzed using an appearance based method similar to eigenspaces (commonly used in face recognition) whereas the usage of local motion analysis is employed to identify different swimming styles.

(v) Medical Applications: Most recently, applicability of action is proposed in various applications of medical science [25]. For example, in the medical area, human motion analysis can aid diagnosis of motor problems by comparing patient motion to normality patterns. It can be done using analysis of action trajectories and their comparison. Another possible medical application is to provide remote assistance to elderly people such as fall detection. Automated aged care support is beneficial for old houses to monitor their residents in best possible way. Action detection and recognition can help generating alarm in situations when patient or the old house residents are unable to inform to fall or inactive situation.

1.1.4 Action Recognition Research Datasets

In this research work , we have used almost all publicly available action datasets. The use of publicly available datasets allows for the comparison of different approaches and gives insight into the inabilities of respective methods. These action data sets are well-known in action recognition community and vary in terms of complexity, type of capturing environment and camera setups. The approaches proposed in this thesis are tested for

these datasets and compared with state-of-the art research work. We discuss the most widely used sets.

(i) Weizmann human action dataset [26] : The human action dataset recorded at the Weizmann institute contains 10 actions (walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack and skip), each performed by 10 persons. The backgrounds are static and foreground silhouettes are included in the dataset. The camera viewpoint is static. In addition to this dataset, two separate sets of sequences were recorded for robustness evaluation which include walking movements viewed from different angles. The second set shows fronto-parallel walking actions with slight variations (carrying objects, different clothing, different styles). This dataset is captured in controlled environments and now considered as a primitive but widely used benchmark action recognition dataset. Silhouettes and volumetric voxel representations are part of the dataset.

(ii) KTH human motion dataset [27] : The KTH human motion dataset contains six actions (walking, jogging, running, boxing, hand waving and hand clapping), performed by 25 different actors. Four different scenarios are used: outdoors, outdoors with zooming, outdoors with different clothing and indoors. There is considerable variation in the performance and duration, and somewhat in the viewpoint. The backgrounds are relatively static. Apart from the zooming scenario, there is only slight camera movement. This is comparatively large dataset and used by the majority of the proposed action recognition approaches as benchmark.

(iii) INRIA XMAS multi-view dataset [28] : IXMAS dataset is the widely recognized multiple view action dataset used by the majority of view-invariant action recognition approaches. This dataset contains actions captured from five viewpoints. A total of 11 persons perform 14 actions (check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw over head and throw from bottom up). The actions are performed in an arbitrary direction with regard to the camera setup. The camera views are fixed, with a static background and illumination settings. Silhouettes and volumetric voxel representations are part of the dataset.

(iv) The UCF sports action dataset [29] : This dataset contains 150 sequences of sport motions (diving, golf swinging, kicking, weightlifting, horseback riding, running, skating, swinging a baseball bat and walking). Bounding boxes of the human figures are provided with the dataset. For most action classes, there is considerable variation in action performance, human appearance, camera movement, viewpoint, illumination and background.

(v) Hollywood human action dataset [30]: The Hollywood human action dataset contains eight actions (answer phone, get out of car, handshake, hug, kiss, sit down, sit up and stand up), extracted from movies and performed by a variety of actors. A second version of the dataset includes four additional actions (drive car, eat, fight, run) and an increased number of samples for each class. One training set is automatically annotated using scripts of the movies, another is manually labeled. There is a huge variety of performance of the actions, both spatially and temporally. Occlusions, camera movements and dynamic backgrounds make this dataset challenging. Most of the samples are at the scale of the upper-body but some show the entire body or a close-up of the face.

(vi) WVU Multiview Action Dataset [31]: The dataset is collected as part of the research work on real-time human action recognition in a camera network at West Virginia University, USA. The multi-camera network system consists of 8 cameras that provide completely overlapping coverage of a rectangular region R (about 50 x 50 feet) from different viewing directions. It contained 11 actions, each performed by 10 actors three times and captured from five different views. These actions include nodding head,

clapping, waving 1 hand, waving 2 hands, punching, jogging, jumping jack, kicking, picking, throwing and bowling. This dataset is relatively new but has full potential to be used for testing of multi-view action recognition performance.

1.2 Challenges and Objectives

Visual recognition of human actions constitutes one of the most challenging problems in computer vision. Over the years, several techniques have been developed, yet it is widely recognized that effective solutions are needed to be proposed and investigated. It is due to the nature of problem that combines the unpredictable human behavior, complex human motion dynamics, strong variations in camera environment especially viewpoint, occlusion and noise, presence of anthropometric differences and uncertainty associated with computational vision. It is now understood that human body has no less than 244 dof (degree of freedom) and therefore, modeling this non-rigid dynamics is extremely difficult task [63]. Anthropometry [64] is another important factor as action dynamics is effected due to changes in ethnicity, class, gender, culture and style, circumstances and choice. These immense challenges make action recognition a daunting vision problem worth consideration of extensive research.

The gigantic sum of challenges drive us to find solutions. However, it is more beneficial to short-list and emphasis certain important aspects to thoroughly investigate a problem. Therefore, we select two important factors effecting action recognition performance for investigation in this thesis. These important factors include: (i) Viewpoint Variations and (ii) Contextual Environment.

Viewpoint: Machine vision is greatly dependent on the camera viewpoint. It is due to the fact that different viewpoints of the same action result in different motion patterns and scene projections and the same action may look quite different when observed from different viewpoints. Therefore, to make practical use of action recognition, it is inappropriate to place restrictions on the possible viewpoint of the camera which makes view-invariant action recognition a quite challenging problem. Except in some specific application, it is unreasonable to assume constant camera viewpoint. Unfortunately, majority of action recognition approaches are dependent on camera viewpoint and restricted to single viewpoint and this daunting problem deserves extensive investigation. The problem of view-invariance is thoroughly investigated in this thesis and different algorithm have been proposed to achieve view-invariance in action recognition.

Contextual Environment: Contextual information is important for interpreting human actions especially when actions exhibit interactive relationship with their context. A broad classification of these environments is: (i) controlled environment, (ii) uncontrolled environment and (iii) exceptional environment. The earlier work on action recognition has been done considering a controlled contextual environment and action recognition research is performed on action datasets capture in some specific environment. The recent trend is to deal uncontrolled environment like YouTube video. Another challenge is to target most challenging contextual environments like action recognition in night vision video sequences. This thesis deals action recognition in all three kind of environments which include controlled, uncontrolled environments and a challenging case study of nighttime action recognition.

1.3 Motivations

Our interests and desired benefits boost our confidence and develop our motivations. The motive behind the research work in this thesis is based on the very artifact of interest to

work and advantages of the related research work. The factors that motivated us for the exploration and research in the area of human action understanding as as follows:

(1) Action recognition is very active research area in computer vision. Majority of low and mid-level vision problems have been solved by computer vision scientists and now computer vision is focusing on solving higher level vision problems. Visual action recognition is a high level computer vision problem that encapsulates the knowledge, scope, achievements and challenges by computer vision. On one hand, the solutions for action recognitions are based on low and mid-level vision solutions and on the other hand, it can help to achieve higher level computer vision goals.

(2) An important personal drive for learning a new field is its usefulness and applicability in solving real world problems. Automated action recognition is very rich in its application and useability. Action recognition has various important applications. These application areas include but not limited to human-computer interfaces, content based video indexing, video surveillance, robotics and medical science. The application areas are one of the most important motivational factors behind the research on action analysis, detection, understanding and recognition. In addition, various other application area are being sorted and suggested by computer vision researchers.

(3) The challenges being faced by robust action recognition are also recognized by other areas of computer vision. In case, if we find the solutions for dealing with these challenges and problems, other related fields of computer area can get benefit from it. The major problems of handling noise, occlusion, temporal variations, feature extraction, robust matching, avoiding tracking, intra-class variation handling and accurate classifications methods are universally recognized by computer vision, image processing and machine learning research communities. The similar challenges are faced by visual action recognition during the development of its solutions. Therefore, the research on action recognition is indirectly beneficial to other areas of computer vision and related scientific disciplines.

(4) Over the last few years, several approaches have been devised to address automated recognition of human actions. These proposed approaches vary in their accuracy and complexity. Despite these solutions, various research gaps are rightly pointed out in our research which motivated us to work in this area. In this thesis, we tried to address these research gaps by proposing approaches and solutions that can handle these problem areas. The detailed discussion of these research gaps and our contribution to propose novel solutions would be described in the next section. However, these research gaps are related to unexplored or unfulfilled exploitation and utilization of important visual cues for automated recognition of human actions in video sequences.

1.4 Contributions

During this research work, we have found important research gaps in the area of visual action recognition and proposed novel solutions with far-reaching effects on development of the respective field. These research gaps and respective solution are the artifacts of this thesis.

Important visual cues and their meaningful representations are of fundamental importance for visual recognition. We explore rich visual cues from geometrical relationships, spatio-temporal patterns and features, frequency domain signal analysis and contextual associations of actions to derive action representations for machine recognition. Similar to the physical world that is composed of rich structures and documents consist of a large number of units (such as characters, words, phrases, and sentences), images and videos can also be considered as a collection of elements (pixels, voxels, edges, patches, etc.). Salient visual cues expressed into these elements help in development of better representations.

We have used these visual cues to develop meaningful representations for better human action understanding. Some of the contributions being claimed in this research work are as follows:

1.4.1 Temporal Order Invariance for view-invariant action recognition (Chapter 3)

Action is known as a spatio-temporal phenomenon consisting of various local variations and patterns (e.g. spatial and temporal gradients). These local space-time patterns can be characterized by defining space-time interest points and action elements can be represented through these interest points. A pioneer work in this direction is spatio-temporal corners [17]. Later on, similar other features are found with improved performance for action recognition. These features are characterized by their computational simplicity, robustness to occlusion, elimination of low level object detection and tracking, and existence of scalable matching schemes. After feature extraction from action sequences, representation framework like bag of visual words [115, 123, 30] is utilized for action matching. While similarity/dissimlarity of these interest points is sought during matching same or different actions, their mutual relationships like space time organization and ordering is ignored.

In addition, these features lack view-invariance and therefore related action recognition approaches are not view-invariant. Most recently, some techniques [51, 52] have been developed to increase exploitation of spatio-temporal features for achieving view-invariance but these approaches are based on the improvement of matching framework like classifier fusion [51] or extended vocabularies rather than a global view-invariant characteristic [52].

To address these limitations and complications, we propose a novel spatio-temporal action matching framework based on discriminative combination of 3D features, named spatio-temporal ordered packets (STOPs), that combines space-time features along with their geometric ordering information into spatio-temporal volumes avoiding complex ordering constraints. Packaging features into volumes ensures that discriminative power of matching is enhanced as whole packets are matched across videos instead of individual features. At the same time matching is made much robust by developing simple matching criteria that take into consideration spatio-temporal order of features within each volumetric packet.

Considering the fact, that an 'action' is essentially a spatio-temporal construct, ignoring temporal order in which spatio-temporal features occur can affect matching performance drastically, especially where various actions have many overlapping low level features. Therefore, we focus on global analysis of human actions and seek a view-invariant representation. We based our approach on the following conjecture: "*The temporal order of actions elements within an action is invariant to viewpoint variations*". We define action elements in terms of local spatio-temporal interest points and define spatio-temporal order preservation constraint in matching framework. Spatio-temporal cuboid features [18] are taken as space-time interest points as these features are based on maximization of discrimination between behaviors. For each action class, we define a feature fusion table. A feature fusion table is a defined data structure to encapsulate multiple training examples against multiple viewpoints for a single action class. It is achieved through features fusion based on principal component analysis. A matching score is then calculated based on global temporal order constraint and number of common features. Finally, the action label of class with maximum value of matching score is assigned to the query action.

Related publications:

- On Temporal Order Invariance for view-invariant action matching, IEEE Transaction on Circuits and Systems for Video Technology, vol. 22, 2012.

1.4.2 Tracking-free and Training-free solutions for View-invariant Action Recognition (Chapter 4)

Multiple view geometry has alleviated many hard problems in computer vision [32, 33, 34]. Estimation of essential matrix and then fundamental matrix from stereo image pair goes back to Longuet-Higgins and eight point algorithm [35]. Therefore, inspired from multi-view geometry, a successful series of incremental work related view invariant action recognition is addressed in [36-44] which is based on the consideration of action point trajectories by a stationary camera and exploitation of epipolar geometry between trajectories of different views of the same action. One of the major benefit of these geometrical based methods is that such methods do not need any training. The basic idea originated with the use of affine epipolar geometry constraints in a series of work [36, 38] which showed that the maxima in space-time curvature of a 3D trajectory are persevered in 2D image trajectories.

The main drawback of these approaches is the assumption of affine cameras. For projective camera model, trajectories of 13 anatomical landmarks are matched by [42] under viewpoint, anthropometric and temporal transforms. Another related work is the use of the point triplets with homography, rank constraint [40] and fundamental ratios [41] which consider that the motion of an articulated body can be decomposed into rigid motion of planes defined by triplet of body points. The main drawback of the all above approaches is the decoupling of tracking and matching. It is assumed that tracking of the landmark points on human body has been performed and trajectories are available. Despite its success, it is hard to achieve as basic assumption is very strong. Due to occlusion and noise, the detection of landmark points is not always robust resulting in manual interventions. As a result detection of landmark points and their tracking is performed manually and epipolar geometry rank constraints are applied on manually obtained trajectories by almost all the representative geometrical based methods [36-44] which lack automation and to make practical use of geometrical solutions, this problem is needed to be addressed.

Recently, [45] has used optical flow based dense correspondences for calculation of static fundamental matrix and showed that it is effective than a sparse set of correspondences. We try to exploit this estimation to solve view invariant recognition of actions in video sequences without tracking. We intend to tackle these drawbacks by proposing a new approach, AVITAR (Achieving View-invariant tracking-free Action Recognition). We explore how dense optical flow can be employed to compensate strong assumptions of landmark point extraction and tracking in epipolar geometry based view invariance action recognition. Taking into consideration that human action is a spatio-temporal phenomenon, we apply constraints on optical flow to be spatio-temporally consistent. Spatio-temporally consistent optical flow helps us in devising spatio-temporally consistent flow fundamental matrix and by defining rank constraints on flow fundamental matrix we are able to derive a dissimilarity score for action sequences.

We proceed incrementally by defining two variants of our approach: (1) We extract actor body silhouettes from original video sequences and calculate spatio-temporally consistent optical flow between respective frames of two videos and then fit epipolar geometry. As fundamental matrix remains same for static scenes, we can calculate action similarity score between two actions being performed in time domain, (2) In addition, we observed that silhouette extraction is not robust in all circumstances especially in case of noise and occlusion. Therefore, we remove pre-processing step of silhouette extraction theocratically by maximizing the exploitation of epipolar geometry. We take action representation in static camera environment as a case of dynamic scene where background is stationary and actor is dynamic. As scene is not entirely static, we get inspiration from structure and motion recovery for scenes consisting of both static and dynamic parts, also known as multi-body segmentation from perspective views without knowing which measurement

belongs to which part of the scene. As we consider only static background and dynamic actor, it is simplified to two-body fundamental matrix, also known as segmentation matrix [46]. It has already been shown [47] that such matrix can linearly be computed from image measurements after embedding all the image points in high dimensional space. Based on these investigations, we derive a new similarity measure for matching actions across different views, without prior segmentation of actors.

Related publications:

- On dynamic view geometry for view-invariant action matching , in Proc. IEEE CVPR 2011.
- AVITAR: Achieving View-invariant Tracking-free Action Recognition, submitted to IEEE Transaction on Image Processing, 2012.

1.4.3 Fast Frequency-domain View-invariant Action Recognition (Chapter 5)

One of the most successful approaches is the application of space-time pattern templates. Earlier work includes temporal matching of periodicity information from a set of optical flow frames by [48] and highly cited [13] which presents a two component temporal template of motion energy image (MEI) and motion history image (MHI). These representations encode, respectively, where motion occurred and the history of occurrences. Another work presents actions as space-time shapes [26] induced by the silhouettes in the space-time volume. It considers space-time saliency utilizing properties of the solution to the Poisson equation. A similar work [49] enforces space-time consistency between template and the target employing a rank based constraint. However, majority of these template-based approaches suffer from high computational overhead.

Recently, the utilization of correlation filters is investigated for recognizing action instances with promising results. The representative work in this regard is the development of Action MACH [29] that has generalized traditional *Maximum Average Correlation Height* (MACH) filter to 3D MACH by including temporal dimension. However, the major gain is in terms of low computational cost as response of the filter can be analyzed in frequency domain. Despite its success, some researchers [50] have indicated inherent discrepancies in MACH filters and questioned their effective utilization for action recognition. One of the weaknesses of MACH filters is their ineffectiveness to encapsulate inter-class variability. Therefore, these filters are trained only for one class at a time and separate MACH filter is needed for every class. Secondly, MACH filters overemphasize average training sample, a biased treatment of low frequency components and behave like average filter and may loose finer details of the training set. They emphasize high energy (low frequency) components and attenuate low energy (high frequency) components of the training set leading to poor intra-class discrimination. In addition, as action datasets are normally misaligned in space and time, they create problems in learning and testing as synthesized filters are not shift-invariant.

We address above mentioned weaknesses and propose an extended *spatio-temporal distance classifier correlation filter* (Action ST-DCCF filter) for action recognition. Our approach offers following advantages: (i) A single Action ST-DCCF filter successfully captures inter-class variability and avoids overemphasize on average training sample by empirically setting contributions of low as well as high frequency information. (ii) Secondly, it presents a different interpretation of correlation filters as method of applying a *spatio-temporal transformation* to the data and transformation matrix is restricted to being Toeplitz ensuring *shift invariance*. It measures similarity between an ideal transformed reference and testing action using a shift-invariant mean square distance measure handling

misalignments and (iii) Another benefit is that resulting decision boundaries are quadratic which are more 'selective' for choosing feature space portions for assigning to various classes and utilize entire correlation plane rather than emphasizing only single point like correlation peak. These advantages of Action ST-DCCF filter can potentially improve action recognition performance.

Related publications:

- Action recognition using spatio-temporal distance classifier correlation filters, In Proc. DICTA 2011.

1.4.4 Contextual Action Recognition in Nighttime Videos (Chapter 6)

The need for understanding actions in context is discussed by different researchers. Scene context is used for event recognition by [54] but it only applies to static images. Recognizing actions in context is discussed by [53] which is formulated on bag-of-features framework and scene-action SVM- based classifier. It is focused on annotated actions in movies and uses script mining for visual learning. A similar approach [55] captures generic object based context by detectors and their descriptors are used as input for supervised learning. More recently, modeling of scene and object context is discussed by [56] for Hollywood2 action datatset. All above approaches target action recognition in high-resolution action videos in movies. One typical benefit available to these approaches is the ease of finding visual interest points and detectors related to actors and their context. Contextual clues become even more crucial when videos are captured in unfavorable conditions like extreme low light nighttime scenarios. These conditions encourage the use of multi-senor imagery and context enhancement.

None of the above approaches discuss nighttime visual context and recognition of actions at nightime. Mostly recently, human action activity recognition is discussed in [57, 58] which focus recognition in infra-red spectrum. However, these approaches ignore action contexts which is not properly captured by infra-red sensors and can not be categorized as contextual action recognition approaches.

We argue that contextual action recognition is not possible using single sensor platform due to the limitations of individual sensor to grab all available visual information about the scene. This situation motivates the use of multiple sensors often of complementary nature.

We explore the importance of contextual knowledge for recognizing human actions in multi-sensor nighttime videos. Information fusion is utilized for encapsulating visual information about actions and their context. Space-time action information is contained using 3D fourier transform of fused action silhouette volume. In parallel, SIFT context images are extracted and fused using principal component analysis based feature fusion for each action class. Contextual dissimilarity is penalized by minimizing context SIFT flow energy.

Related publications:

- Contextual Action Recognition in Nighttime video sequences, In Proc. DICTA 2012

1.4.5 Contextual Enhancement of Nighttime Videos (Chapter 6)

We explore that robust action recognition in multi-sensor scenario is not possible without context enhancement of nighttime video sequence which involves multi-sensor color fusion of multi-sensor videos.

The goal of video fusion is to create a single enhanced video sequence from complementary video inputs that is more suitable for the purpose of human visual perception, object detection and target recognition. Over the years, several image fusion techniques

are developed which vary in their complexity, robustness and quality. One major trend in image fusion research is to sacrifice complexity to gain quality. However, opposed to images, complexity criterion has more significance in video domain which is intended for real time use. Therefore, video applications do not encourage algorithmic complexity and require simple and efficient information fusion.

Color is another important requirement in addition to fusion but colorization of fused grayscale imagery is a daunting task. Most recently, various manual and semi-automatic colorization techniques have been reported in the literature to solve this difficulty. A highly cited work is colorization based on optimization [60] which needs user defined color scribbling. It proves to be an attractive method which requires neither precise image segmentation, nor accurate region tracking based on the idea that neighboring pixels in space-time with similar intensities should have similar colors. However, one shortcoming of this method lies in the requirement that input images are annotated with user defined color scribbles and thus lacks full automation. Another popular work is colorization based on color transfer [62] using statistical analysis to impose one images color characteristics to another image. It uses a de-correlated color space $\ell\alpha\beta$ and swatches for color transfer from target color image. This technique has the same drawback that it requires manual selection of a color target image and swatches. In addition, color space conversions and swatches make additional burden in terms of complexity.

Despite these shortcomings, above approaches have transformed the cumbersome work of manual colorization into semi-automatic colorization. Due to their successful application in colorization and color correction, these techniques are extended for colorizing night vision imagery [181] presenting a software based approach to night vision offering a cheaper and reliable solution. Therefore, it is highly desirable that fully automated colorization should be introduced to facilitate real-time video processing for night vision applications.

The quality assessment of color image fusion comes in the category of blind quality evaluation methods because of the absence of any reference image with optimal fusion and colors. Various blind objective quality measures for grayscale image fusion are available in the literature. To the best of our knowledge no appropriate objective quality measure exists to address diversity of color image fusion frameworks in night vision applications.

We propose a software based approach which overcomes above mentioned limitations by simultaneously fusing information from forward looking infra-red and low light visible sensors and introducing automatic colorization for context enhancement at nighttime. Firstly, corresponding frames from complementary video streams are fused and pseudo-colorized using RGB color channel integration. Then, efficient color morphing technique is used in RGB color space avoiding any color space conversion. Automation is introduced by integrating source color image selection with contextual features and colorfulness characteristics. A night vision system named $SCENT$ is developed based on proposed approach. Quality evaluation shows that our approach not only gives promising fusion and color quality but also proves to be the efficient in terms of execution time.

We propose a novel color image fusion quality measure, CFOI (Color Fusion Objective Index) which encapsulates the powers of color image quality, image colorfulness and fused information index. In addition, it evaluates the gradient structure preservation in color fused image.

Related publications:

- Automating video fusion and colorization for context enhancement at nighttime, Information Fusion, 2012.

- Automated multi-sensor color video fusion for nighttime video surveillance, In Proc. IEEE ISCC 2010

- A novel color image fusion QoS measure for multi-sensor night vision applications, In Proc. IEEE ISCC 2010
- SCARF: semi-automatic colorization and reliable image fusion, In Proc. DICTA 2010.

1.5 Organization of the Thesis

The organization of this thesis is based on the above discussed contributions. All proposed approaches have been arranged according to their uniqueness in different chapters. Every chapter addresses a new visual cue and supports how it can help in providing a better solution. The detail of organizational arrangement of chapters is illustrated in figure 1.1.

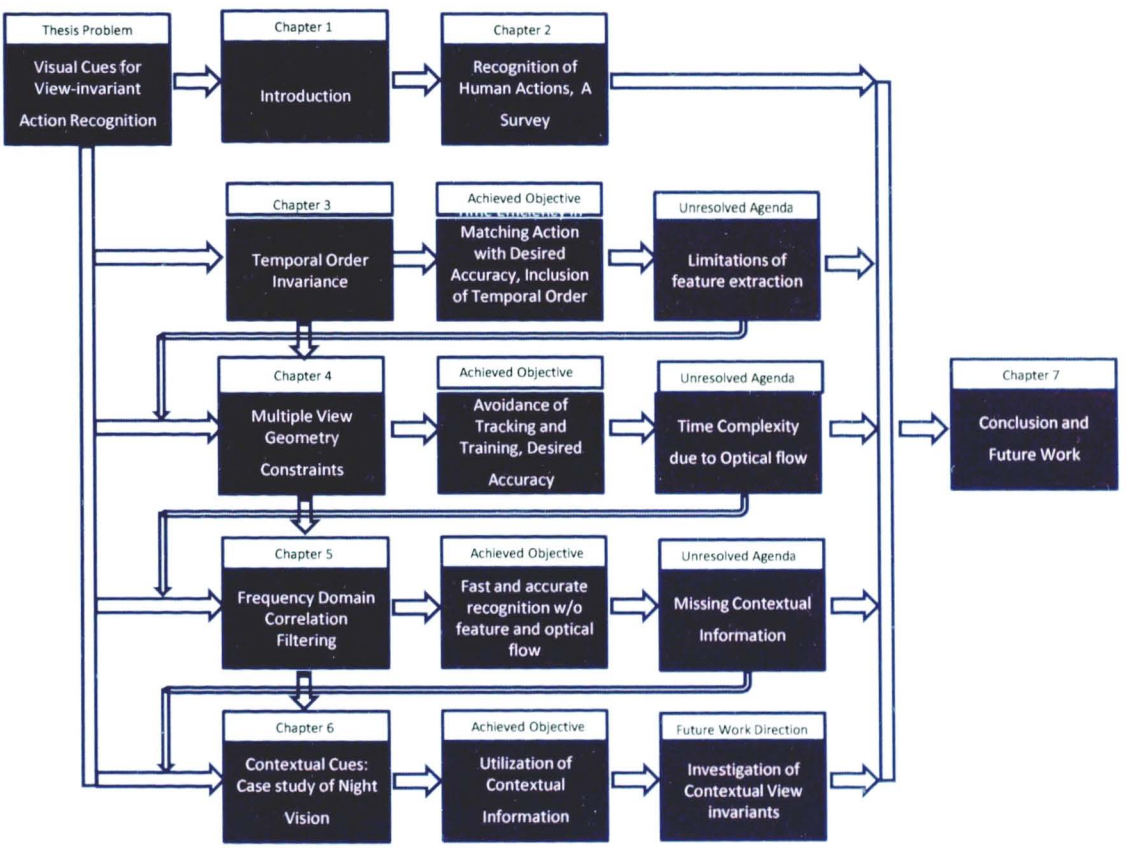


Figure 1.1: Organizational Chart showing the flow of research work in this thesis.

Chapter 2 presents detailed description of literature survey related to the research work in this thesis. Section 2.1 describes space-time feature based approaches for view-invariant action recognition. Section 2.2 presents geometry based action recognition and gives detail about epipolar geometry constraints. Section 2.3 discusses action recognition approaches based on frequency domain correlation filtering and finally techniques about contextual action recognition and context enhancement are mentioned in section 2.4.

Chapter 3 presents importance of order information in action representation. It proposes 3D STOPs (spatio-temporal ordered packets) for improved action recognition and discusses temporal order invariance for view-invariant action recognition.

Chapter 4 discusses multiple-view and dynamic scene geometry and proposes solutions for view-invariant action recognition based on rank constraints. We discuss how we derive

action matching scores across different viewpoint by defining rank constraints on flow fundamental matrices in static and dynamic two-body scenarios.

Chapter 5 discusses correlation filters and frequency domain filtering and proposed spatio-temporal distance classifier correlation filters. The design and performance of these spatio-temporal action filters is presented in detail. It also discusses view clustering for achieving view-invariant action recognition.

Chapter 6 discusses the importance of context for action recognition in challenging scenarios such as night. It proposes contextual action recognition at nighttime based on contextual enhancement of multi-sensor visual information. The context enhancement is proposed based on color video fusion of infra-red and CCD video sequences. In addition, quantitative quality evaluation of such techniques is also proposed in this chapter.

All the chapters present and discuss respective experimental setup, results and conclusion. However, to sum up the detailed results obtained in this thesis, we present conclusions of thorough investigation of all the proposed techniques for action recognition in chapter 7. In addition, we discuss future research directions for the continuity of this research work.

1.6 Conclusions

In this chapter, we introduced our research problem, background introduction, its significance and applications, our motivation to work on this problem and organizational layout of this thesis. In the next chapter, we would present literature survey about visual recognition of human actions in detail and would classify and review previous approaches. In addition, we would provide a linkage of our research contributions towards the removal of the shortcomings present in previous approaches .

Chapter 2

Visual Recognition of Human Actions: A Survey

Research on vision based human action analysis and recognition is increasingly becoming the subject of attention in multiple disciplines such as praxeology [1, 7], psychology [65], cognitive neuroscience [66, 67] and computer vision [68, 69]. Praxeology presents action as an important tool for rational investigation of human decision-making while cognitive science presents action as an atomic unit of human activity to understand cognition. In a similar manner, computer vision emphasizes the analysis of human actions to achieve its high level visual perception. It implies that visual perception of human action is a critical function which is meant to interpret actor's intentions and purpose of his behavior while simultaneously dealing with the complex nature of non-rigid action dynamics, anthropometric variations and other related cognitive challenges. The recognition of movement can be performed at various levels of abstraction. Different taxonomies have been proposed based on full or partial body movements, view dependent or view invariant action representation or modeling and recognition frameworks.

Earlier work on human action recognition goes back to framework of [73, 74] that used a simple representation of stick figures to analyze different poses of an actor. It is followed by a series of work studying different frameworks to analyze, understand and recognize human actions. Until now, several approaches have been proposed on this subject and it is extremely difficult to mention all of them individually. There are several existing surveys within the area of vision based human motion analysis and recognition. Recent surveys include [12, 69, 75, 77, 78, 79, 80, 81] which review action recognition approaches on different bases. This thesis explores important visual cues that can play a primal role in recognizing actions in different challenging circumstances. Rich visual cues from geometrical relationships, spatio-temporal patterns and features, frequency domain signal analysis and contextual associations of actions are used to derive action representations for machine recognition. View dependency of action representations as a core problem is considered, explored and addressed in this work. In this chapter, we analyze the recent approaches in computer vision that are closely related to our work. These approaches are categorized based on the action modeling framework. These frameworks include (i) spatio-temporal feature based framework (Chapter 3), (ii) geometrical modeling (Chapter 4), (iii) frequency domain filtering (Chapter 5) and (iv) contextual action recognition (Chapter 6).

2.1 Feature based Approaches

The use of local spatio-temporal features and their ordering constraint is an important visual cue explored in this thesis. This section reviews approaches that utilize local spatio-temporal features.

Local spatio-temporal features are fast becoming representation of choice for action recognition due to their computational simplicity, robustness to occlusion and minor viewpoint changes, elimination of low level object detection and tracking, and existence of scalable matching schemes. These space time interest points are discriminative as they are computed in a way that they capture not only the pixel intensity information but also the motion information (e.g. statistics of optical flow) in the vicinity of the interest points. After feature extraction from action sequences, representation framework like bag of visual words [115, 123, 30] is utilized for action matching. While similarity/dissimlarity of these interest points is sought during matching same or different actions, their mutual relationships like space-time organization and ordering is ignored.

Various space-time feature detectors [18, 17, 91, 133, 92] and descriptors [135, 136, 53, 96, 133] have been proposed in the past few years. Feature detectors usually select spatio-temporal locations and scales in video by maximizing specific saliency functions. The detectors usually differ in the type and the sparsity of selected points. Feature descriptors capture shape and motion in the neighborhoods of selected points using image measurements such as spatial or spatio-temporal image gradients and optical flow. Laptev [17] evaluated the repeatability of space-time interest points as well as the associated accuracy of action recognition under changes in spatial and temporal video resolution as well as under camera motion. Similarly, Willems et al. [133] evaluated repeatability of detected features under scale changes, in-plane rotations, video compression and camera motion. Local space-time descriptors were evaluated by Laptev et al. [136], where the comparison included families of higher-order derivatives (local jets), image gradients and optical flow. Dollar et al. [18] compared local descriptors in terms of image brightness, gradient and optical flow. Scovanner et al. [96] evaluated the 3D-SIFT descriptor and its two-dimensional variants. Jhuang et al. [91] evaluated local descriptors in terms of the magnitude and orientation of space-time gradients as well as optical flow. Klser et al. [135] compared space-time HOG descriptor with HOG and HOF descriptors [53]. Willems et al. [133] evaluated the extended SURF descriptor.

Space-time feature detectors: The Harris3D detector was proposed by Laptev and Lindeberg in [17], as a space-time extension of the Harris detector [9]. The authors compute a spatio-temporal second-moment matrix at each video point using independent spatial and temporal scale values. They proposed an optional mechanism for spatio-temporal scale selection. The Cuboid detector is based on temporal Gabor filters and was proposed by Dollar et al. [6] and has become popular feature detector over the years. The Hessian detector was proposed by Williams et al. [26] as a spatio-temporal extension of the Hessian saliency measure used in [2, 18] for blob detection in images. The detector measures the saliency with the determinant of the 3D Hessian matrix. The position and scale of the interest points are simultaneously localized without any iterative procedure. In order to speed up the detector, the authors used approximative box-filter operations on an integral video structure. Each octave is divided into 5 scales, with a ratio between subsequent scales in the range 1.2-1.5 for the inner 3 scales. The determinant of the Hessian is computed over several octaves of both the spatial and temporal scales.

Space-time feature descriptors: The HOG/HOF descriptors were introduced by Laptev et al. in [136]. To characterize local motion and appearance, the authors compute histograms of spatial gradients and optic flow accumulated in space-time neighborhoods of detected interest points. Normalized histograms are concatenated into HOG, HOF as well as HOG/HOF descriptor vectors and are similar in spirit to the well known SIFT

descriptor. The HOG3D descriptor was proposed by Klser et al. [135]. It is based on histograms of 3D gradient orientations and can be seen as an extension of the popular SIFT descriptor [100] to video sequences. Gradients are computed using an integral video representation. Regular polyhedrons are used to uniformly quantize the orientation of spatio-temporal gradients. Williams et al. [133] proposed the extended SURF (ESURF) descriptor which extends the image SURF descriptor [180] to videos. Dollar et al. [18] proposed the Cuboid descriptor along with the Cuboid detector. Principal component analysis (PCA) is used to project the feature vector to a lower dimensional space.

Considering the fact that an 'action' is essentially a spatio-temporal construct, ignoring temporal order in which spatio-temporal features occur, can affect matching performance drastically, especially where various actions have many overlapping low level features. In addition, the discriminative power of spatial relationships between these low level features play an important role when dealing with actions that closely mimic each other e.g., jogging and running. To overcome these shortcomings several competing approaches have been proposed recently.

For instance, naive ordering is imposed by dividing the space-time volume of a video into space-time bins much like spatial pyramid matching used for images [108, 53, 109, 123]. However, as opposed to images, content of a video can vary drastically depending on how input video is segmented (along temporal scale), and location and speed of action, thus making such rigid binning scheme hard to generalize. Binning neighborhoods of interest points at various scales for descriptor computation is another approach [126, 127]. However, this approach also suffers from the rigidity imposed by fixed scales at which computations are performed. Schemes for large scale image retrieval [124, 125] present similar inspiring ideas like bundling spatial features and geometric verification based on area ratio of triangle generated by two visual words. Again, these approaches are restricted to spatial domain as their application goal is entirely different from action recognition which focus more on temporal ordering. In addition, triangle based spatial ordering is difficult to visualize in video domain.

Recently, Kovashka et al.[110] proposed an feature centric approach where each feature maintains orientation and location information of neighboring features at various scales in spirit of shape context feature [112]. Scales are computed in such a way that makes the local groupings of feature discriminative in terms of action specific distance metrics. This approach mitigates issues of feature discrimination and spatio-temporal ordering to some extent, however, we believe feature groups should be a construct grounded in or associated with the object (or space-time segment) inducing the action as oppose to isolated feature points. This will have the following benefits:

- No need to search across scales for locating discriminative groupings. They will always be associated with the space-time segment that implicitly captures the shape and motion of the object performing an action
- Ordering will have natural explanation in terms of subunits of an action. For instance, hand going up and down for hand waving action
- A multi-feature representation that fuses space-time regions with local interest points.

Part based action representations: Unit formation is fundamental to visual perception. Part based action representations present unit formation of actions. These approaches can be categorized into two types: first type utilizes visual appearance of parts and geometrical constraints. In contrast, the other type relaxes structural constraints and represents action video sequence as a set of independent features. The latter approach is popular and widely used due to its computational simplicity. It represents action formation in terms of spatio-temporal features [17] that are marked by their elimination of

tracking, robustness to occlusion and scalable matching like bag-of features framework. The major weakness inherent in these approaches is due to the ignorance of mutual relationships of features like space time organization and temporal ordering in bag-of-features framework. It led to the study of temporal structure of actions.

Temporal structures of actions: The importance of temporal order is investigated by different researchers. Temporal composition of different motion segments for recognizing human activities is studied by [115]. It adopts an action representation based on spatio-temporal interest points (STIPs) [17] while video sequence is decomposed into temporal segments of various length. Discriminative subsequence mining is proposed in [138] to find optimal discriminative subsequence patterns represented by spatio-temporal cuboid detectors [18]. Finally, visual words arranged into temporal bins are presented for classification. Short subsequences called action snippets of 1-7 frames long are proposed in [131] to alleviate temporal segmentation of actions using form and motion features. Most recently, ACTOM sequence modeling is proposed by [141] that represents temporal structure of actions as a sequence of histogram of actom-anchors visual features using spatio-temporal interest points (STIPs) [13]. However, these actom are manually annotated at training level.

Above mentioned approaches utilize temporal order information for improved action recognition performance. Therefore, their objective is to overcome the weaknesses of bag-of-features framework like spatial bag-of-word [139]. Unfortunately, the success of these approaches is marginal as these approaches are not view-invariant. However, our objective is different from above approaches as we explore global temporal order within human actions to seek view-invariant action recognition which has not been investigated by any previous work. Subsequently, we propose a novel notion of *temporal order invariance* for scalable framework for view-invariant action recognition.

We focus on the global analysis of human actions and seek a view-invariant representation. We based our approach on the following conjecture: "*The temporal order of actions units within an action is invariant to viewpoint variations*". We define action units in terms of local action dynamics and motion variations encapsulated by local spatio-temporal interest points and define spatiotemporal order preservation constraint in matching framework. Spatiotemporal cuboid features [18] are taken as space-time interest points as these features are based on maximization of discrimination between behaviors. For each action class, we define a feature fusion table. A matching score is then calculated based on global temporal order constraint and number of matching features. Finally, the action label of class with maximum value of matching score is assigned to the query action.

2.2 Geometry based Approaches

Another visual cue that has been explored in this work is geometrical coherence and modeling for view invariant action recognition. In this section, we only focus on geometrical action recognition approaches in literature.

Multiple View Geometry and its Applications: Multiple view geometry has alleviated many hard problems in computer vision [32, 33, 34]. Estimation of essential matrix and then fundamental matrix from stereo image pair goes back to Longuet-Higgins and eight point algorithm [35]. The fundamental matrix is the algebraic form of the intrinsic geometry between two views. Within just a few years, applications of the fundamental matrix were found in 3D scene reconstruction, stereo camera applications, image alignment, video synchronization. In [32], Hartley and Zissemman summarized the work related to the fundamental matrix between multiple views of the same static scene. Based on the significance of the fundamental matrix, researchers tried to extend its use to other types

of applications. By relaxing constraints on static camera, researchers explored new ways of geometric coherence.

Avidan and Shashua [82] considered the case where objects can move freely along lines or conics in 3D and there is no constraint on the camera motion. In this case, the 3D motion trajectories can be recovered from sequence of images. Similarly, constraints on the object motion along lines in 3D were introduced in [82] but allowed a scene to contain both stationary and moving objects. By recovering the camera motion, static and moving objects can be automatically segmented out for direct application to 3D scene reconstruction. By relaxing the constraint on the constant speed of moving objects and imposing planer motion constraint, [72] proposed a multi-view C-tensor similar to multi-focol tensor. In [84] a case of moving stereo system is considered observing rigid objects that move arbitrary along plane. At each time instant the moving stereo system gives a 3D view of the scene. A tensor for matching 3D views of a scene is proposed that is analogous to the fundamental matrix between two views. Similarly, [85] describe a space time projection model for Galilean camera and propose a mapping function between the videos of two Galilean cameras when the scene is planer. It proposes a normalized linear algorithm for estimating the parameters of the fundamental matrix relating Galilean cameras.

In recent years, inspired from multi-view geometry, a successful series of incremental work related view invariant action recognition is addressed in [36, 37, 38, 39, 40, 41, 42, 43, 44] which is based on the consideration of action point trajectories by a stationary camera and exploitation of epipolar geometry between trajectories of different views of the same action. One of the major benefits of these geometrical based methods is that such methods do not need any training. The basic idea originated with the use of affine epipolar geometry constraints in a series of work [36, 38] which showed that the maxima in space-time curvature of a 3D trajectory are persevered in 2D image trajectories.

View Invariance Action Recognition: The research on view-invariance action recognition now spans almost a decade and several representations have been sorted out. Major representations include trajectory based approaches [36, 37, 38, 39, 40, 41, 42, 43, 44], spatio-temporal templates [28, 49], view invariant features [71] and the exploitation of space-time interest points [51, 52]. This increasing interest is due to the objective of achieving unconstrained action recognition intended to various applications like video surveillance, human computer interaction, video search and retrieval. It demands action recognition approaches to be stable to view changes to exploit their practical use.

Action is known as a spatio-temporal phenomenon consisting of various local variations and patterns (e.g., spatial and temporal gradients). These local space-time patterns can be characterized by defining space-time interest points and action elements can be represented through these interest points. A pioneer work in this direction is spatio-temporal corners [17]. Later on similar other features [96, 18] are proposed with improved performance for action recognition. These features are characterized by their computational simplicity, robustness to occlusion, elimination of low level object detection and tracking, and existence of scalable matching schemes. However, these features lack view-invariance and therefore related action recognition approaches are not view-invariant. An exception is the work of [71] that has proposed self-similarity matrices as view-invariant features that prove useful in matching actions across different views.

In addition, some work has emerged with the aim of proposing a view-invariant matching framework in which local space time features play a building role. An approach [51] based on local partitioning and hierarchical classification of the 3D Histogram of Oriented Gradients (HOG) descriptor to represent sequences of images into a data volume. Action classification is achieved through a hierarchy of classifiers. Another work [52] uses view knowledge transfer based on bipartite graph to model two view-dependent vocabularies

and apply bipartite graph partitioning to co-cluster two vocabularies into visual-word clusters called bilingual-words (i.e., high-level features) to bridge the semantic gap across view-dependent vocabularies. This approach is based on bag-of-feature framework which has inherent property of its orderless formation.

Coming back to geometrical methods that are more closer to our work, we find various interesting approaches. For stationary camera based view invariant action recognition system, a series of work [36, 38] presents a view invariant action representation consisting of dynamic instants and intervals. Motion is represented by a sequence of dynamic instants where a dynamic instant is an instantaneous entity representing a significant change of any motion characteristic (i.e., speed, direction, acceleration) and is detected by identifying maxima in the spatio-temporal curvature. Then with affine camera model, a rank constraint is derived to match different trajectories generated by different or same actions. Dynamic time warping is used to synchronize trajectories with temporal variations. This work is able to match same actions from different viewpoints but use of manual trajectories and affine camera model limits its practical use.

Recently, matching trajectories of anatomical landmarks using projective camera model is proposed by [43]. It uses rank constraint based matching score based on condition number of the observation matrix formed from 13 landmark points and shows robustness to viewpoint variations, anthropometric and temporal transforms. The main drawback is again the manual use of landmark detection and their trajectories.

Invariant space trajectories (ISTs) are proposed in [37] with plane formed by five landmark points on human body. It models actions in terms of view-invariant canonical body poses and trajectories in 2D invariance space, leading to an effective way to represent and recognize human actions from a general viewpoint but the way of generating invariant space trajectories is not automated.

The application of multiple view geometry of moving cameras is explored by [44] which forms rank constraint on 27 point correspondences based observation matrix. The subject of moving camera scenario is quite interesting but the basic framework also assumes that trajectories are available.

The idea of point triplets is utilized by another series of work [40, 41] that states that motion of an articulated body can be decomposed into rigid motions of planes defined by triplets of body points. Using the fact that the homography induced by the motion of a triplet of body points in two identical pose transitions reduces to the special case of a homology, it uses the equality of two of its eigenvalues as a measure of the similarity of the pose transitions between two subjects, observed by different perspective cameras and from different viewpoints. A view-invariant matching framework utilizes fundamental ratios in [41] that ratio of 2×2 sub-matrix in fundamental matrix is invariant to view changes. These approaches can match two different views of an action, however they assume that tracking has already been performed and trajectories for point triplets are available which is a really strong assumption.

Therefore, the majority of geometrical view-invariance action recognition approaches [36, 37, 38, 39, 40, 41, 42, 43, 44] require detection of individual body parts or salient features and their tracking over a long period of time. All approaches try to exploit constraints on homography or fundamental matrix but its exploitation is suffered due to their initial assumption of available trajectories. An alternative approach is the use of space time interest points [17] which avoids tracking but extraction of such feature points in case of self-occlusion and noise is not robust and suffers from same dilemma as trajectory tracking approaches. In addition, view invariant local space time interest points is still absent in literature as it has been established previously that there exist no invariants for 3D to 2D projection.

In forthcoming chapters , we try to maximize the exploitation of multiple view geometry in devising a view invariant action recognition approach. Inspired from the use of instantaneous flow correspondences to derive flow fundamental matrix [45], we use spatio-temporally consistent optical flow on actor silhouettes to get observation matrix and apply rank constraint to derive action matching score across varying viewpoints. Getting robust extraction of silhouettes in case of occlusion and noise is again a hardship. Therefore, to further remove this pre-processing step, we explore dynamic scene geometry.

Two-view geometry of multiple moving objects is an active research problem. The first generalization of eight point algorithm to multiple motions was not known until recently [33]. The pioneer work which discussed two- body segmentation from two perspective views is presented in [46]. A generalization of this work into multiple moving objects in two perspective views introducing multi-body fundamental matrix is given in [47]. A similar approach for describing the geometry of dynamic scene is presented in [72]. Due to above mentioned work, the concept of multi-body segmentation and multi-body structure from motion has established. Our objective is somewhat different as we want to use the properties of multi-body fundamental matrix for devising a dynamic scene geometry based representation for achieving view invariant action recognition.

2.3 Template based Approaches

The second important visual cue addressed in this thesis is the use of space time pattern templates. In this section, we review action recognition approaches based on space time templates and frequency domain filtering.

The application of space-time pattern templates is a successful action recognition approach. Temporal template matching emerged as an early solution to the problem of action recognition, and a gamut of approaches which fall under this general denomination has been proposed over the years. Early advocates of temporal matching based approaches, such as Polana and Nelson [48], developed methods for recognizing human motions by obtaining spatio-temporal templates of motion and periodicity features from a set of optical flow frames. These templates were then used to match the test samples with the reference motion templates of known activities. Essa and Pentland [86] generated spatio-temporal templates based on optical flow energy functions to recognize facial action units. Efros et al. [102] proposed an approach to recognizing human actions at low resolutions which consisted of a motion descriptor based on smoothed and aggregated optical flow measurements over a spatio-temporal volume centered on a moving figure. This spatial arrangement of blurred channels of optical flow vectors is treated as a template to be matched via a spatio-temporal cross correlation against a database of labeled example actions.

In order to avoid explicit computation of optical flow, a number of template-based methods attempt to capture the underlying motion similarity amongst instances of a given action class in a non-explicit manner. Shechtman and Irani [49] avoid explicit flow computations by employing a rank-based constraint directly on the intensity information of spatio-temporal cuboids to enforce consistency between a template and a target. Given one example of an action, spatio-temporal patches are correlated against a testing video sequence. Detections are considered to be those locations in space-time which produce the most motion consistent alignments.

Representative work includes temporal matching of periodicity information from a set of optical flow frames [48], a two component temporal template of motion energy image (MEI) and motion history image (MHI) [13], space-time shapes induced by the silhouettes in the space-time volume [49] and space time behavior based correlation [26]. Bobick et al [13] computed Hu moments of motion energy images and motion-history images to

create action templates based on a set of training examples which were represented by the mean and covariance matrix of the moments. Action Recognition was performed using the Mahalanobis distance between the moment description of the input and each of the known actions. However, majority of these template based approaches suffer from high computational overhead due to template matching.

Given a collection of labeled action sequences, a disadvantage of these methods is their inability to generalize from a collection of examples and create a single template which captures the intra-class variability of an action. Effective solutions need to be able to capture the variability associated with different execution rates and the anthropometric characteristics associated with individual actors. Recent popular methods which employ machine learning techniques such as SVMs and AdaBoost, provide one possibility for incorporating the information contained in a set of training examples.

To overcome the problems faced by these template based methods, the utilization of correlation filters is investigated for recognizing action instances with promising results. The representative work in this regard is the development of Action MACH [29, 121] that has generalized traditional 2D *Maximum Average Correlation Height* (MACH) filter to 3D MACH by including temporal dimension. However, the major gain is in terms of low computational cost as response of the filter can be analyzed in frequency domain. Despite its success, some researchers [50] have indicated inherent discrepancies in MACH filters and questioned their effective utilization for action recognition. One of the weaknesses of MACH filters is their ineffectiveness to encapsulate inter-class variability. Therefore, these filters are trained only for one class at a time and separate MACH filters are needed for each action class. Secondly, MACH filters overemphasize average training sample, a biased treatment of low frequency components and behave like average filter and may loose finer details of the training set. They emphasize high energy (low frequency) components and attenuate low energy (high frequency) components of the training set leading to poor intra-class discrimination. Thirdly, as action datasets are normally misaligned in space and time, they create problems in learning and testing as synthesized filters are not shift-invariant. Finally, action recognition frameworks based on these correlation filters are not view-invariant that is very desirable aspect for unconstraint action recognition.

In forthcoming chapters, we address above mentioned weaknesses and propose a new view-invariant action recognition approach based on our extended *space-time distance classifier correlation filter* (Action DCCF filter) for view invariant action recognition. Our approach offers following advantages: (i) It provides view-invariance, (ii) Action DCCF filter successfully captures inter-class variability and avoids overemphasize on average training sample by empirically setting contributions of low as well as high frequency information. (iii) It presents a different interpretation of correlation filters as method of applying a *spatio-temporal transformation* to the data, restricted to being Toeplitz ensuring *shift invariance.* It measures similarity between an ideal transformed reference and testing action using a shift-invariant mean square distance measure handling misalignments and (iv) another benefit is that resulting decision boundaries are quadratic which are more 'selective' for choosing feature space portions for assigning to various classes and utilize entire correlation plane rather than emphasizing only single point like correlation peak. These advantages of VIEW DCCF filter can potentially improve performance of view-invariant action recognition.

2.4 Context based Approaches

Visual cues from the contextual background is the last visual cue addresses in this thesis. For this purpose, a case study of night vision vision data is used. In this section, we review contextual action recognition and context enhancement at night time.

Contextual information is important for interpreting human actions especially when actions exhibit interactive relationship with their context. Contextual clues become even more crucial when videos are captured in unfavorable conditions like extreme low light nighttime scenarios. These conditions encourage the use of multi-senor imagery and context enhancement.

The need for understanding actions in context is discussed by different researchers. Scene context is used for event recognition by [54] but it was only applied to static images. Recognizing actions in context is discussed by [53] which was formulated on bag-of-features framework and scene-action SVM based classifier. It is focused on annotated actions in movies and uses script mining for visual learning. A similar approach [55] captures generic object based context by detectors and their descriptors are used as input for supervised learning.

More recently, modeling of scene and object context is discussed by [56] for Hollywood2 action datatset. All above approaches target action recognition in high-resolution action videos in movies. One typical benefit available to these approaches is the ease of finding visual interest points and detectors related to actors and their context.

We present actions in night vision scenario which offers real challenges due to extreme low light conditions. None of the above approaches discuss nighttime visual context and recognition of actions at nigh-time. Most recently, human action activity recognition is discussed in [57, 58] which focuses on recognition in infra-red spectrum. However, these approaches ignore action contexts which are not properly captured by infra-red senors and can not be categorized as contextual action recognition approaches.

We argue that contextual action recognition is not possible using single sensor platform due to the limitations of individual sensor to grab all available visual information about the scene. This situation motivates the use of multiple sensors for context enhancement.

Contextual Enhancement: The operational requirement to fuse night vision imagery is due to the limitations of individual sensor to grab all available visual information about the scene [59]. A common multi-sensor night vision system uses infrared images in case of forward looking infrared cameras and low light images in case of low light visible cameras. The infrared images are maps of infra-red radiation emission which is partly governed by the temperature of the objects. Therefore, such sensors prove good for perceiving hot targets in a busy background, seeing through fog, and monitoring paths through a cluttered forest. However, they are not much effective during thermal crossover periods at night or after long periods of rain and capturing scenery such as trees, leaves and grass in natural scene. On the other hand, low light visible cameras are able to capture surrounding environment but mostly fail to capture specific targets especially hot bodies like a person in camouflage. In addition, even in case when targets are not hiding, low light conditions make their observation obscure.

To solve this problem, image fusion is used which extracts meaningful information from complementary sensor images and combines visual information into a single output image. Over the years, several image fusion techniques are developed which vary in their complexity, robustness and quality. One major trend in image fusion research is to sacrifice complexity to gain quality. However, oppose to images, complexity criterion has more significance in video domain which is intended for real time use. Therefore, video applications do not encourage algorithmic complexity and require simple and efficient information fusion. Furthermore, to meet real time surveillance needs video representation is necessary which gives complete spatio-temporal visual information compared to limited spatial information presented by still images. The video fusion is a process of visual information integration from a number of registered video sequences without loss of information and introduction of distortion. The goal of video fusion is to create a single enhanced video

sequence from complementary video inputs that is more suitable for the purpose of human visual perception, object detection and target recognition.

Color is another important requirement in addition to fusion but colorization of fused grayscale imagery is a daunting task. Most recently, various manual and semi-automatic colorization techniques have been reported in the literature to solve this problem. A highly cited work is colorization based on optimization [60] which needs user defined color scribbling. It proves to be an attractive method which based on the idea that neighboring pixels in space-time with similar intensities should have similar colors and requires neither precise image segmentation, nor accurate region tracking. However, one shortcoming of this method lies in the requirement that input images are annotated with user defined color scribbles and thus lacks full automation. Another popular work is colorization based on color transfer [61] using statistical analysis to impose color characteristics from source color image to another image. It uses a de-correlated color space $\ell\alpha\beta$ and swatches for color transfer from target color image. This technique has the same drawback as it requires manual selection of a color target image and swatches. In addition, color space conversions and swatches make additional burden in terms of complexity.

Despite these shortcomings, above approaches have transformed the cumbersome work of manual colorization into semi-automatic colorization. Due to their successful application in colorization and color correction, these techniques are extended for colorizing night vision imagery [62, 181] presenting a software based approach to night vision offering a cheaper and reliable solution. Therefore, it is highly desirable that fully automated colorization should be introduced to facilitate real-time video processing for night vision applications.

Image fusion is a well established research area within the domain of digital image processing. The past decade gave rise to a considerable number of different approaches to multi-sensor image fusion which vary in their complexity, robustness and quality. In general, depending on the level at which fusion takes place, image fusion can be divided into three categories known as pixel level, feature level and decision level fusion [59]. We focus on pixel level fusion. The most reliable framework for pixel level image fusion is multi-resolution based fusion [155] and the most popular approach for image fusion is wavelet based fusion [156]. But majority of image fusion approaches deal with grayscale images with fused grayscale output and there are few approaches which deal with colorization of fused imagery as well. Our work is related to approaches which include fusion as well as grayscale colorization in their framework. Therefore, in this section we only mention that work which is closely related to our approach.

In absence of original color sensors, colors are transferred to grayscale images following any of these approaches: (1) manual colorization [60, 157],(2) false-colorization [158, 159] and (3) color transfer [62, 181, 182, 160]. Manual colorization [60, 157] gives promising results but it is time consuming due to intensive manual intervention. In addition, these approaches do not deal with multisensory fusion. Therefore, fusion and colorization is possible in separate steps. False colorization approaches decrease manual work but provide noisy and false colorization far from natural day-like appearance. Color-transfer based approaches are advantageous as they result in near natural colors as compared to false colorization techniques.

Color transfer is introduced by Reinhard et al.[61] for transferring colors between two color images using a color space $\ell\alpha\beta$ based on correlation minimization of three color coordinate axes. Originaly it did not deal grayscale image colorization as only objective was the color correction. Welsh et al.[62] extended this method for transferring colors to grayscale images. On similar footing, Toet [181] employed color transfer technique for colorizing grayscale intensified nighttime imagery. Both approaches work in $\ell\alpha\beta$ color space involving multiple intermediate color space conversions. Wang et al.[182] presented

a color fusion algorithm based on color transfer in YUV space and showed that it is less computation intensive. In a similar way, Li and Wang [160] presented color transfer and fusion algorithm based on a linear IUV color space to overcome the harmfulness of the logarithmic transformation to image fusion introduced during the conversion between RGB and $\ell\alpha\beta$ transformation but this approach is computationally expensive due to the use of wavelet transform making it less appropriate for extension to real time video domain. All above approaches deal color transfer using different color space conversions which contribute to their complexity. In addition, they are restricted to work in image domain and do not deal with video processing and its complexity.

Literature review additionally indicates that a fundamental lack of automation is always present in all above techniques because they never define the selection criteria for color source image in color transfer step which mainly defines the quality of final fused image. All approaches use arbitrary target color image based on mere subjective visual perception. To deal with this shortcoming, we propose automated selection of source color image based on global scene context features and colorfulness. Different color image retrieval methods can be considered as alternatives but they are usually complex due to object level details and not suitable in real time video surveillance applications. We rather focus on main theme or context of images.

Visual context recognition has a long history of research with earlier work in cognitive science. In computer vision, context has been used for understanding of static scenes. There are two approaches possible for scene recognition. Traditional conceptions in computational vision have portrayed scene recognition as a progressive reconstruction of input from local image characteristics (edges, surfaces), successively integrated into decision layer of increased complexity. A new paradigm adopted by [161] suggests the recognition of real world scenes from the encoding of the global configuration, ignoring most of the detail and object information. The later approach has the advantage that it does not need any segmentation of the scene nor object and region detection. Conditional to the scene category specification, objects in the scene are independent and context can be defined in terms of overall scene category. Therefore, our source image retrieval approach is built upon the idea that scene can be categorized without going into detail and decomposing them into objects. It uses global scene representations for inferring the main context or theme of the image [162].

We try to address above mentioned shortcomings by proposing an automated night vision system, $SCENT$ (system for color exploitation at nighttime). The major contributions are as follows: It presents a fully automated color night vision system by fusing and colorizing grayscale videos simultaneously. It uses color morphing in RGB color space without any color space conversion. It additionally introduces a simplistic retrieval mechanism for source color image based on global scene characteristics like context and colorfulness. To develop better image fusion algorithms, reliable fusion quality assessment is crucial. Therefore, quality evaluation of image fusion algorithms is widely investigated. Various quality measures have been proposed for multi-sensor grayscale image fusion techniques; but no appropriate quality measure has been devised for the objective quality evaluation of multi-sensor color image fusion. We propose a novel color image fusion quality measure, Color Fusion Objective Index (CFOI) which encapsulates the powers of color image quality, image colorfulness and fused information index. In addition, it evaluates the gradient structure preservation in color fused image. An application of information fusion in nighttime imagery is utilized for experimentation. Experimental results show that CFOI captures fusion of all three factors, important for color image fusion (colors, high frequency edges and low frequency common information). In particular; it is not biased toward any of them as previous standards like Petrovic [164] and IQI [165]. In addition, it deals with colors which were not consideration of previous image fusion quality metrics.

2.5 Conclusions

In this chapter, we reviewed computer vision literature and presented a survey of previous approaches about human action analysis and recognition. We arranged our discussion according to the contributions claimed in this research work. We pointed out different drawbacks and weaknesses of previous works and briefly mentioned how we are going to address these shortcomings in this thesis. Next chapter is our first contribution chapter which discusses feature based approaches to recognize actions across different viewpoints. It targets temporal ordering of local space-time features and proposes two incremental solutions to recognize actions in video sequences.

Chapter 3

Action Analysis using Space-time Features

Good order is the foundation of all good things.

~ Edmund Burke (1729 - 1797)

The research work presented in this chapter has been published as:

1. Anwaar-ul-haq, I. Gondal and M. Murshed, *On Temporal Order Invariance for view-invariant action matching*, IEEE Transactions on Circuits and Systems for video technology, Vol (22), DOI 10.1109/TCSVT.2012.2203213, 2012.

Visual recognition of human actions is a complex phenomenon due to non-linear dynamics of actions, anthropometric variations and strong dependency on camera viewpoint. In this chapter, we explore how non-linear action dynamics can be represented by local space-time features (also known as spatio-temporal or 3D features). We discuss how bundling and ordering constraints can enhance distinctive characteristics of local space-time features for increasing action recognition performance. In addition, we investigate effects of viewpoint variations on action representation which greatly affects action recognition performance. It is due to the fact that action scene captured from different viewpoints contains different representations of same action posing a high-level challenge to computer vision (Figure 3.1). Action recognition approaches which counter the effects of view variations and recognize actions despite viewpoint changes are referred as view-invariant action recognition approaches. In this chapter, we discuss that temporal order of action instances has profound effect on action representations and introduce the notion of temporal order invariance by exploitation of temporal order of local spatio-temporal features.

Local spatio-temporal features are fast becoming representation of choice for action recognition due to their computational simplicity, robustness to occlusion and minor viewpoint changes, elimination of low level object detection and tracking, and existence of scalable matching schemes. These space time interest points are discriminative as they are computed in a way that they capture not only the pixel intensity information but also the motion information (e.g. statistics of optical flow) in the vicinity of the interest

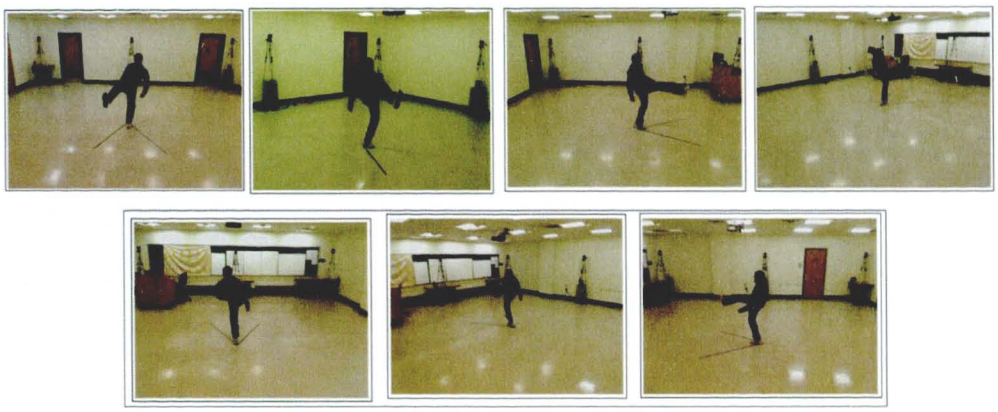


Figure 3.1: An illustration of same action (Kicking) from seven different cameras with different viewpoints by same actor. It shows how strongly viewpoints variations effect the description of an action.

points. After feature extraction from action sequences, representation framework like bag of visual words [115, 123, 30] is utilized for action matching. While similarity/dissimlarity of these interest points is sought during matching same or different actions, their mutual relationships like space time organization and ordering is ignored.

Considering the fact that an 'action' is essentially a spatio-temporal construct, ignoring temporal order in which spatio-temporal features occur can affect matching performance drastically, especially where various actions have many overlapping low level features. In addition, the discriminative power of spatial relationships between these low level feature play an important role when dealing with actions that closely mimic each other e.g. jogging and running.

To overcome these shortcomings several competing approaches have been proposed recently. For instance, naive ordering is imposed by dividing the space-time volume of a video into space-time bins much like spatial pyramid matching used for images [108, 53, 109, 123]. However, as opposed to images, content of a video can vary drastically depending on how input video is segmented (along temporal scale), and location and speed of action, thus making such rigid binning scheme hard to generalize. Binning neighborhoods of interest points at various scales for descriptor computation is another approach [126, 127]. However, this approach also suffers from the rigidity imposed by fixed scales at which computations are performed. Schemes for large scale image retrieval [124, 125] present similar inspiring ideas like bundling spatial features and geometric verification based on area ratio of triangle generated by two visual words. Again, these approaches are restricted to spatial domain as their application goal is entirely different from action recognition which focus more on temporal ordering. In addition, triangle based spatial ordering is difficult to visualized in video domain. Recently, Kovashka et al.[110] proposed an feature centric approach where each feature maintains orientation and location information of neighboring features at various scales in spirit of shape context feature [112]. Scales are computed in such a way that makes the local groupings of feature discriminative in terms of action specific distance metrics.

However, we believe that, to mitigate issues of feature discrimination and spatio-temporal ordering, feature group should be a construct grounded in or associated with the object (or space-time segment) inducing the action as oppose to isolated feature points. This will have the following benefits:

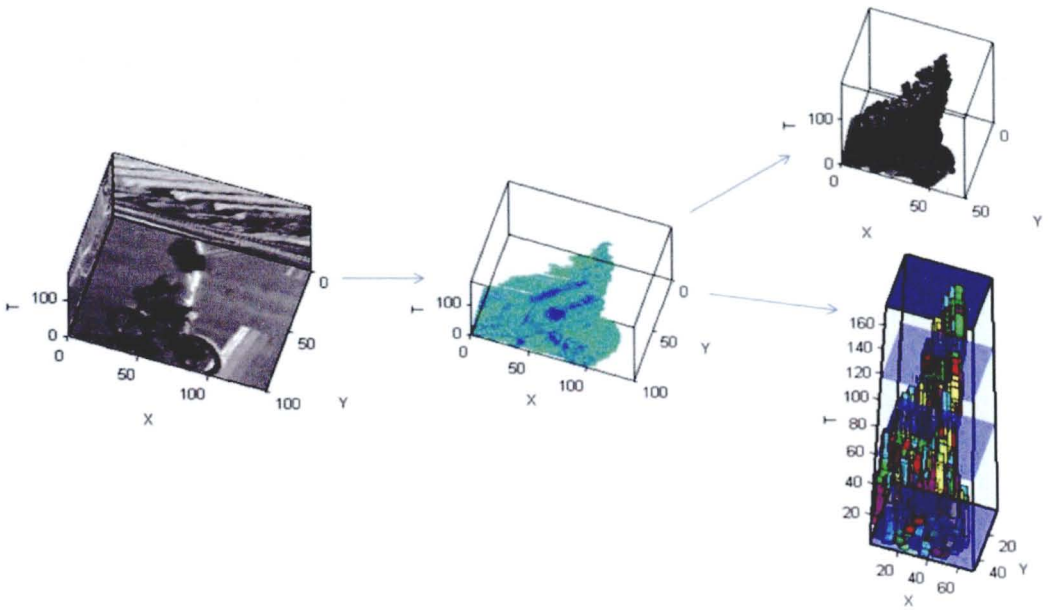


Figure 3.2: Left to Right: Action volumes of v-cycling from You Tube data set, the respective maximally stable volume (MSV) and two different views of STOP features which encapsulate $spatio-temporal$ cuboids inside $MSVs$.

- No need to search across scales for locating discriminative groupings. They will always be associated with the space-time segment that implicitly captures the shape and motion of the object performing an action,
- Ordering will have natural explanation in terms of subunits of an action. For instance, hand going up and down for hand waving action,
- a multi-feature representation that fuses space-time regions with local interest points.

To address these limitations and complications, we propose a novel spatio-temporal action matching framework based on discriminative combination of 3D features, named spatio-temporal ordered packets (STOPs), that combines space-time features along with their geometric ordering information into spatio-temporal volumes avoiding complex ordering constraints. Packaging features into volumes ensures that discriminative power of matching is enhanced as whole packets are matched across videos instead of individual features. At the same time matching is made much robust by developing simple matching criteria that take into consideration spatio-temporal order of features within each volumetric packet.

Figure 3.2 provides an illustration of volumetric packets for cycling from Hollywood data set [123]. Volume computation is carried out by computing maximally stable volumes [16], while cuboid features [18] within these volumes are used as local space-time features. We also propose an indexing scheme for fast matching of videos that extends the bag-of-word model to bag-of- volumetric packets and demonstrate that this representation is much more discriminative. Extensive experimentation is performed on four challenging action data sets (Weizmann [49], KTH [27], Hollywood [123], and UCF YouTube [30]), covering both constrained and unconstrained setting. Recognition performance is comparable to or exceeds existing action recognition approaches. We also demonstrate robustness against various parameters including noise and partial occlusion.

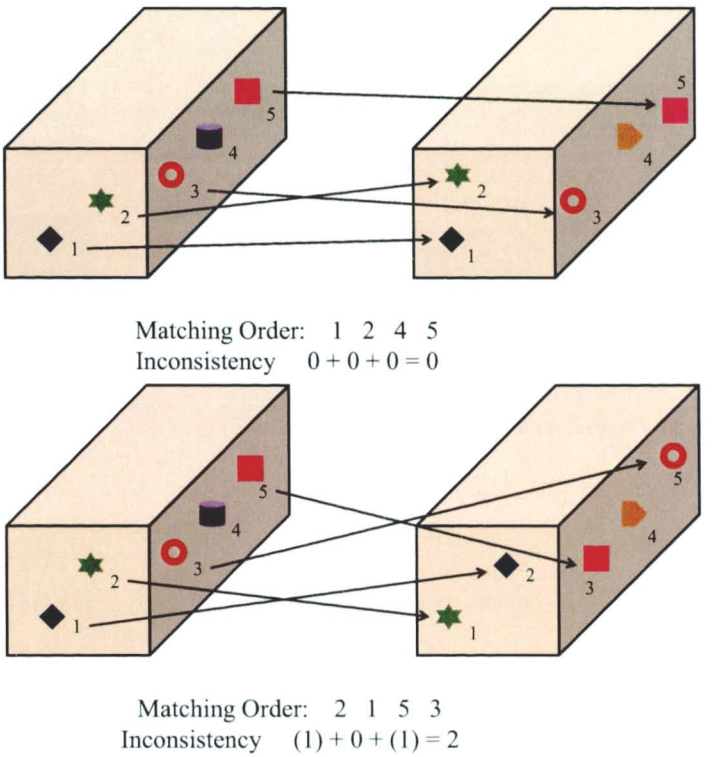


Figure 3.3: The spatio-temporal ordering constraint and Indexing for volumetric packets. (Above) Consistent relative ordering between matched spatio-temporal features (Below) Inconsistent ordering between spatio-temporal features.

3.1 Spatio-temporal Ordered Packets (STOPs)

The key motivation of spatio-temporal ordered packets (STOPs) is to enhance discriminative power of local spatio-temporal features by combining them with sub-video volumes representing shape and motion of the object(s) present in a video. In simple terms, STOP is a local group of spatio-temporal features within an arbitrary shaped sub-volume of a video.

For computing volumetric STOPs, we use a combination of spatio-temporal cuboid features [18], grouped together inside a maximally stable volume [16].

Formally, let $C = \{c_i\}$ be the i-th spatio-temporal cuboid feature and $V = \{v_j\}$ be the j-th volume extracted from a video, then a volumetric packet feature p_j is defined as: $p_j = \{c_i | c_i \in v_j, c_i \in C\}$, where $c_i \in v_j$ denotes spatio-temporal cuboid feature, c_i, that falls inside the maximally stable volume v_j. A set of all STOPs extracted from a video is represented by $P = \{p_j\}$. Note that a cuboid feature may belong to multiple packets as underlying maximally stable volumes may overlap each other.

Next we describe in detail how volumes and features are computed and how features are encapsulated within a volume. To avoid any confusion in the use of isolated term 'feature', we will refer to maximally stable volumes as 'envelope feature' and cuboids as 'enclosed features' for rest of the chapter.

3.1.1 Enclosed Features - Spatio Temporal Cuboids

Spatio-temporal cuboid features are used as enclosed features. Spatio-temporal interest points are located by convolving video signal S with a set of separable linear filters [18], and local maxima of the response function is used as interest point location around which

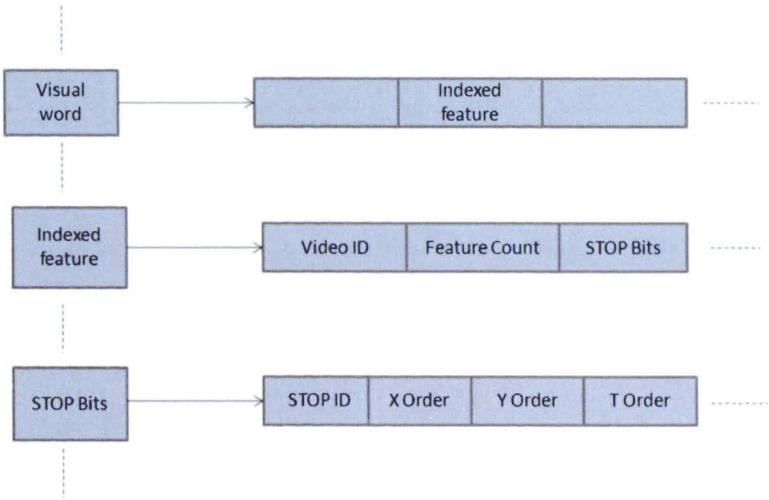


Figure 3.4: Inverted file Index: The structure shows how geometrical order information is stored in this data structure.

descriptors are computed. The response function R_f has the following form :

$$R_f = (S * g * h_{ev})^2 + (S * g * filter_{od})^2, \tag{3.1}$$

$$filter_{ev}(t; r, w) = -cos(2\pi tw)e^{-t^2/\tau^2}, \tag{3.2}$$

$$filter_{od}(t; r, w) = -sin(2\pi tw)e^{-t^2/\tau^2}, \tag{3.3}$$

where $g(x, y; \sigma)$ is a 2D Gaussian smoothing kernel applied along spatial dimensions, and $filter_{ev}$ and $filter_{od}$ are the quadrature pair of 1D Gabor filter applied temporally. Response function R_f has two important parameters, σ and τ, which correspond to spatial and temporal scales of the detector, respectively.

These spatio-temporal interest points effectively discern local variations in intensities under periodic frequency components, respectively. In addition, they capture other significant variations and motions like spatio-temporal corners.

Descriptors: After interest point detection, descriptors are computed for cuboid patches centered at detected interest points (x, y, t, σ, τ). The spatial size, $\Delta_x(\sigma)$, $\Delta_y(\sigma)$, of the cuboids are a function of σ while the temporal size, $\Delta_t(\tau)$, is a function of τ. For each interest point, a cuboid is extracted which contains the spatio-temporally windowed pixel values. These cuboids have a side length of approximately six times the scale at which they are detected. There are different possible ways to define these descriptors: The simplest one requires flattening the cuboid into a vector, although the resulting vector is potentially sensitive to small cuboid perturbations. The second method involves histogramming the values in the cuboid. Such a representation is robust to perturbations but casts away all positional information. Local histograms, like 2D SIFT descriptor [100], offer a reasonable compromise. We have utilized this approach. The cuboid is divided into a number of regions and a local histogram is created for each region. The objective is to achieve robustness to small perturbations while retaining some positional information. Finally, we compute a low dimensional representation of the descriptor using locality preserving projection (LPP)[134]. The final descriptor has 100 dimensions that are sufficient to comprise discrimination information. LPP preserves locality information and better preserves the discriminative power of descriptors.

3.1.2 Envelope Features - Maximally Stable Volumes

The goal of maximally stable volume (MSV) detection is to compute stable connected region volumes from an input video sequence. A MSV volume is a connected component in (x,y,t) space which has homogeneous intensity distribution inside and high intensity difference at its boundary. MSV volume is particularly attractive for the current problem as their computation does not require any elaborate contour tracking or object detection algorithms. They can handle topological changes in the region shape due to articulated body or camera motion. By performing computation at various scales, both fine and large scale structures can be extracted. In addition, they can be computed automatically in real time with low computational complexity.

We used MSV detection by interpreting the input video as connected weighted graph, where video voxels are taken as nodes and edges are relations between voxels with 6, 18 or 26 neighborhood. A data-structure named component tree is built. Each node of the component tree contains a volume. The tree structure allows calculating a stability value for every node analyzing the change in size of the volume while moving the component tree upwards. The most stable volumes, i.e. the nodes with the highest stability values are returned as the detection result. The stability criterion ρ is defined by the area variation as:

$$\rho(V_i^T;\Delta) = \frac{|V_{j-\Delta}^T| - |V_{k+\Delta}^T|}{|V_i^T|}, \tag{3.4}$$

where V_i^T is connected volume obtained by thresholding intensity value at intensity T. Here operator $|.|$, is the cardinality and Δ is the importance parameter which sets the threshold range and the number of component tree levels for computation of stability. Higher values means more stable MSV. Higher values also result in detection of fewer volumes.

3.2 Bag-of-Volumetric Packets

In this section, we describe the indexing and matching scheme that exploits weak spatio-temporal geometric consistency while matching volumetric packets. We call it bag of volumetric packets. We start by quantizing local spatio-temporal descriptors into visual words. Visual vocabulary is learned using hierarchical K-means and Kd-tree is used to quantize descriptors. For each volumetric packet, its enclosed cuboid descriptors are quantized using this vocabulary. For volumetric packet matching, we exploit two weak geometrical constraints. First constraint ensures, a minimum number of common visual words exist between two packets, while the second constraint ensures that spatial and temporal configuration of visual words is consistent. Next, this matching scheme is described in more detail.

3.2.1 Volumetric Packet Matching

Let $Q = q_j$ and $T = t_i$ denote a query and a target volumetric packet with enclosed visual words, $q_j \in W$ and $t_i \in W$, where W is the visual vocabulary. A matching score S between these two packets is defined as;

$$S(q,t) = m(q,t) - \gamma.g(q,t), \tag{3.5}$$

where $m(q,t)$ is the matching score obtained by counting the number of concurring visual words, while $g(q,t)$ is the geometric score obtained by looking at the spatio-temporal

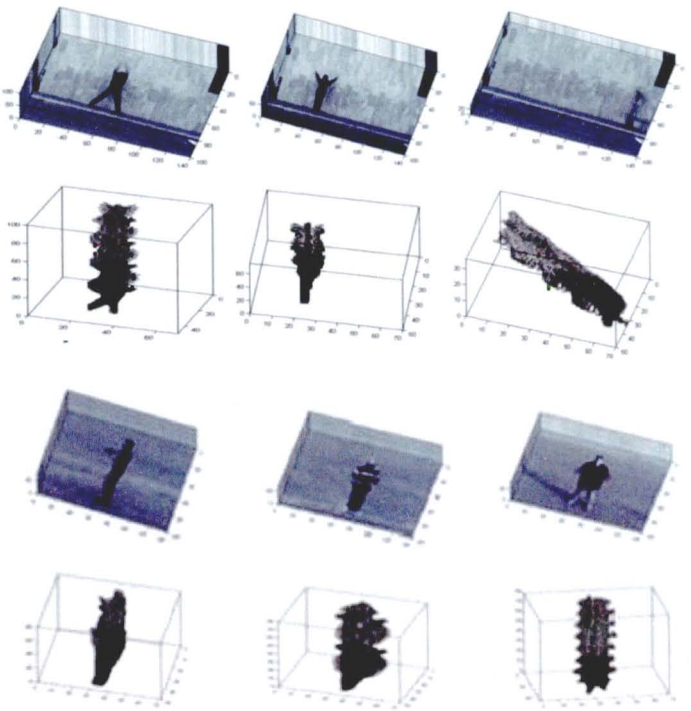


Figure 3.5: Top: Action volumes of jack, wave2 and run from Weizmann dataset and their respective STOP features. Below: Action volumes of boxing, waving and clapping from KTH dataset and their STOP features. (Note that horizontal axis is X, vertical axis is Y and temporal axis is T. and lack of clarity is due to scale of volumes in low resolution.)

consistency of matched visual words. Here, γ is weight parameter which controls the influence of defined constraints on final matching.

Now, we explain how each component of score, S, is computed. Enclosed features in $Q = \{q_j\}$ and $T\{t_i\}$ are sorted according to their x, y and t coordinates, respectively, and any $t_i \in T$ that does not have matching $q_i \in Q$ is discarded. Then first component of S is computed by counting the number of common visual words in two packets:

$$m(q,t) = |\{t_i\}|. \tag{3.6}$$

Geometric component exploits a weak spatio-temporal constraint between query and target packet using relative spatio-temporal ordering information. It penalizes matching score between packets where visual words do not obey the spatio-temporal order. Since, coordinates x, y and t are used for defining the ordering, for every visual word $t_m \in T$ and its corresponding visual word, $q^{'} \in Q$, we denote the geometric location by $L_q[t_i]$ and find inconsistency I as :

$$I^{Dim}(q,t) = \sum_m f(L_q[t_m] > L_q[t_{m+1}]), \tag{3.7}$$

where Dim is any spatio-temporal dimension, x, y or t. Here, f, is the indicator function measuring the consistency or relative ordering in two respective packets. The final geometric score is defined as;

$$g(q,t) = min(I^X(q,t), I^Y(q,t)) + I^T(q,t) \tag{3.8}$$

This geometric score finds the geometrical inconsistency in relative ordering for two packets Q and T. Note that due to importance of temporal consistency, temporal inconsistency term is weighted more than spatial inconsistency terms. Figure 3.3 shows relative ordering used for enforcing weak spatio-temporal geometric constraint and inconsistency computation.

3.2.2 Action Matching

In order to find the best matching action volume for a query action, we use a voting-based framework that utilizes inverted file index for efficient and fast matching.

The structure of the inverted file is shown in figure 3.4. Each visual word has a single entry in the index and points to structure containing information such as video in which it has appeared, appearance count, volumetric packet it belongs to within the video, and its spatial-temporal order. This indexing scheme efficiently stores order information of all STOPs in a video sequence. We use in total 29 "STOP bits" to encode information on STOP features. Of these 29, 5 bits are used to represent "STOP ID", 8 each for "X Order", "Y Order" and "T Order". This allows 32 STOP features per videos with maximum of 256 visual words per STOP.

Each visual word in the query action video casts its vote for a potential matching video. The matched videos are ranked based on the votes and the best ranked video is selected as the correct match. The voting scheme for finding the best matching action is summarized as follows:

Algorithm 1 Given a target database of n videos,$v = 1.....n$ with m visual words, as p_{iv} is visual word i of video v, constituent visual word q_i' of a query video, t_f as term frequency and d_f as document frequency of visual words [114], then the best matching video j in database based on similarity score σ is determined as:

1: Initialize $\sigma_v = 0$ for all dataset videos.
2: For each query visual word and for each visual word in target video j.
3: Update $\sigma_j = \sigma_j + \frac{t_f}{d_f} S(q_i', t_{iv})$ using equation 3.5.
4: Best matching video is the video with the highest value of σ.
5: Assign action label of query to the best matched video.

3.3 Experimental Results and Discussion

A comprehensive set of experiments are performed on four standard human action data sets. The data sets represent actions performed both in constrained and unconstrained settings and represent different set of challenges when it comes to recognizing actions.

3.3.1 Dataset and Experimental Setup

The data sets used for our experimentations include Weizmann [49], KTH [27], Hollywood [123], and UCF YouTube [30]. First two data sets are well known data sets and present controlled experimental settings, and therefore, can be used to benchmark our algorithm against existing algorithms. Hollywood and Youtube data sets are relatively challenging as they contain actions performed in presence of clutter, interacting objects, camera motion and captured from arbitrary viewing angles.

Recognition is performed in leave one out cross validation (LOOCV) setting. Each action video is used as a query once and the best matching video is selected using the voting strategy described above. Action label of the best matching video is assigned to the query video. In order to test the contribution of spatio-temporal geometrical consistency,

each experiment is performed with and without using geometrical constraint. Comparison against existing techniques is performed in terms of recognition accuracy. Furthermore, experiments are carried out to investigate the dependence of our algorithm on various algorithmic parameters (e.g. weight of geometric component, volume stability, noise and occlusion). Next, we describe experimental setup for each data set, present results, and discuss the outcomes. Note that horizontal axis is X, vertical axis is Y and temporal axis is T and lack of clarity is due to scale of volumes in low resolution.

Weizmann Action Dataset: This data set contains 90 low-resolution (180×144, 50fps) video sequences of 10 natural actions performed by 9 different actors. Actions are: run, walk, skip, jumping-jack, jump, pjump, side, wave-two-hands or wave2, wave-one-hand or wave1 and bend. For this dataset, a visual vocabulary of 1080 words is learned and an inverted file index is created as described above. Volumetric packets are extracted by encapsulating visual words inside MSVs. Due to simplicity of the data set and lack of clutter, MSV extraction is quite straight forward. We extracted MSVs with minimum 300 voxels and stability factor of 10. MSVs with less than five enclosed visual words are ignored while packet construction. Figure 3.5 presents some representative MSVs and extracted volumetric packets.

KTH Dataset: This data set contain 600 low resolution (160×120, 25fps) video sequences containing six action categories: walking, running, jogging, boxing, clapping and waving. In total, there are 100 video sequences for each action performed by 25 different actors. Every actor performs each action four times in four different backgrounds. As each video contains repetition of actions, we extract sub-volumes of 80 frames each for stable extraction of MSVs and volumetric packets. The choice of number of frames is motivated by the work of Schindler *et al.* [131] which elaborates on how many frames are sufficient to extract meaningful information from action sequences. For this data set, we learned a visual vocabulary of 2400 visual words and corresponding inverted file index is created. Volumetric packets are extracted by encapsulating visual words inside MSVs. MSVs with minimum of 250 voxels and stability factor 10 are used packet construction. Packets with less than five enclosed visual words are ignored. Figure 3.5 shows some representative action volumes and extracted STOP features for boxing, waving and clapping action. We found that MSVs extraction for this data set is very robust resulting in excellent 3D segmentation of action which helped in obtaining improved results.

Hollywood Data set: This data set consists of realistic and challenging video sequences from 32 Hollywood movies. It contains 8 actions, namely, 'AnswerPhone', 'GetoutofCar', 'HandShake', 'HugPerson', 'Kiss', 'SitDown', 'SitUp' and 'StandUp'. These actions mostly contain interactions of an individual with other individuals or objects. Some of these actions are hard to classify without contextual information. In addition, occlusion and pose variation becomes a significant challenge. We again start by computing MSV. In this case the stability parameter was kept at 5 to generate increase number of volumes per video to cater for multiple moving objects. A visual vocabulary of 2000 words is learned and an inverted file index is created. Following the previously defined protocol, we performed experiments with and without using geometrical constraint for packet matching. Best matching action video is retrieved for each query action and its label is assigned.

Youtube Dataset: This dataset contains complex and challenging video collection from YouTube (resolution 320×240) representing action in unconstrained environment (camera shake, cluttered background, variations in viewpoint, scale and illumination). The dataset contains 1600 video sequence of 11 actions: v-shooting (basketball shooting, 141 videos), v-biking (145), v-diving (156), v-golf (142), v-riding (horse riding, 198), v-juggle (football-juggling, 156), v-swing (137), v-tennis(167), v-jumping (trampoline jumping, 119), v- spiking (116) and v-walk-dog (123). A visual vocabulary of 6400 visual words is learned and an inverted file index is created.

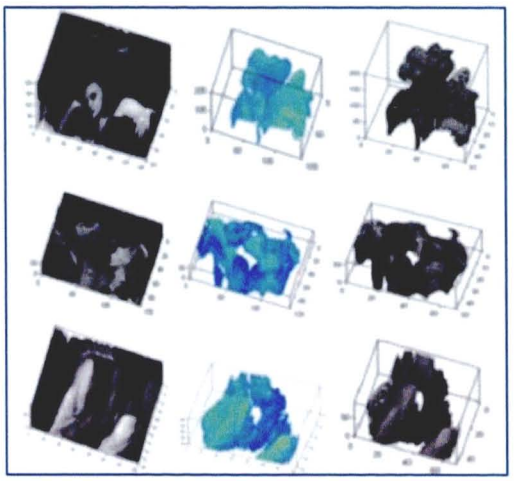


Figure 3.6: (Three Left Columns) Volumetric Packets. First column: Video sequences from Hollywood data set; Middle column: Maximally stable volume corresponding to the sequence; Third column: Volumetric packet with local features encapsulated within the volume.)

3.3.2 Action Recognition Performance

Recognition is performed using LOOCV (leave oneout cross validation) settings (one verses all) and confusion matrices are displayed. First we describe results for Weizmann dataset. First confusion matrix shows results obtained using geometric consistency for packet matching, while the second matrix shows results obtained without using any geometric consistency. We obtained average recognition accuracy of 100% using geometric consistency, while recognition accuracy is 94.44% without it. Note that, in the absence of geometric consistency constraint, there was mix-up in skipping and hand waving actions which we believe is due to the fact that "frequency" information dominates in absence of geometric constraint and causes confusion among these very similar actions. This points to the fact that geometric consistency constraint can play critical role in distinguishing actions that have similar human body motion and action speed. Figure 3.7 and 3.8 present confusion matrices with and without geometric constraint.

Two confusion matrices are displayed. First matrix shows results obtained using geometric consistency for packet matching, while the second matrix shows results obtained without using any geometric consistency. We obtained average recognition accuracy of 95.3% using geometric consistency, while recognition accuracy was 92.16% without it. Average recognition accuracy comparison for KTH dataset with other existing techniques is as follows: Neibles [140]: 83.33%, Dollar[18]: 81.17%, Cao[119]: 95.02%, Liu[87]: 94.01% and our proposed: 95.3% percent. Our results are better than most of the existing approaches. We again observed that geometric constraint helped in removing ambiguities between actions which are very similar in terms of limb motion and speed (e.g. clapping, boxing, or running, jogging etc.). It is important to note that distinguishing these minor variations and getting them right are critical for any action recognition system. From this point of view, getting classification of strikingly different actions (e.g. hand-waving & running) is not interesting and we believe even simple classifier can correctly classify action with such large variations.

The average class precision for each action in Hollywood dataset is shown in figure 3.9. Our mean average precision is 40.4% which is better than the average precision of 38.4% as reported in the original paper [123]. In this experiment, geometric ordering resulted in

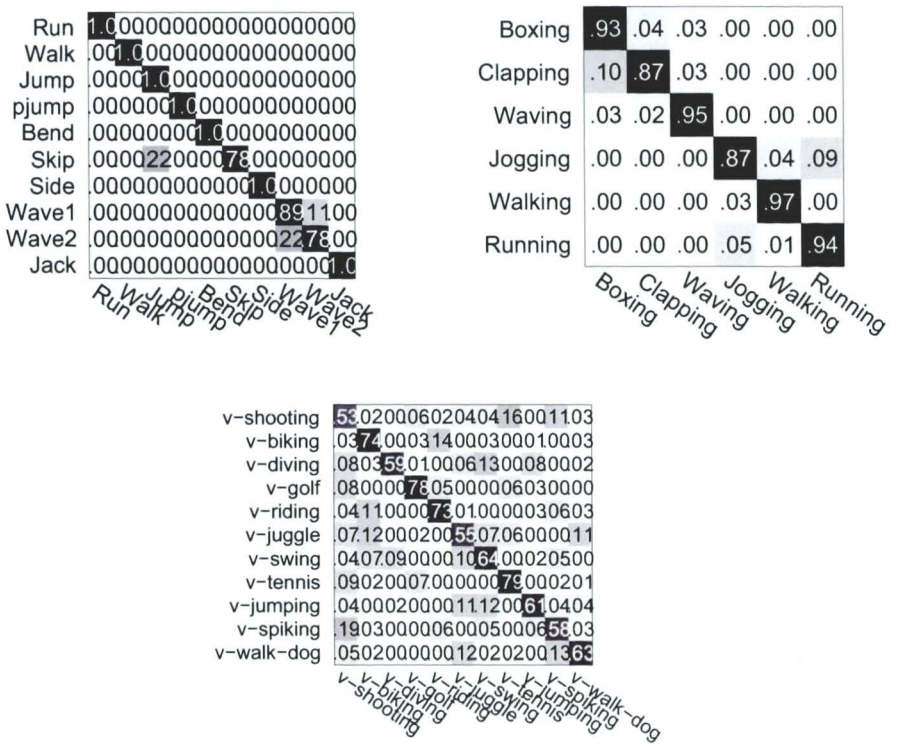


Figure 3.7: (Top Left)Confusion Matrices for Weizmann dataset without using geometric constraint (accuracy 94.44%). (Top Right) Confusion Matrices for KTH data set without using geometric constraint (accuracy 92.16%). (Below) Confusion matrices for YouTube dataset with mean accuracy of 65.6% without geometric constraint.

degradation of average class precision for 'AnswerPhone' and 'Kiss action while improved precision for 'Standup', 'Sitdown', 'Standup', 'HugPerson' and 'HandShake' action are observed.

The last and most realistic dataset is YouTube dataset. On this data set, robust extraction of MSVs was a challenging task. We used a smaller stability value of 5, which increased the number of MSV per video and therefore allows matching to carried out over many packets per video. Again, we conducted experiments with and without using geometrical constraint and computed confusion matrices. We observed that 'v-shooting', 'v-juggling' and 'v-spiking' proved difficult as MSV extraction suffered due to background cluttered environment (e.g. pool, crowd etc.). Another challenge is the presence of jittery motion in some of the videos which requires video stabilization. We used off the shelf VirtualDub Deshaker [120] for removing shaky motion. We obtained mean average accuracy of 66.8% which is better than 64.3% [118] and 65.4% (using only motion features) [30].

3.3.3 Robustness against noise and occlusions

The robustness of the proposed matching scheme is evaluated by introducing noise and partial occlusions to the original video sequences. Introduction of noise and occlusion will effect the computation of MSVs. We wanted to observe how the variations in MSVs effects the matching performance.

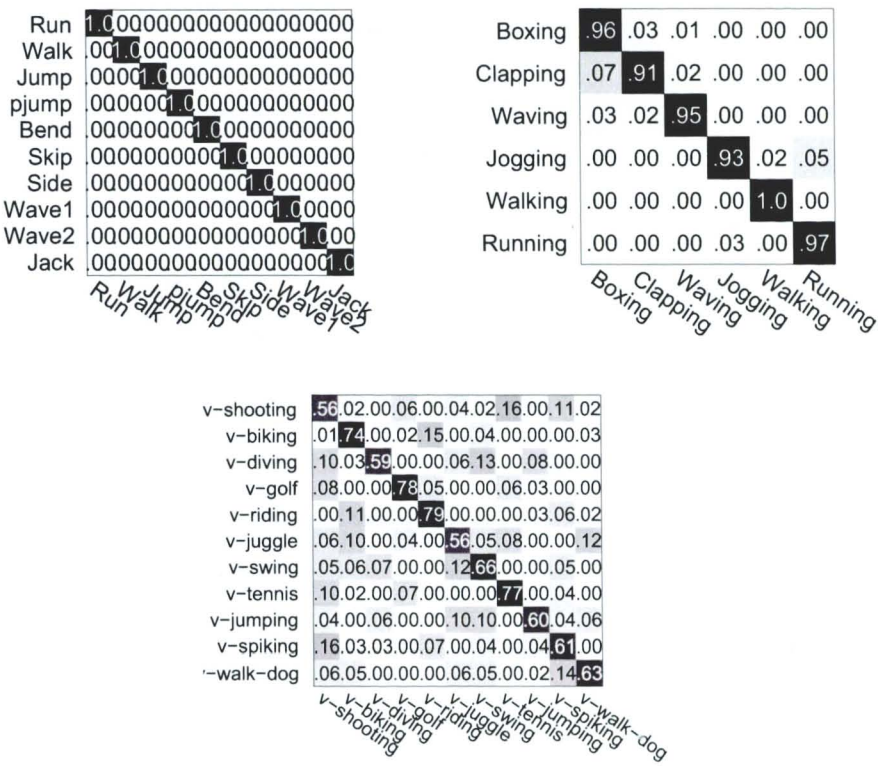


Figure 3.8: ((Top Left) Confusion Matrices for Weizmann dataset with geometric constraint (accuracy 100%). (Top Right) Confusion Matrices for KTH data set with geometric constraint (accuracy 95.3%).(Below) Confusion matrices for YouTube dataset with mean accuracy of 65.6%.

We observed that recognition accuracy remains very high up till noise density levels of 0.25 after which we observed severe degradation in MSV computation. Similarly occlusion, discrimination of volumetric packets decrease in presence of occlusion as MSV starts breaking up into smaller chunks and many local spatio-temporal features get missed. the occlusion is artificial generated creating horizontal and vertical strips (keeping intensity vales equal to zero in this range)according to action video.

Similar to Weizman data set, we tested the robustness of KTH dataset for our algorithm by introducing noise and occlusion. Effects of applied occlusion and noise decreased 9%, 7% percent accuracy for Weizmann and 7%, 5% percent for KTH dataset.

3.3.4 Computation Time

On Intel (R) CoreTM 2 Duo system with 4GB RAM and unoptimized Matlab code, we get average query processing time of 6.2 seconds excluding feature extraction time. The summarized effects of geometric weight on overall accuracy are as follows: The average recognition accuracy of all datasets is 74.9% for weight 2, 75.6% for weight 2.5 and 73.1% for weight 4 which suggests that 2.5 is found suitable empirically.

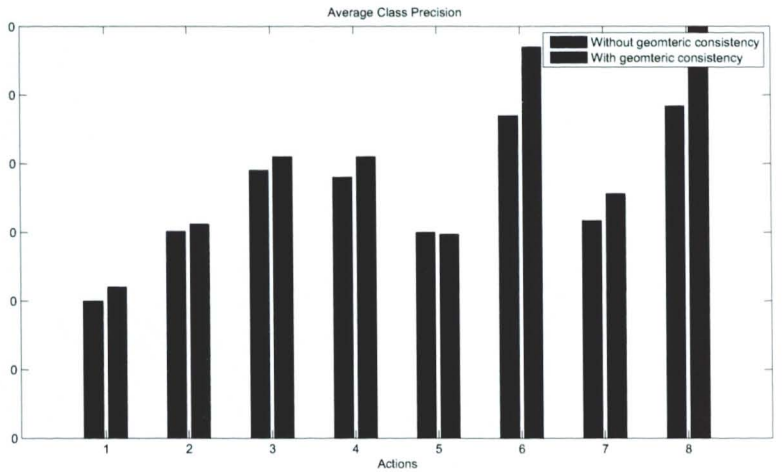


Figure 3.9: Average class precision for 8 actions in Hollywood data set with mean average precision as 36.06% and 40.4% for without and with geometric constraint, respectively. X-axis shows actions 1-8, which include, AnswerPhone, getoutofCar, HandShake, HugPerson, Kiss, SitDown, SitdUp and StandUp actions and Y-axis is the precision for recognition.

3.4 Seeking Temporal Order Invariance for View-invariant Action Recognition

Imagine the sequence of an activity: A standing person bends to pick up a ball, holds it, stands up and throws the ball while five different cameras observe his activity from five different viewpoints. The captured actions differ from each other especially when displayed in digital images due to the differences in viewpoint. This visual difference causes enormous difficulty for recognizing actions in view-invariance sense which motivates us to find an invariant action property unaffected by the variations in viewpoints (see Figure 3.10). However, if we divide the whole activity into constituent actions and note their temporal order or sequence (e.g., 1: bend, 2: pick-up, 3: stand-up and 4: throw), we find a view-invariant property. The temporal order of these actions remains same, no matter from which viewpoint they are captured. No camera can capture 'throw' action before the 'pick-up' action. Similar is the case of individual action instances as constituent action units (e.g., representation of local motion and posture variations) within an action preserve a temporal order irrespective of the camera viewpoints. In this section, we investigate part based action representation and temporal order invariance to devise a view-invariant action recognition approach based on the above conjecture where "view-invariant" action recognition is defined as the visual recognition of actions that is unaffected by viewpoint variations.

Part based action representations: Unit formation is fundamental to visual perception. Part based action representations present unit formation of actions. These approaches can be categorized into two types: first type utilizes visual appearance of parts and geometrical constraints. In contrast, the other type relaxes structural constraints and represents action video sequence as a set of independent features. The latter approach is popular and widely used due to its computational simplicity. It represents action formation in terms of spatio-temporal features [17] that are marked by their elimination of tracking, robustness to occlusion and scalable matching like bag-of-features framework.

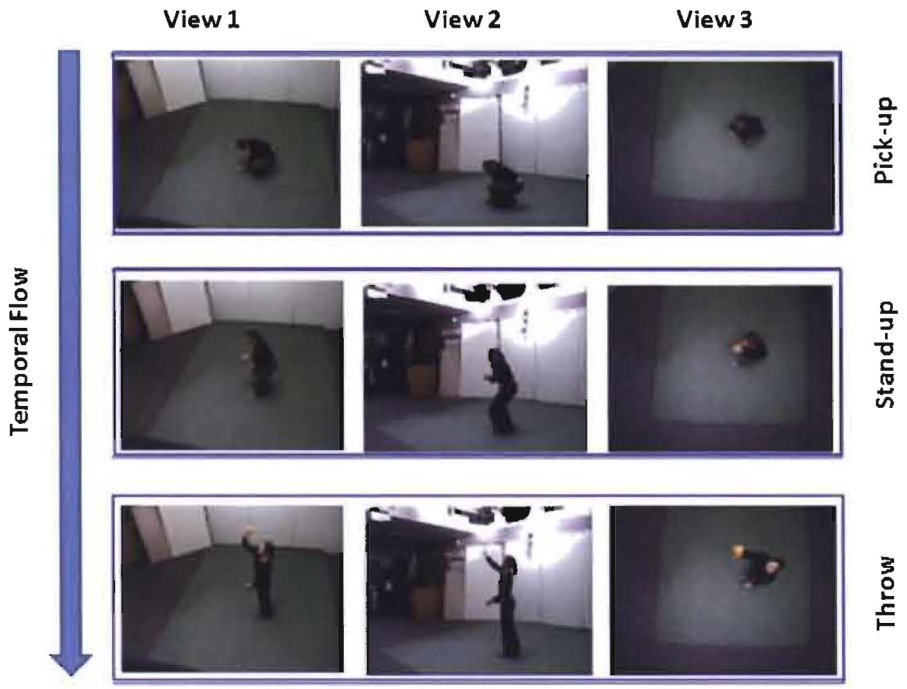


Figure 3.10: Temporal Flow within an Activity: Pick-up, Stand-up and Throw actions from three different cameras with different viewpoints. (Sample frames from IXMAS dataset, only three views are shown). It shows that temporal order of actions remains same with the whole activity irrespective of viewpoint variations.

The major weakness inherent in these approaches is due to the ignorance of mutual relationships of features like space-time organization and temporal ordering in bag-of-features framework. It leads to the study of temporal structure of actions.

Temporal structures of actions: The importance of temporal order is investigated by different researchers. Temporal composition of different motion segments for recognizing human activities is studies by [140]. It adopts an action representation based on spatio-temporal interest points (STIPs) [17] while video sequence is decomposed into temporal segments of various length. Discriminative subsequence mining is proposed in [138] to find optimal discriminative subsequence patterns represented by spatiotemporal cuboid detectors [18]. Finally, visual words arranged into temporal bins are presented for classification. Short subsequences called action snippets of 1-7 frames long are proposed in [131] to alleviate temporal segmentation of actions using form and motion features. Most recently, actom sequence modeling is proposed by [141] that represents temporal structure of actions as a sequence of histogram of actom-anchors visual features using spatio-temporal interest points (STIPs) [17]. However, these actom are manually annotated at training level.

The Contribution: Above mentioned approaches utilize temporal order information for improved action recognition performance. Therefore, their objective is to overcome the weaknesses of bag-of-features framework like spatial bag-of-words [139]. Unfortunately, the success of these approaches are marginal as they are not view-invariant. However, our objective is different from the above approaches as we explore global temporal order within human actions to seek view-invariant action recognition, which has not been investigated by any previous work (to the best of our knowledge). Subsequently, we propose a novel notion of *temporal order invariance* for scalable framework for view-invariant action recognition.

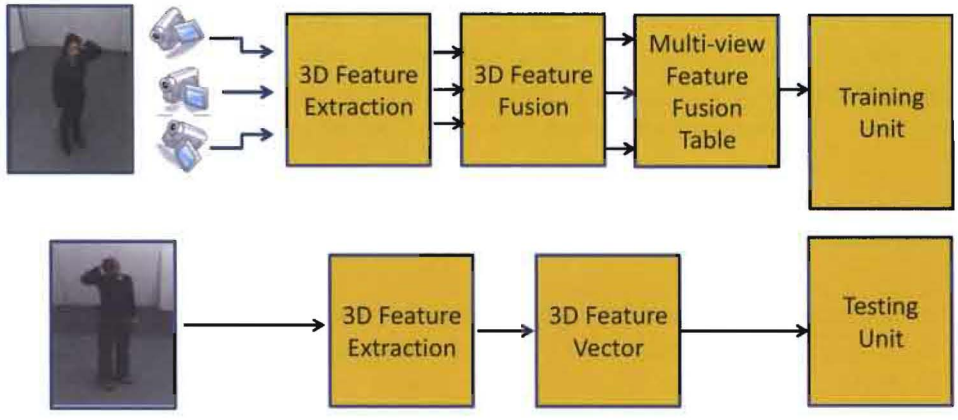


Figure 3.11: The Proposed Framework: (above) Training is performed for all available viewpoints for getting fusion tables for each action class by repeating described steps for each action class in the dataset (sample instance of scratch-head is shown), and (below) Testing sequence for unknown query action video from an arbitrary viewpoint.

In this work, we focus on global analysis of human actions and seek a view-invariant representation. We based our approach on the following conjecture: "*The temporal order of actions elements within an action is invariant to viewpoint variations*". We define action elements in terms of local spatio-temporal interest points and define spatiotemporal order preservation constraint in matching framework. Spatiotemporal cuboid features [18] are taken as space-time interest points as these features are based on maximization of discrimination between behaviors. For each action class, we define a feature fusion table. A feature fusion table is a defined data structure to encapsulate multiple training examples against multiple viewpoints for a single action class. It is achieved through feature fusion based on principal component analysis. A matching score is then calculated based on global temporal order constraint and number of common features. Finally, the action label of the class with maximum value of matching score is assigned to the query action.

3.5 The Proposed Approach

Actions are spatio-temporal patterns which can be characterized by a set of discriminative parts or components. We call these discriminative parts as action elements. These discriminative parts can be detected by spatio-temporal interest points and thus action elements can be represented by small patches around detected interest points based on various measures namely saliency, cornerness, periodicity or motion activity.

Let V be a volume representing a set of consequent input frames, defined on a set of points P where $p = (x, y, t) \in P$ is an individual space-time point or voxel. We intend to find a set of space-time interest points F within this volume and use temporal order information to seek view-invariance. We use spatio-temporal patches around these interest points and build descriptors. To deal with view-invariant action recognition, our proposed approach uses view information fusion with spatiotemporal feature fusion to develop feature fusion tables and enforces geometrical order consistency during matching. The flowchart of our proposed framework is shown in figure 3.11. Training is performed for all actions for all available viewpoints to get fusion tables for each action class. It is archived by repeating described steps (1: spatio-temporal feature extraction, 2: multi-view feature fusion and 3: construction of feature fusion table) for each action class in dataset.

Figure 3.11 shows sample instance of scratch-head only. Testing sequence for unknown query action video from arbitrary viewpoint is used and matching score is calculated for every fusion table and action label of the table with maximum matching score is assigned to the query action. In the following subsections, we give detail of the used space-time features, feature fusion, feature fusion tables and matching framework.

3.5.1 Spatio-temporal Feature Fusion using Principal Component Analysis

For all training video sequences related to the same action captured from the same viewpoint, we use spatio-temporal information fusion. This information fusion is done by fusing spatiotemporal features in training videos using principal component analysis (PCA).

The fusion strategy is simple. An action video sequence contains many spatiotemporal features. (i) We arrange all video sequences of the same action class and the same view into the same group; (ii) We extract cuboid features from video sequences and sort features according to their temporal order; (iii) For all video sequences (same view, same class), we fuse features according to their position in temporal order; (iv) Feature fusion is achieved through PCA. For instance, all features (position 1 in temporal order, 1st feature of all videos) of wave action in view 1 are concatenated into a single feature vector and principal component analysis is used to reduce its dimensionality to a single feature. (v) Finally, fused features for each class are arranged into fusion tables (to be described in the next section).

Suppose we represent a set of training videos as $\mathbf{V}$. For K action classes, $\mathbf{V} = V_1, V_2, \ldots, V_K$. For each of v different views of original datum, we use m spatio-temporal features in a single fusion table and the number of features remains same for all views in a single fusion table. To achieve it, we set m as the minimum number of spatiotemporal features extracted for a training viewpoint. However, value of m varies for each action class as number of spatiotemporal features is different for different actions. Let $f_{i,j}^{k,n}$ denote the j-th feature, by temporal order, of the l-th training video for the i-th viewpoint in the k-th action class for all $1 \leq i \leq v$, $1 \leq j \leq m$, $1 \leq k \leq K$, and $1 \leq n \leq N(k,i)$ where $N(k,i)$ denotes the number of training videos used for viewpoint i in action class k. Features at the same temporal order in all the training videos of the same action class and viewpoint are fused using PCA to obtain a single feature of reduced dimensionality as:

$$F_{i,j}^{k} = PCA(f_{i,j}^{k,1}, \ldots, f_{i,j}^{k,N(k,i)}), \tag{3.9}$$

for all $1 \leq i \leq v$, $1 \leq j \leq m$, $1 \leq k \leq K$.

Spatio-temporal features are key players in this information fusion. However, this fusion framework is independent of the type of spatio-temporal features. We have used cuboid features [18] due to their remarkable success in capturing local variations in action instances as described in previous section.

3.5.2 Multiple View Feature Fusion Tables

The spatio-temporal feature fusion described above is performed separately for each view. We combine information for all views related to an action class. For this purpose, we use spatio-temporal feature fusion table for each action class. A feature fusion table is a defined data structure to encapsulate multiple training examples against multiple viewpoints for a single action class. These tables are kind of feature matrices whose rows are ordered fused space time features related to a view and columns are respective fused feature for different views. We represent collection of training feature fusion table as matrix $\mathbf{T}$. For K action classes, $\mathbf{T} = T_1, T_2, \ldots, T_K$. Each of these matrices have $i = 1, \ldots, v$ rows and

$j = 1, \ldots, m$ columns. Table 3.1 shows an illustration of spatio-temporal feature fusion table.

Table 3.1: The general structure of Feature Fusion Table T_k, $1 \leq k \leq K$. Each of these matrices have $i = 1, \ldots, v$ rows (viewpoints) and $j = 1, \ldots, m$ columns (number of features).

$F^k_{1,1}$	$F^k_{1,2}$	$\cdots$	$\cdots$	$F^k_{1,m}$
$F^k_{2,1}$	$F^k_{2,2}$	$\cdots$	$\cdots$	$F^k_{2,m}$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$F^k_{v,1}$	$F^k_{v,2}$	$\cdots$	$\cdots$	$F^k_{v,m}$

3.5.3 Action Classification

Let $Q = (q_1, \ldots, q_m)$ denote the feature vector from a query action video sequences with m enclosed features. For each fusion table T_k of action class $1 \leq k \leq K$, Q is matched against every row of T_k and a matching score S between them is calculated. The maximum matching score among all the rows of the corresponding feature fusion table is selected as the matching score of the corresponding class. It describes which view in respective multi-view fusion table is closer to the viewpoint of the query action. However, it should be noted that matching score does not correspond to the sum of all rows of a fusion table; but a single row with the maximum score. The class with the maximum matching score is considered as the action class and its label is attached to the query action.

Considering action class k, the matching score $S_{k,i}$ for the i-th viewpoint, obtained from the i-th row $T_{k,i} = (F^k_{i,1}, \ldots, F^k_{i,m})$ of the feature fusion table T_k, is defined as:

$$S_{k,i}(Q, T_{k,i}) = M(Q, T_{k,i}) - \gamma G(Q, T_{k,i}) \tag{3.10}$$

where $M(Q, T_{k,i})$ is the matching score obtained by counting the number of concurring features and $G(Q, T_{k,i})$ is the geometric score obtained by looking at the temporal inconsistency of the matched features. Here, γ is the weight parameter, which controls the influence of the defined constraints on the final matching. A sensitivity analysis on γ is presented in Table 3.3.

Similarity Score Calculation: Here, we explain how each component of score, $S_{k,i}$ is computed. Enclosed features in Q and $T_{k,i}$ are sorted in temporal order according to their t, x, y coordinates, in order. Then, the first component of $S_{k,i}$ in equation 3.7 is computed by counting the number of matching spatio-temporal features:

$$M(Q, T_{k,i}) = \sum_{j=1}^{m} \mathbf{I}_{q_j = F^k_{i,j}} \tag{3.11}$$

where $\mathbf{I}_B$ is the *Boolean-to-integer* conversion function defined as $\mathbf{I}_B = 1$ if B is true; 0 otherwise.

Geometric component exploits a weak temporal constraint between the query and target viewpoint using relative temporal ordering information. It penalizes matching score where corresponding spatio-temporal features do not obey the temporal order. Let vector $L_{Q,T_{k,i}}$ denote the temporal order of matching in $T_{k,i}$ of the matched features of Q, in their temporal sorted order, where $|L_{Q,T_{k,i}}| \leq m$ denotes the number of matched features.

We can then use $L_{Q,T_{k,i}}$ to measure the temporal inconsistency as:

$$I^t(Q,T_{k,i}) = \sum_{l=1}^{|L_{Q,T_{k,i}}|-1} \mathbf{I}_{L_{Q,T_{k,i}}(l)>L_{Q,T_{k,i}}(l+1)} \tag{3.12}$$

where the superscript t is used to identify temporal ordering. The final geometric inconsistency score is defined as

$$G(Q,T_{k,i}) = I^t(Q,T_{k,i}). \tag{3.13}$$

This geometric score finds the geometrical inconsistency in relative ordering of matching spatio-temporal features.

3.6 Temporal Order Invariance: Experimentation

A comprehensive set of experiments is performed on publicly available multi-view action datasets. These are standard multi-view human action datasets and pose significant challenges to action recognition. In next subsections, we give a brief description of the datasets, experimental settings, recognition accuracy, performance comparison and discuss the effects of important parameters.

3.6.1 Multi-view Action Datasets

In this experimentation, we have used Multi-view WVU Action Dataset for initial investigation and used well-known Inria multi-view IXMAS Dataset [15] for extensive validation of our approach.

IXMAS Action Dataset: The Inria Xmas Motion Acquisition Sequences (IXMAS) is widely used dataset for view-invariant action recognition. It contained 11 actions, each performed by 10 actors three times and captured from five different views. We collected 1650 video sequences. These actions include check-watch, cross-arms, scratch-head, sit-down, get-up, turn-around, walk, wave, punch, kick and pick-up. The variations in viewpoints between five different cameras pose significant challenge.

WVU Multiview Action Dataset: The dataset was collected as part of the research work on real-time human action recognition in a camera network. The multi-camera network system consists of 8 cameras that provide completely overlapping coverage of a rectangular region R (about 50 x 50 feet) from different viewing directions. It contained 11 actions, each performed by 10 actors three times and captured from eight different views. These actions include nodding head, clapping, waving 1 hand, waving 2 hands, punching, jogging, jumping jack, kicking, picking, throwing and bowling. This dataset is available from [31].

3.6.2 Experimental Setup

Most of the previous approaches use leave one out cross validation (LOOCV) setting. Therefore, we use the same setting to facilitate performance comparison. Each action video is used as a query once and matched with each multi-view fusion table to calculate a matching score. Every multi-view fusion table has an action label. Action label of the best matching (the highest matching score) fusion table is assigned to the query video. The confusion matrices are displayed to demonstrate the recognition accuracy. Multi-class SVM is used for classification using one-against rest approach.

3.6.3 Performance Comparison

Performance comparison against the existing techniques is performed in terms of recognition accuracy. Recognition accuracy is calculated for each individual camera setting and average recognition accuracy for all cameras is displayed.

Average Recognition Accuracy: The confusion matrix in figure 3.12 shows results for WVU action dataste while figure 3.13 shows results for IXMAS action dataset. This performance is calculated by averaging the performance accuracy of all camera viewpoints of WVU and IXMAS dataset. We achieve 92.04% performance accuracy for WVU and recognition accuracy of 83.5% for IXMAS. This performance for IXMAS is comparable to the existing techniques of [52], [51] and [71] which show average recognition accuracy of 82.8%, 83.4% and 71.2% respectively.

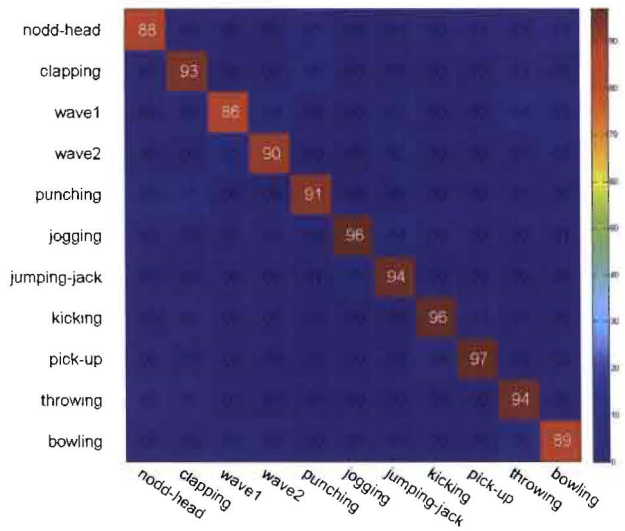


Figure 3.12: Confusion matrix for WVU dataset which shows average recognition accuracy of all viewpoints.

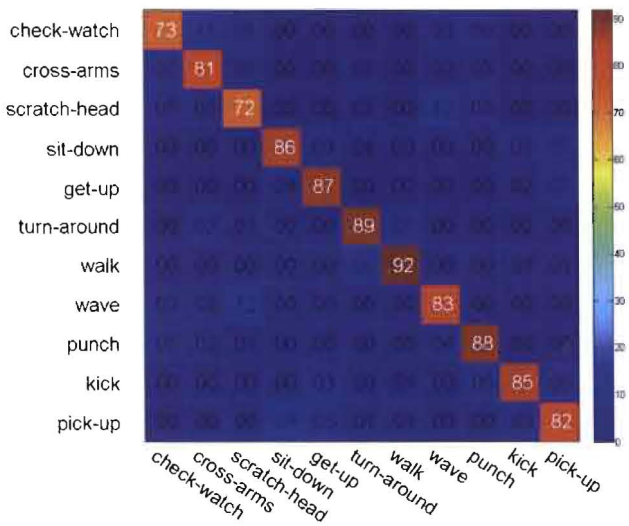


Figure 3.13: The Confusion Matrix for IXMAS Action dataset (with geometric consistency) with recognition accuracy (83.51%).

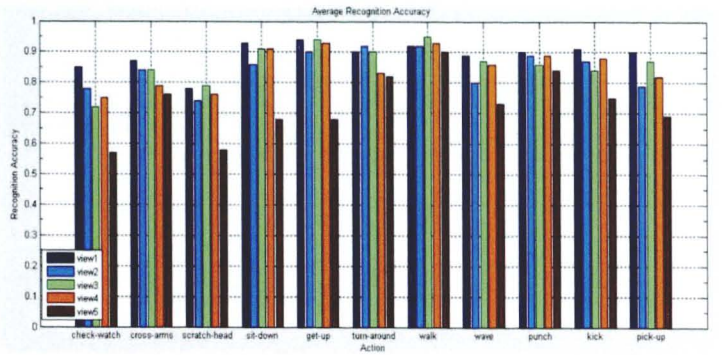


Figure 3.14: Recognition performance of IXMAS dataset from five different cameras with different viewpoints with geometric consistency ON.

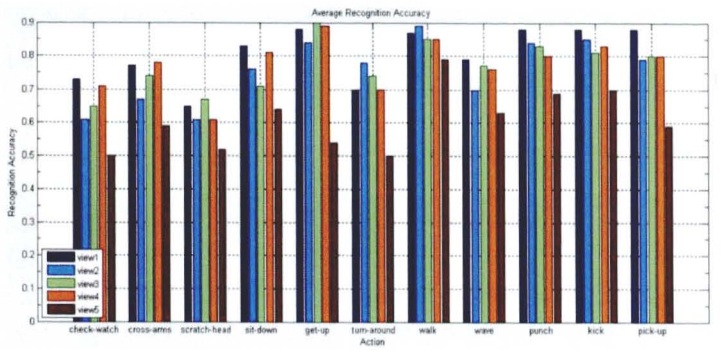


Figure 3.15: Actions instances of IXMAS dataset from five different cameras with different viewpoints with geometric consistency OFF.

Recognition Accuracy vs Camera Viewpoints: In addition to average recognition accuracy, we have shown recognition against individual camera settings and compared it with other exiting techniques. Table 3.2 shows performance comparison with the existing techniques.

Figure 3.14 and 3.15 show recognition performance of individual actions vs five camera viewpoints. Majority of the action classes show higher recognition results except action captured from camera5. One explanation is that this camera setting is exactly above the actor and due to the nature of complex action dynamics, it is difficult to comprehend actions properly by spatio-temporal features.

To effectively test the performance of our approach, we have included further experimentation on multi-view WVU dataset. We used seven views for training and eighth one for testing and displayed our result in fusion table from (T1$\cdots$T11) are trained for seven different views and labeled for respective action classes, nodding head, clapping, waving 1 hand, waving 2 hands, punching, jogging, jumping-jack, kicking, picking, throwing and bowling. It shows that score is maximum for the respective query class, a throwing action for different view. Table 3.2 indicates performance comparison with other competitive approaches.

Figure 3.16: Action matching score for individual Action (Throwing) in WVU dataset. Fusion table from (T1—T11) are trained for seven different views and stand for respective action classes, nodding head, clapping,waving 1 hand,waving 2 hands, punching, jogging, jumping-jack, kicking, picking, throwing and bowling. It shows that score is maximum for the respective query class, a throwing action for different view.

Table 3.2: Performance comparison with the existing techniques. Average recognition is the average performance for all five cameras.

Method	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
Ours	**89.2**	84.7	**86.9**	87.0	69.7	**83.5**
Ref.[52]	86.6	81.1	80.1	83.6	**82.8**	82.8
Ref.[51]	86.7	**89.9**	86.4	**87.6**	66.4	83.4
Ref.[71]	74.8	74.5	74.8	70.6	61.2	71.2

3.6.4 Importance of Geometrical Order Consistency

The most important setting is the matching with order consistency constraint. To know the importance of geometry constraint, we turn-off the geometric consistency. We obtained average recognition accuracy of 74.27% for IXMAS action dataset which shows a degradation of 9.2% when order consistency is OFF. This large degradation shows the importance of order consistency in devising a view-invariant action representation and greatly validates our conjecture that : "*The temporal order of actions units within an action is invariant to viewpoint variations*".

3.6.5 Impact of Important parameters

An important parameter in our approach is the value of γ which effects the contribution of geometrical consistency. The performance increases as γ increases within a small range then for a long range of γ, the performance remains stable and then again shows a degradation in performance. The reason is that as the value of γ increases to a certain range, the order consistency constraint would be overemphasized leading to degradation of performance. Table 3.3 indicates performance against three different values of γ for IXMAS dataset.

The extraction of spatio-temporal cuboid features is an important part of our approach. The spatial size, $\Delta_x(\sigma)$, $\Delta_y(\sigma)$, of the cuboid is a function of σ while the temporal size, $\Delta_t(\tau)$, is a function of τ. We use standard values of $\sigma = 2$ and $\tau = 4$ [18]. In our

Table 3.3: The effect of different values of γ on average recognition accuracy comparison for IXMAS dataset.

Values of γ	Accuracy
1.5	80.2
2.5	83.5
4.5	75.1

experimentation, adequate value of minimum number of features m was found to be 5. Below this value recognition accuracy begins to suffer. At $m = 4$, average accuracy for IXMAS dataset is decreased by 9.8 percent.

3.6.6 Limitations and Average Computation Time

In case of very complex action dynamics and difficult camera viewpoint, number of spatio-temporal features decrease and too few features can effect performance. Therefore, one limitation is the high dependence on spatio-temporal feature extraction. Some viewpoints make their extraction really difficult like camera 5 viewpoint in IXMAS. We have used a consistency constraint that is not affine invariant. Therefore, improvements are possible by devising new feature extraction technique and consistency constraint.

On Intel (R) CoreTM 2 Duo system with 4GB RAM and un-optimized Matlab code, we get average run-time of 6.3 seconds excluding feature extraction time. To calculate feature extraction time, we utilize 1200 frames of resolution 480 $\times$ 360, and get 0.8 frames/sec for cuboid features.

3.7 Conclusions

In this chapter, we proposed the concept of temporal order invariance and investigated our conjecture that : "*The temporal order of action units within an action is invariant to viewpoint variations*". To ensure global temporal order in part-based action representation, we utilize spatio-temporal features, feature fusion and geometrical order constraint. For each action class, we construct a feature fusion table to facilitate feature matching. A matching score is then calculated based on global temporal order consistency constraint and number of matching features. Finally, the action label of the class with maximum value of matching score is assigned to the query action. Experimentation is performed on challenging multiple view *IXMAS* and *WVU* action datasets with encouraging results comparable to the existing view-invariant action recognition techniques. Our framework is independent of the type of spatio-temporal detectors. The reason for selection of cuboid features is their remarkable success in part based action recognition approaches and robustness to noise and occlusion.

However, in case of very complex action dynamics and difficult camera viewpoint, number of spatio-temporal features decrease and too few features can effect performance. In addition, if some motion is present in background context of videos, it also generates spatio-temporal features which may lead to wrong interpretation of actions. To deal these limitations of feature extraction, we explore epipolar geometry. In the next chapter, we exploit epipolar geometry for extracting view-invariant action recognition without trajectory tracking, feature extraction and training by using spatio-temporal optical flow.

Chapter 4

Action Analysis using Epipolar Geometry

> Nobody untrained in geometry may enter my house.
>
> ~Plato (428 BC - 348 BC)

The research work presented in this chapter has been published as:

1. *Anwaar-ul-Haq, Iqbal Gondal and Manzur Murshed,* "On dynamic scene geometry for view-invariant action matching", *In Proc.* CVPR 2011.
2. *Anwaar-ul-Haq, Iqbal Gondal and Manzur Murshed,* "AVITAR: Achieving view-invariant tracking-free action recognition", submitted to IEEE Transaction on Image processing, 2012.

In previous chapter, we proposed the concept of temporal order invariance to solve view-invariant action recognition and pointed out its limitations due to feature extraction. An alternative way of achieving view-invariant action recognition is the exploitation of geometrical models between different views of the same action. However, geometrical approaches heavily rely on tracking. For instance, these approaches consider detection of landmark points on human body and their tracking by assuming that motion trajectories for all landmark points are available throughout the course of an action (figure 4.1). Unfortunately, due to occlusion and noise, detection and tracking of these landmark points is not robust. To alleviate this problem, majority of the work assumes that point trajectories are manually marked which is a clear drawback and lacks automation claimed by computer vision. This chapter presents important visual cues extracted from geometrical constraints and flow correspondences to avoid landmark point detection and their tracking for extraction action dissimilarity measures.

Our geometrical model is based on multiple view geometry fitting between action instances. The benefits of using multiple view geometry is that it simplifies hard problems related multiple views [32, 33, 34]. Estimation of essential matrix and then fundamental matrix from stereo image pair goes back to Longuet-Higgins and eight point algorithm

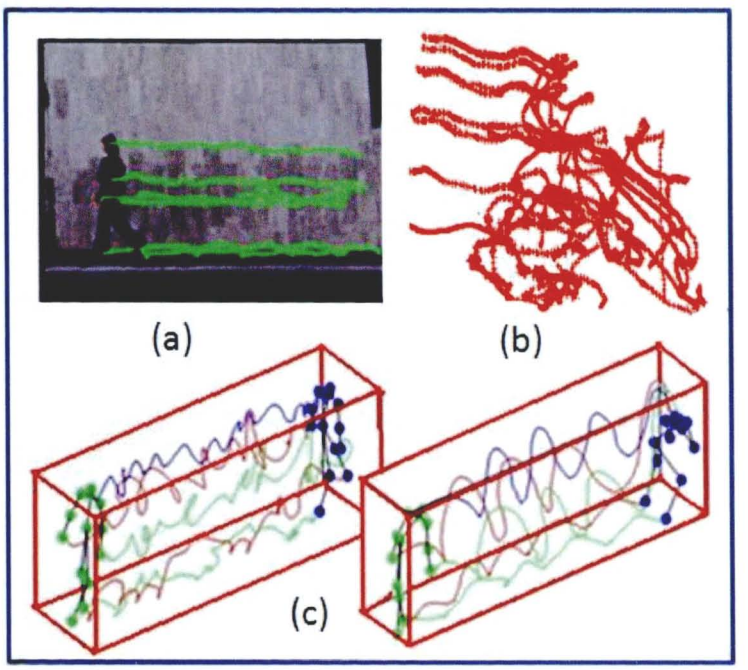


Figure 4.1: Traditional trajectory based action representations which show landmark detection on actor body and tracking of landmark points. (a) walking action tracking, (b) complex trajectories for an activity and (c) trajectories within an action volume.

[35]. Therefore, inspired from related geometrical models, a successful series of incremental work related view invariant action recognition is addresses in [36, 37, 38, 39, 40, 41, 42, 43, 44] which is based on the consideration of action point trajectories and exploitation of geometry between trajectories of different views of the same action. One of the major benefit of these geometrical based methods is that such methods do not need any training. The basic idea originated with the use of affine epipolar geometry constraints in a series of work [36, 38] which showed that the maxima in space-time curvature of a 3D trajectory are persevered in 2D image trajectories.

The main drawback of these approaches is the assumption of affine cameras. For projective camera model, trajectories of 13 anatomical landmarks are matched by [42] under viewpoint, anthropometric and temporal transforms. Another related work is the use of the point triplets with homography, rank constraint [40] and fundamental ratios [41] which consider that the motion of an articulated body can be decomposed into rigid motion of planes defined by triplet of body points. The main drawback of the all above approaches is the decoupling of tracking and matching. It is assumed that tracking of the landmark points on human body has been performed and trajectories are available. Despite its success, it is hard to achieve as basic assumption is very strong. Due to occlusion and noise, the detection of landmark points is not always robust resulting in manual interventions. As a result detection of landmark points and their tracking is performed manually and epipolar geometry rank constraints are applied on manually obtained trajectories by almost all the representative geometrical based methods [36, 37, 38, 39, 40, 41, 42, 43, 44] which lack automation and to make practical use of geometrical solutions, this problem is needed to be addressed and its solution becomes the objective of our work.

The **novelty** of our work is the development of view-invariant action dissimilarity measures without any tracking. It avoids the use of salient point detection on human

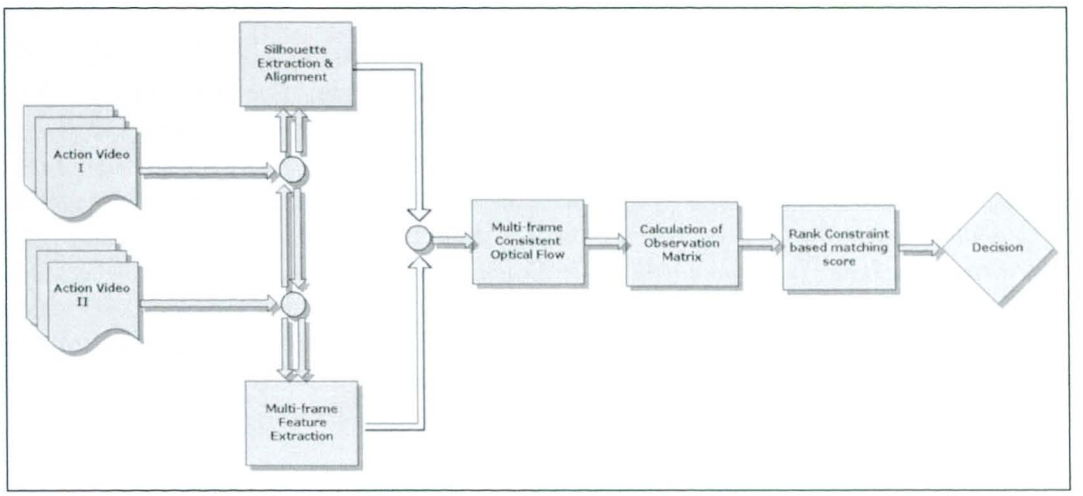


Figure 4.2: Flow diagram of general framework of our approach, AVITAR (achieving view-invariant tracking-free action recognition). It shows how two video sequences are processed to calculate action matching score is calculated from a series of steps. First, based on option (AVITAR1—AVITAR2), silhouettes are extracted or multi-frame feature matches are calculated and corresponding optical flow is used for calculation of flow correspondences then score is calculated based on rank of corresponding observation matrix.

body, trajectory calculation and trajectory matching which are long standing assumptions in geometrical based action recognition methods. This is achieved using spatio-temporal optical flow and rank constraint defined to establish epipolar geometry between video sequences containing similar or dissimilar actions. In addition, we explore dynamic scene geometry using two-body epipolar constraint which facilitates to work on original action volumes without prior segmentation of actors. We show that multi-body flow fundamental matrix captures the geometry of dynamic scenes and helps in devising an action matching score across different views.

Given two video sequences captured from unknown viewpoints and containing unknown actions by same or different actors, our objective is to determine that actions are same or different. We further extend it to develop a framework for action retrieval and recognition. We discuss that we have to modify our geometrical model according to available action representations. In addition, the use of rank constraints can save computational efforts such as the calculation of fundamental matrix and only observation or measurement matrix is sufficient to determine if epipolar geometry can be established.

Taking into consideration that human action is a spatio-temporal phenomenon, we apply constraints on optical flow to be spatio-temporally consistent. Spatio-temporally consistent optical flow helps us in devising spatio-temporally consistent flow fundamental matrix and by defining rank constraints on flow fundamental matrix we are able to derive a dissimilarity score for action sequences. We proceed incrementally by defining two variants of our approach: (1) We extract actor body silhouettes from original video sequences and calculate spatio-temporally consistent optical flow between respective frames of two videos and then fit epipolar geometry. As fundamental matrix remains same for static scenes, we can calculate action similarity score between two actions being performed in time domain, (2) In addition, we observed that silhouette extraction is not robust in all circumstances especially in case of noise and occlusion. Therefore, we remove pre-processing step of silhouette extraction theocratically by maximizing the exploitation of epipolar geometry.

The flow diagram illustrating the general framework of our approach is shown in figure 4.2.

We take action representation in static camera environment as a case of dynamic scene where background is stationary and actor is dynamic. As scene is not entirely static, we get inspiration from structure and motion recovery for scenes consisting of both static and dynamic parts, also known as multi-body segmentation from perspective views without knowing which measurement belong to which part of the scene. As we consider only static background and dynamic actor, it is simplified to two-body fundamental matrix, also known as segmentation matrix [46]. It has already been shown [47] that such matrix can linearly be computed from image measurements after embedding all the image points in high dimensional space. Based on these investigations, we derive a new similarity measure for matching actions across different views, without prior segmentation of actors.

Our **contributions** are threefold:

- We try to address strong assumption of landmark point detection and tracking in geometrical based methods and propose a tracking-free training-free approach for view-invariant action matching maximizing the exploitation of multiple view geometry. Therefore, rather than decoupling the problem of tracking and matching, we solve the problem in a single go (Section 4.1),

- We explore and introduce a novel application of multi-body fundamental matrix and propose a novel similarity score for action matching based on the property of segmentation matrix or two-body fundamental matrix (Section 4.2). It helps establishing view invariant action matching framework without any preprocessing on original video sequences,

- We apply optical flow on stereo images (corresponding frames of two action videos) to achieve observation matrix but apply consistency constraint on four images (one image in advance) to get spatio-temporally consistent optical flow. Spatio-temporally consistent optical flow based on loop consistency of four images (combination of consecutive and corresponding images) is introduced. It help in devising a robust observation matrix free from outliers caused by redundancy in optical flow (Section 4.3).

4.1 Calculating Action Matching Score using Static Fundamental Matrix

The epipolar geometry is the intrinsic projective geometry between two views and fundamental matrix encapsulates this intrinsic geometry, a 3×3 matrix of rank 2. In this section, we show how epipolar geometry can be employed to extract action matching score. However, unlike previous trajectory based approaches [36, 37, 38, 39, 40, 41, 42, 43, 44], actions are not represented by trajectories.

4.1.1 Action Representation

We represent the current pose and posture of an actor in terms of all body points in 3D space $A = A_1, A_2 \cdots A_n$ where $A_i = (x, y, z)$, n is equal to size of frame (number of pixels). We calculate flow correspondences based on dense and consistent optical flow between aligned and normalized actor silhouettes from respective frames of two video sequences.

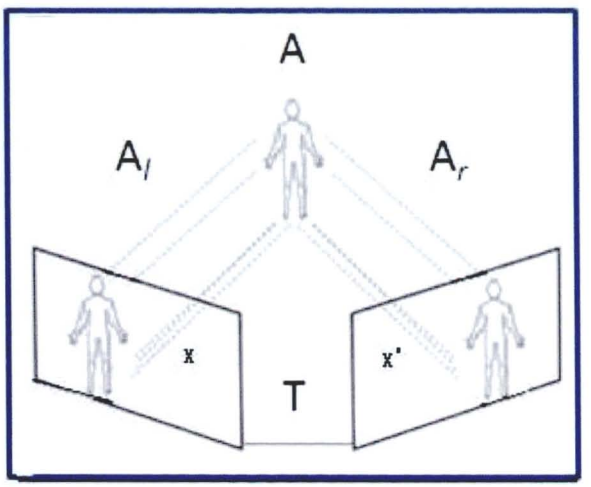


Figure 4.3: Static epipolar geometry with two fixed cameras. A denotes 3D point (point on actor body), A_r and A_l denote projections on right and left image planes. As cameras are static, fundamental matrix should be satisfied between respective frames if sufficient correct correspondences are available.

4.1.2 Establishing Fundamental Matrix

Assume two static cameras view a $3D$ point A as shown in figure 4.3 The vectors A_l and A_r refer to same 3D point in left and right camera frames. The vectors a_l,a_r are the projection of $3D$ point A to the left and right reference camera frames.

The left and right reference frames of camera are related by extrinsic parameters defining a rigid transformation in 3D space as [32, 34]:

$$A_r = R(A_l - T), \tag{4.1}$$

where R is the rotation from left to right camera reference frame, and T is translation vector connecting centers of two cameras.

The equation of epipolar plane through A can be written as the coplanarity condition of vectors P_l,T, and $A_l - T$ or

$$(A_l - T)^T T \times A_l = 0. \tag{4.2}$$

Using equation 4.1 , we get

$$(R^T A_r)^T \times A_l = 0 \tag{4.3}$$

As vector product can be written as a multiplication by a rank deficit matrix,

$$T \times A_l = S A_l \tag{4.4}$$

where

$$S = \begin{bmatrix} 0 & -Tz & Ty \\ Tz & 0 & -Tx \\ -Ty & Tx & 0 \end{bmatrix} \in \mathbb{R}^{3\times 3}. \tag{4.5}$$

By virtue of this fact, we can write equation 4.3 as:

$$A_r^T R S A_l = 0, A_r^T E A_l = 0, E = RS \tag{4.6}$$

The matrix E is called the essential matrix. By perspective projection, we can write:

$$a_l = \frac{f_l}{Z_l} A_l, a_r = \frac{f_r}{Z_r} A_r \tag{4.7}$$

where f_l, f_r are focal lengths and Z is the distance between A and baseline T of stereo system. Therefore, we can write equation 4.6 as:

$$a_r^T R S a_l = 0, a_r^T E a_l = 0, E = RS \tag{4.8}$$

Let M_l and M_r are intrinsic camera parameters. We can write a relationship for image coordinates in left and right views of the scene as $A_r = M_l p_l$ and $A_l = M_r p_r$, where $A_r = (x, y, 1)$ and $A_l = (x', y', 1)$ are homogeneous pixel coordinates. Putting these relations, we can write equation 4.7 as:

$$A_l^T (M_r^{-T} E M_l^{-1}) A_r = A_l F A_r = 0, \tag{4.9}$$

which presents epipolar constraint where $F \in \mathbb{R}^{3\times3}$ is the fundamental matrix.

4.1.3 Derivation of Action Matching Score

Action matching score is based on the assumption that in case of static camera, fundamental matrix must be satisfied between two action instances represented in respective frames of two video sequences containing only intensity based human silhouettes. To determine that if fundamental matrix is satisfied, it is not necessary to calculate it as rank of measurement matrix is enough to determine that fundamental matrix exists. Therefore, defining a rank constraint, we can derive action matching score. The only worry is the calculation of measurement matrix based on reliable correspondences. Rather than using manual correspondences or trajectories, we use dense flow correspondences that further eliminate the need of tracking.

Optical flow establishes relationship between moving actor body between two respective frames of two video sequences. We use flow correspondences between respective frames as $A_r = (x, y, 1)$ and $A_l = (x + u, y + v, 1)$. To find meaningful correspondences, we assure that both silhouettes are spatially aligned and normalized which we attain through aligning centroid of region of interest (ROI) in respective frames. In addition, rather than using binary values we use intensity values for calculating reliable optical flow.

For x_i where $i = 1 \cdots n, n \geq 8$, we can write an expression $OF = 0$ where $O \in \mathbb{R}^{n\times3}$ is the observation or measurement matrix based on image correspondences between corresponding silhouette frames as:

$$\begin{pmatrix} x_1^2 + u_1 x_1 & x_1 y_1 + u_1 y_1 & x_1 + u_1 & x_1 y_1 + v_1 x_1 & y_1^2 + v_1 y_1 & y_1 + x_1 & x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_n^2 + u_n x_n & x_n y_n + u_n y_n & x_n + u_n & x_n y_n + v_n x_n & y_n^2 + v_n y_n & y_n + x_n & x_n & y_n & 1 \end{pmatrix} F = 0. \tag{4.10}$$

For unique solution, rank of O has to be eight. Unfortunately, due to noise it may not be exactly eight. In this case, the smallest $9th$ singular value of O should be close to zero. Based on the property that fundamental matrix F remains the same for static scenes, we can use this property for matching different view actions across different video sequences without any tracking.

4.2 Calculating Action Matching Score using Multi-body Fundamental Matrix

In previous section, we have assumed that actor silhouettes are available which require fine segmentation of actors from their background scene. In this section, we explore how we can use epipolar geometry without segmentation of actors from their background. The need for extending epipolar geometry based action matching score to two-body case is the removal of actor silhouette extraction as pre-processing assumption which is difficult to achieve in scenarios of noise and occlusion. Therefore, compared to previous section which uses extracted actor silhouettes, here we use original video frames containing complete scene including actor as well as background.

Let (X_1, X_2) be image point pair associated with two frames of a scene belonging to any of n independent moving objects. According to multi-body epipolar geometry constraint [47], there exists fundamental matrix $F_i \in \mathbf{R}^3$ such that following constraint is satisfied:

$$\prod_{i=1}^{n}(X_2{}^T F_i X_1 = 0) \tag{4.11}$$

regardless of the object to which this image pair belongs.

4.2.1 Establishing Two-body Fundamental Matrix

Now, imagine the simplest case of multi-body epipolar geometry with $n = 2$. Fig. 4.4 represents two frames of a scene with static (background) and dynamic (actor) points. Image points pairs (x_1^a, x_2^a), (x_1^b, x_2^b) as subscript is image number and superscript is the type of object; actor (dynamic) and background (static). Equation can be written as:

$$((x_2^a)^T F_1 x_1^a)((x_2^b)^T F_2 x_1^b) = 0 \tag{4.12}$$

This equation is no longer bilinear but rather bi-quadratic of any point X one of the points associated to either actor or background. Furthermore , the equation is no longer linear in F_1 and F_2 but rather bilinear in F_1 and F_2. However, if sufficiently many image correspondences are given, we can still recover F_1 and F_2 despite the fact that we do not know the object or motion to which each image pair belongs [47].

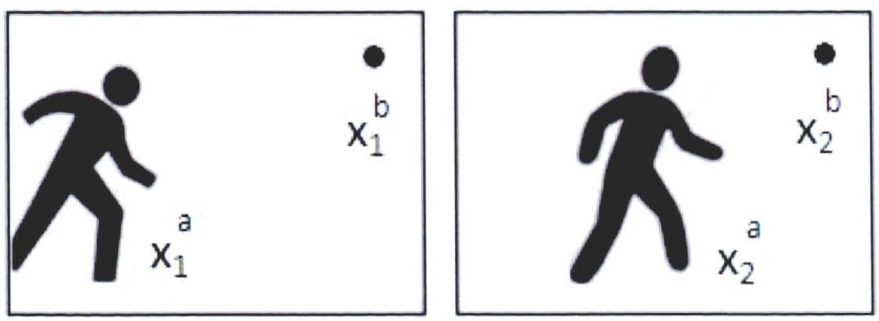


Figure 4.4: Two views of two independent objects in each image, one static (belonging to background) and other dynamic (belonging to actor) and two-body epipolar geometry is explored in this scenario.

To convert a nonlinear problem into linear problem, polynomial embedding of image points to high dimensional space can be used. Veronese map of degree 2 can be used. Let

$X = (x, y, z)$ be any image point belonging to either actor or background, Veronese map of degree 2 for X is given as:

$$\nu_2(X) = [x^2, xy, xz, y^2, yz, z^2] \in \mathbb{R}^6. \tag{4.13}$$

Definition 1: Veronese Map

The Veronese Map of degree n $\upsilon_n : \mathcal{P}^2 \to \mathcal{P}^{M_n - 1}$ be the nth order lifting giving:

$$\upsilon_n(X) = [x^n, x^{n-1}y, x^{n-1}z, \cdots, z^n]^T \tag{4.14}$$

with total of

$$M_n = \binom{n+2}{2} = \frac{(n+1)(n+2)}{2} \tag{4.15}$$

different monomials.

The Veronese map can convert the multi-body epipolar constraint into a bilinear expression. The Knonecker product of Veronese map $\nu_2(X_1) \in \mathbb{R}^6$ and $\nu_2(X_2) \in \mathbb{R}^6$ is a vector in $\mathbb{R}^{36}$ whose entries are exactly the same as monomials given in (18):

$$\nu_2(X_1) \otimes \nu_2(X_2) = [m_1, m_2, \ldots, m_{36}]^T \in \mathbb{R}^{36} \tag{4.16}$$

where $m_i's$ are the monomials sorted in the degree-lexicographic order:

$$\begin{pmatrix} x_1^2x_2^2, x_1^2x_2y_2, x_1^2x_2z_2, x_1^2y_2^2, x_1^2y_2z_2, x_1^2z_2^2 \\ x_1y_1x_2^2, x_1y_1x_2y_2, x_1y_1x_2z_2, x_1y_1y_2^2, x_1y_1y_2z_2, x_1y_1z_2^2 \\ x_1z_1x_2^2, x_1z_1x_2y_2, x_1z_1x_2z_2, x_1z_1y_2^2, x_1z_1y_2z_2, x_1z_1z_2^2 \\ y_1^2x_2^2, y_1^2x_2y_2, y_1^2x_2z_2, y_1^2y_2^2, y_1^2y_2z_2, y_1^2z_2^2 \\ y_1z_1x_2^2, y_1z_1x_2y_2, y_1z_1x_2z_2, y_1z_1y_2^2, y_1z_1y_2z_2, y_1z_1z_2^2 \\ z_1^2x_2^2, z_1^2x_2y_2, z_1^2x_2z_2, z_1^2y_2^2, z_1^2y_2z_2, z_1^2z_2^2 \end{pmatrix} \tag{4.17}$$

These 36 monomials are 'basis' in the space $\mathbb{R}^{36}$. Finally, two-body epipolar constraint now can be written as:

$$\nu_2(X_2)^T F \nu_2(X_1) \tag{4.18}$$

We call $F \in \mathbb{R}^{6\times 6}$ as two-body fundamental matrix.

4.2.2 Derivation of Action Matching Score

This section uses rank constraint on polynomial embedding of flow correspondences for matching actions across different viewpoints. The action matching score is calculated by satisfying the existence of two-body fundamental matrix but we do not need to calculate two-body fundamental matrix as rank of measurement or observation matrix is sufficient to tell if two-body fundamental matrix is satisfied. However, as this observation matrix is calculated by complete two-body scenario (dynamic actor and static background) rather than single body (actor silhouettes), defined rank constraint is also different. Reliable flow correspondences in this case are possible by using feature matching based consistent optical flow to be described in next section.

Given a collection of N image point pairs $(X'^j, X^j)_{j=1}^N$, the vector f satisfies the system of linear equations:

$$Of = \begin{bmatrix} (v_2(X'^1) \otimes v_2(X^1))^T \\ (v_2(X'^2) \otimes v_2(X^2))^T \\ \vdots \\ (v_2(X'^N) \otimes v_2(X^N))^T \end{bmatrix} f = 0. \tag{4.19}$$

In order to recover F uniquely from above, we need:

$$rank(O) = M_n^2 - 1 \tag{4.20}$$

In our special case, $n = 2$, therefore, for linear solution to exist, F must have at most rank thirty five according to [47, 72], and we can take the smallest singular value of O as similarity measure for view invariant action recognition. Unfortunately, due to noise, the rank of matrix F may not be exact. In this case, the smallest singular value of O should be close to zero. This is the similarity score derived from two-body epipolar geometry to be employed on original action volumes without any pre-processing on videos for matching actions.

4.3 Seeking Spatio-temporally Consistent Flow Correspondences

Actions are spatio-temporally dynamic patterns for which we get dense stereo flow correspondences. Several optical flow algorithms are now present in literature. However, due to temporal variations and movements of actor body parts, majority of optical flow algorithms are unreliable in our case. We want these flow correspondences to remain consistent within consecutive stereo pairs of respective action sequences that is possible only by using multi-image spatio-temporally consistent optical flow. Most recently, spatio-temporally consistent optical flow [88] have been proposed based on loop consistency of three images (two consecutive from one sequence and third from respective stereo sequence). To accommodate these algorithms to serve our need, we had to extend loop consistency to four image (two respective frames from each stereo sequence).

Similarly, three image based feature matching [89] is extended to four-image feature matching to be used for feature based spatio-temporally consistent optical flow. This extension increases performance in terms of average angular (AAE) and average endpoint error (AEE) as shown in Fig. 4.5 which presents and compares motion field, AAE, AEE for standard stereo video sequences of waving. Standard stereo video sequence is used as its benchmark (ground truth motion field) is available.

4.3.1 Multi-frame Spatio-temporally Consistent Optical Flow : Four Frame Case

We consider a stereo video setup with two cameras providing un-calibrated and not necessary synchronized image sequences. We refer two frames as $I_1 = \Omega \subset \mathbb{R}^2 \longrightarrow \mathbb{R}$ and $I_2 = \Omega \subset \mathbb{R}^2 \longrightarrow \mathbb{R}$ and the forward flow between them is $f_{1,2} : \Omega \subset \mathbb{R}^2$ and backward flow as $f_{2,1} : \Omega \subset \mathbb{R}^2$. We build our four image based optical flow based system by enforcing the symmetry between forward and backward flow. It means that if a point x in the first image I_1 does not become occluded, following its flow to the second image I_2 and then returning with the backward flow should remain exactly the same at the starting point. This forward and backward flow can be written with a symmetry condition as:

$$f_s(x) = f_{1,2}(x) + f_{2,1}(x + f_{1,2}) \approx 0. \tag{4.21}$$

If we extend it further and consider four temporally and spatially neighboring frames i.e. I_1, I_2, I_3 and I_4. For a point that is visible in all four frames, a loop from I_1 over I_2, I_3 and I_4 and going back to I_1 should end exactly at the starting place. This loop consistency can be written as:

$$\begin{aligned} f_l(x) &= f_{1,2}(x) + f_{2,3}(x + f_{1,2}) + f_{3,4}(x + f_{1,2} + f_{2,3}) \\ &\quad + f_{4,1}(x + f_{1,2} + f_{2,3} + f_{3,4}) \approx 0. \end{aligned} \tag{4.22}$$

All four flows involved in this loop consistency are unknown initially as an iterative strategy is defined by following the $TV - L^2$ (total variation- L^2 norm) framework [88], an optimized differential optical flow framework for stereo videos. Accordingly, the update is defined as :

$$f_{i,j}^{k+1} = f_{i,j}^{k} + \alpha df_{i,j}. \tag{4.23}$$

where $\alpha = \psi(f_s)\psi(f_l)$ is weighting parameter with $\psi = 1 - exp(\frac{-\|x\|_2^2}{d})$ with constant $d > 0$. To update the flow $f_{i,j}^k$, other unknown flow are kept fixed. Using a quadratic energy function E with differentiable L^2 norm

$$E_q = \|I_i - I_j(x + f_{i,j}^k) + \nabla I_j df_{i,j}\|_2^2 + \frac{1}{\Theta}\|f_{i,j}^k + df_{i,j} - \tilde{f}_{i,j}\|_2^2. \tag{4.24}$$

setting $\frac{\delta E_q}{\delta df_{i,j}} = 0$, resulting 2×2 linear system is solved for the update $df_{i,j}$. Θ is auxiliary variable and ∇ is smoothness parameter. The current estimate $f_{i,j}^{k+1}$ is used to calculate TV-optimized version $\tilde{f}_{i,j}^k$. Then all other unknown flow fields are updated. Flow fields are updated only when the symmetry and loop consistency constraints are satisfied.

4.3.2 Spatio-temporally Consistent Optical Flow based on Multi-frame Matching

Multi-image feature point matching can be helpful for optical flow calculation on unsynchronized stereo sequences. Usually, feature matching is performed between two images at a time like nearest neighbor matching that compares the distance of the nearest neighbor to the distance of the second nearest neighbor and only accepts a match if their ratio is below a threshold [100] but recently three-image feature matching has been proposed by [89]. We build upon this idea to extend feature matching framework to four spatio-temporally neighboring stereo video frames described in Appendix A. Now, we can include matched features into optical flow for stereo sequences described in the previous section.

4.4 Robustness to Anthropometric Variations, Occlusion and Noise

In addition to temporal synchronization, our approach shows considerable robustness to anthropometric variations, occlusions and noise. Here is the detail how we deal these factors:

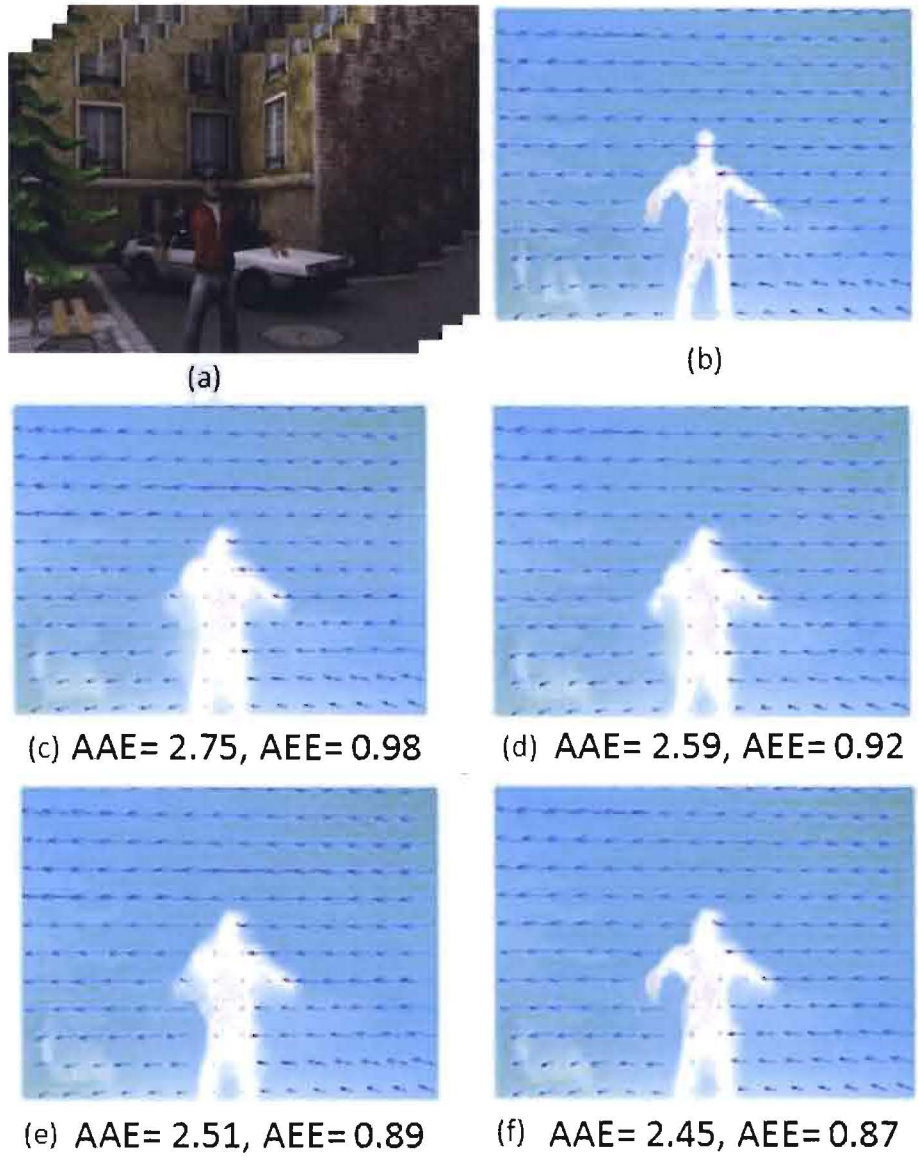


Figure 4.5: Spatio-temporally Consistent Optical Flow Multi-frame Setup: It utilizes four frames, two consecutive frames from each stereo video sequence. It is calculated between two stereo frames but temporal consistency constraint is used obtain only temporally consistent flow values. Fig. shows the original setup but as AVITAR 1 uses intensity based silhouettes, it utilizes silhouettes.

Algorithm 2 AVITAR (A generic snapshot of pseducode for matching actions in two videos.)

```
1: procedure CALCULATESCORE(V1, V2, nframes)
2:     for i ← 1, frames do
3:         I1 ← V1[i]
4:         I2 ← V2[i]
5:         I3 ← V2[i + 1]
6:         I1 ← V1[i + 1]
7:         [U, V] ← consistentflow(I1, I2, I3, I4)
8:         [X1, X2] ← findMatches(I1, I2, U, V)
9:         O ← calObservationMatrix(X1, X2)
10:        E ← SVD(O)
11:        Score[i] ← min(E)
12:    end for
13: end procedure
```

4.4.1 Dealing Anthropometric Variations

Human action is very complex in nature which is affected by different anthropometric variations. To deal these variations, we base our approach on posture constraint [43] in multiple view geometry which articulates that fundamental matrix is satisfied between two actors if their postures are same irrespective of their anthropometric variations (scale, clothing etc.).

Postural constraint is based on the conjecture that two actors performing the same action have similar postures at a corresponding time instant giving a clue that actions can be recognized by measuring the dissimilarity of postures based on epipolar geometry.

Fundamental matrix does not encapsulate only the relative position in different views but relative poses of actors and their anthropometric variations as well. Therefore, fundamental matrix should be satisfied between similar postures of different actors.

4.4.2 Dealing Noise

Noise is inherent problem of almost every system. In our case, dense flow correspondences may contain noise. It is handled in two fashions: (i) Consistent optical flow based on loop consistency removes outliers and shows robustness to noise, (ii) Only 8 correct correspondences in case of static F and 35 in case of two-body F are required to find a unique solution.

Ideally the smallest singular value of A should be zero but in case of noise it can deviate from zero. Even if it is close to zero, it demonstrates good differentiation characteristics which is utilized by our approach.

4.4.3 Dealing Occlusion

In real world scenarios, occlusion is unavoidable. Occlusion handling is not directly addressed in this work but our algorithm has sufficient support for occlusion handling due to loop consistency constraint in consistent optical flow calculation.

Figure 4.6: Effect of anthropometric variations. Epipolar geometry is fitted between points on two subjects with physical differences in structure but with similar pose. In this experiment, we click points on body left subject and respective epipolar line is automatically drawn on right subject passing through respective body point. It validates posture constraint.

4.4.4 Dealing Temporal Synchronization

Temporal un-synchronization of actions may be caused due to execution of actions or the different frame rates of the camera. To deal these anthropometric variations, we base our approach on posture constraint [43] in multiple view geometry which articulates that fundamental matrix is satisfied between two actors if their postures are same irrespective of their anthropometric variations (scale, clothing etc.).

Postural constraint is based on the conjecture that two actors performing the same action have similar postures at a corresponding time instant giving a clue that actions can be recognized by measuring the dissimilarity of postures based on epipolar geometry. Fundamental matrix does not encapsulate only the relative position in different views but relative poses of actors and their anthropometric variations as well. Therefore, fundamental matrix should be satisfied between similar postures of different actors. This scenario is presented in figure 4.6 in which epipolar geometry is satisfied between two different actors with different clothing and body variations but having same posture.

4.5 Experimental Results

4.5.1 Datasets and experimental Set-up

For experimentation and performance comparison, we have used standard action datasets: (1) IXMAS Dataset [15], and (2)Multi-view WVU Action Dataset [31]. These are benchmark datasets for view-invariant action recognition. Our experimentation includes both variants of our approach: (1) AVITAR 1: which uses static F based matching score using

algorithm 1 and (2) AVITAR 2: which uses two-body F based matching score using algorithm 2. We present recognition results for our approach in terms of confusion matrices and video retrieval results. We also do robustness analysis to different effects of noise, occlusion and viewpoint variations. First, we give a brief description of our data sets:

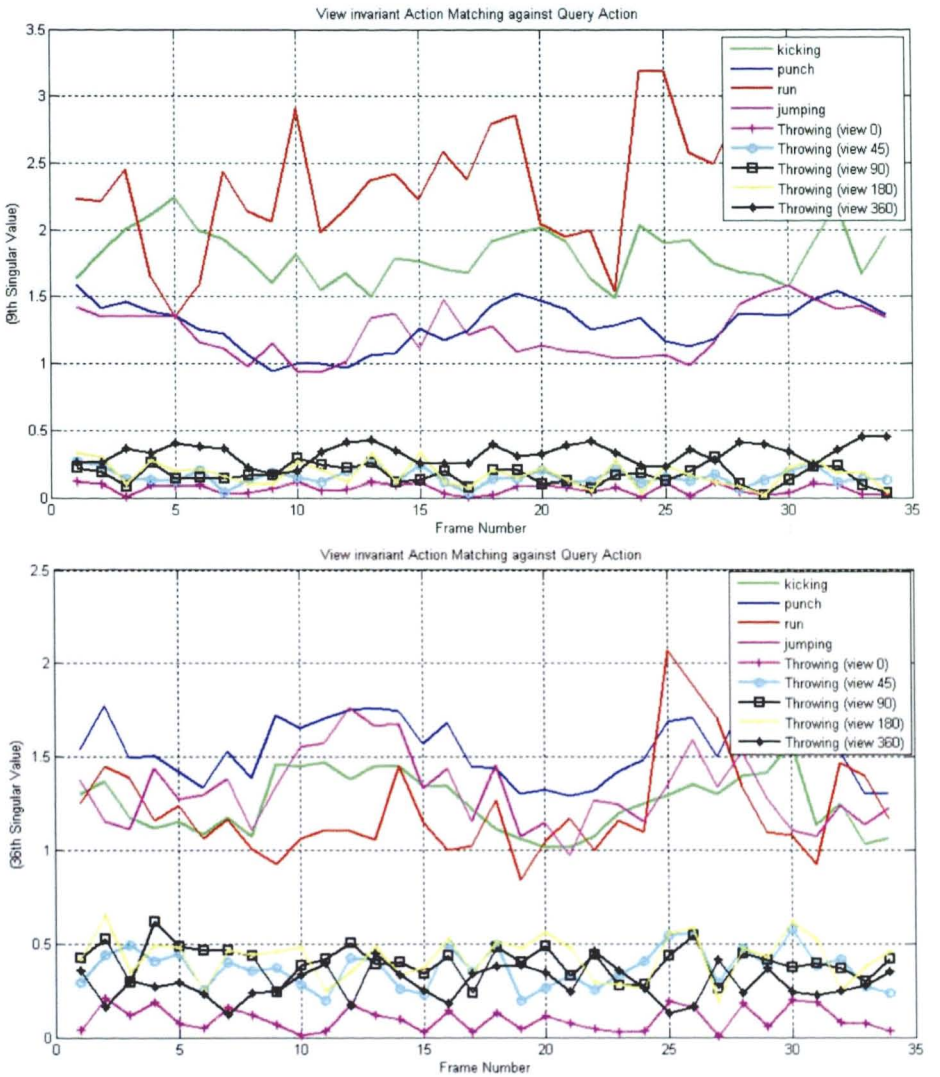


Figure 4.7: The video matching results against the query action sequence for WVU multi-view dataset

IXMAS Action Dataset: The Inria Xmas Motion Acquisition Sequences (IXMAS) is widely used data set for view-invariant action recognition. It contained 11 actions, each performed by 10 actors three times and captured from five different views. We collected 1650 video sequences. These actions include check-watch, cross-arms, scratch-head, sit-down, get-up, turn-around, walk, wave, punch, kick and pick-up.

The variations in viewpoints between five different cameras pose significant challenge especially in case of cam4 which is placed above the actor. All actions are temporally synchronized.

WVU Multiview Action Dataset: The dataset was collected as part of the research work on real-time human action recognition in a camera network. The multi-camera network

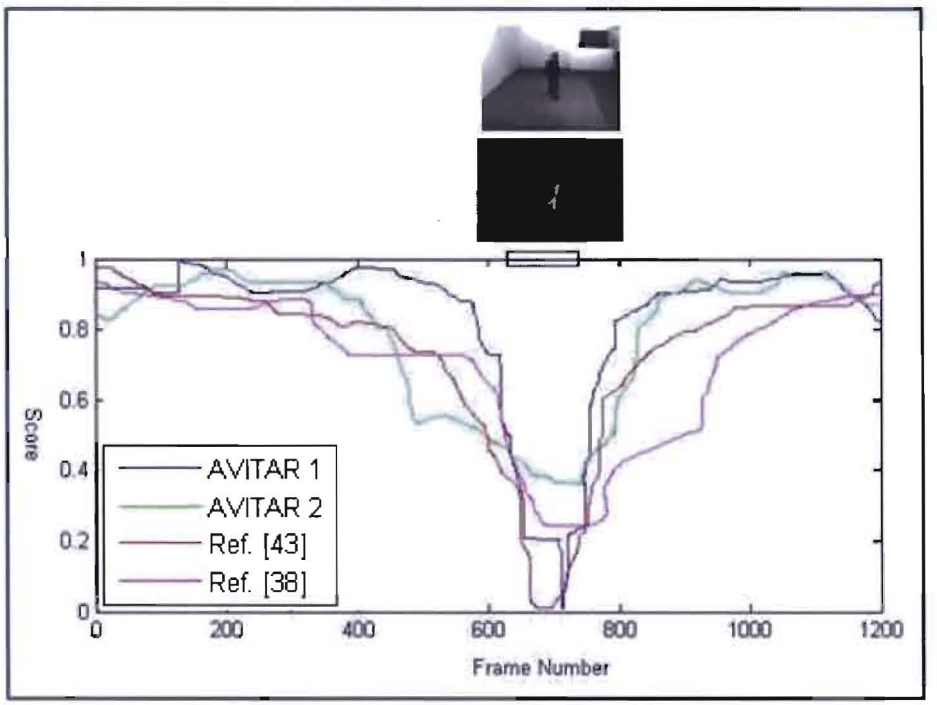


Figure 4.8: Video Retrieval results for walking action of Alba action against different view long video sequence which contains 1200 frames.

system consists of 8 cameras that provide completely overlapping coverage of a rectangular region R (about 50 $\times$ 50 feet) from different viewing directions.

It contained 11 actions, each performed by 10 actors three times and captured from five different views. These actions include nodding head, clapping, waving 1 hand, waving 2 hands, punching, jogging, jumping jack, kicking, picking, throwing and bowling. However, we exclude *standing* sequences as standing do not represent any action.

4.5.2 Action matching and retrieval

In these experiments, we show how action similarity scores demonstrate their discriminative property to match similar actions in the presence of different actions in the same view as well as same action captured from different viewpoints. for better understanding, we present both microscopic and macroscopic analysis of our experiments. We show that discriminative property of our measures is helpful in retrieval of similar actions against an action query similar to leave one out cross validation strategy.

In microscopic analysis, we take only few frames (35 frames for each action sequence) and calculate action similarity scores proposed above in AVITAR 1 and AVITAR 2 that uses the smallest singular value of respective observation or measurement matrix displayed in figure 4.7. In this experiment, 9 action sequences are matched against test sequence of *throwing* action. These 9 sequences are comprising four (1 for each action of kicking, punching, running and jumping) and five sequences of throwing action captured from different viewpoints (0,45,90, 180, 360). X-axis in figure 4.7 shows frame numbers and Y-axis shows respective matching score. It shows that throwing sequences from all viewpoints are clearly differentiable (contain lower values) against test throwing action by different actor. This experiment is performed severalty for AVITAR 1 and AVITAR 2.

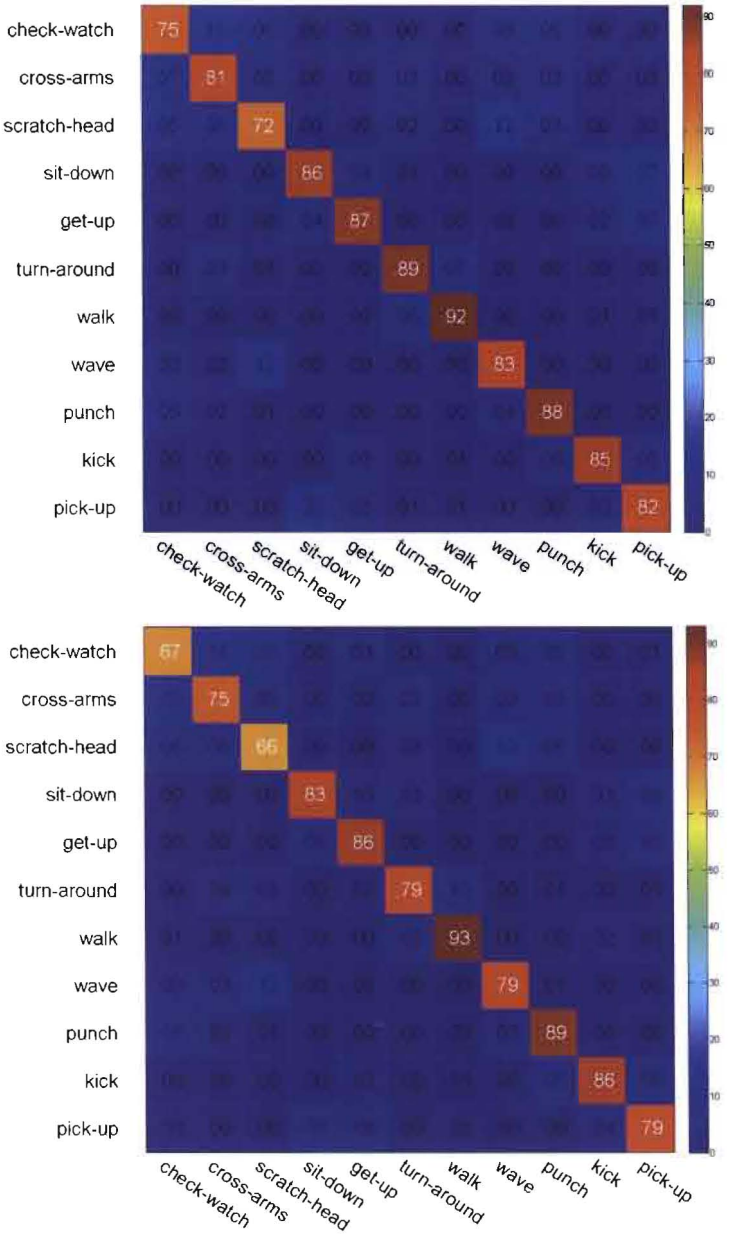


Figure 4.9: Confusion matrix for IXMAS dataset against AVITAR1 and AVITAR2

In macroscopic analysis, we additionally retrieve individual action from long video sequence comprising all 11 actions in the sequence. We use Alba action sequence in IXMAS multi-view data set for this experiment. The query action is taken as walking action and retrieval results against different view test sequence based on the smallest singular-values of measurement matrices are shown in figure 4.8. X-axis in figure 4.8 shows frame numbers and Y-axis shows respective matching score. AVITAR1, AVITAR 2 and comparable trajectory based approaches [36, 37, 38, 39, 40, 41, 42, 43, 44] are used in this experiment. The smallest singular values of observation matrix should be lower (value approaching to zero) when matching action segment starts and higher elsewhere. Therefore, action segment can be retrieved for values below a given threshold. A threshold of 0.4 is set for this experiment. All approaches were able to extract action segment. However, advantage of our approaches is the achievement of goal without tracking.

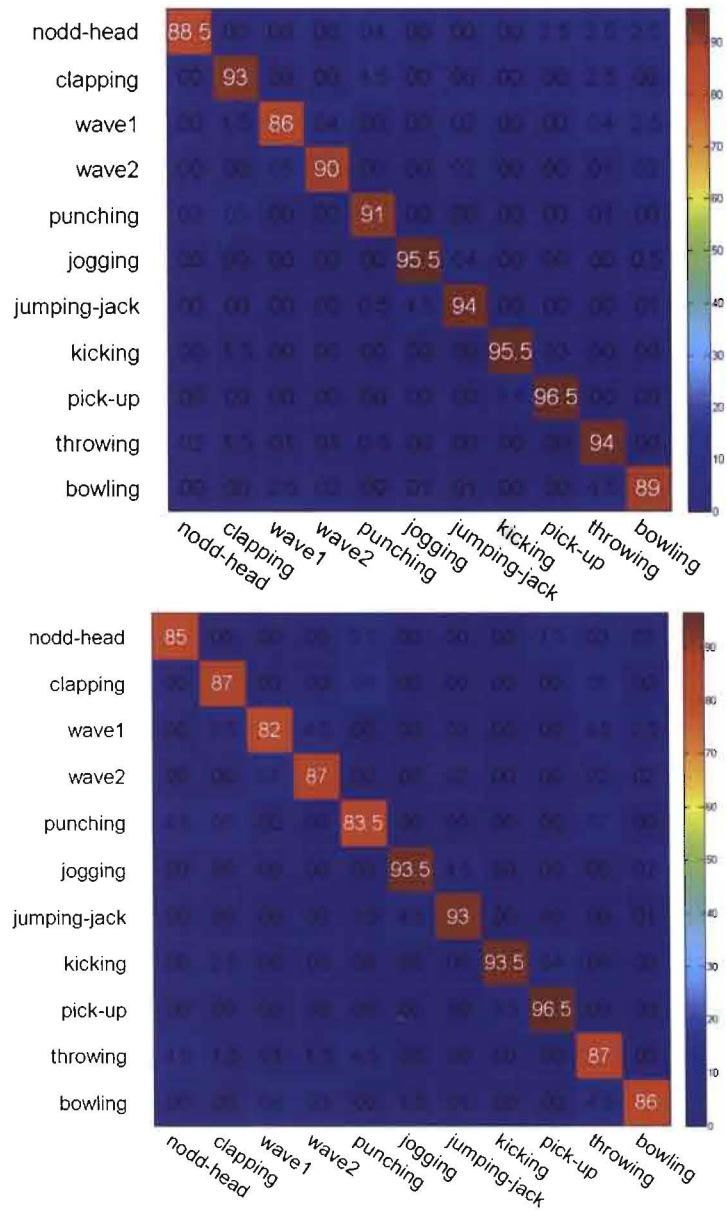


Figure 4.10: Confusion matrix for WVU dataset against AVITAR1 and AVITAR2

4.5.3 Action recognition

For action recognition, We use entire data from IXMAS and WUV dataset. We calculate the confusion matrices in both cases. Confusion matrices are standard way of representing recognition accuracy. We divide each action into fixed number of frames and repeat recognition based on average value of matching scores form complete sequence. We use leave one out cross validation (LOOCV) strategy and match each action from arbitrary viewpoint against all other actions including same action from all viewpoints. The diagonal values comprising higher average recognition results show success of our approaches. These confusion matrices are shown in figure 4.9 and figure. 4.10 respectively.

For WVU multiview dataset, we take actions from viewpoint V1 as query and match it against all 7 different views separately. We calculate action matching scores using both algorithms, (AVITAR 1, AVITAR 2). Finally, we calculate confusion matrices based on average recognition accuracy against all five cameras which are shown in figure 4.11. V5 and

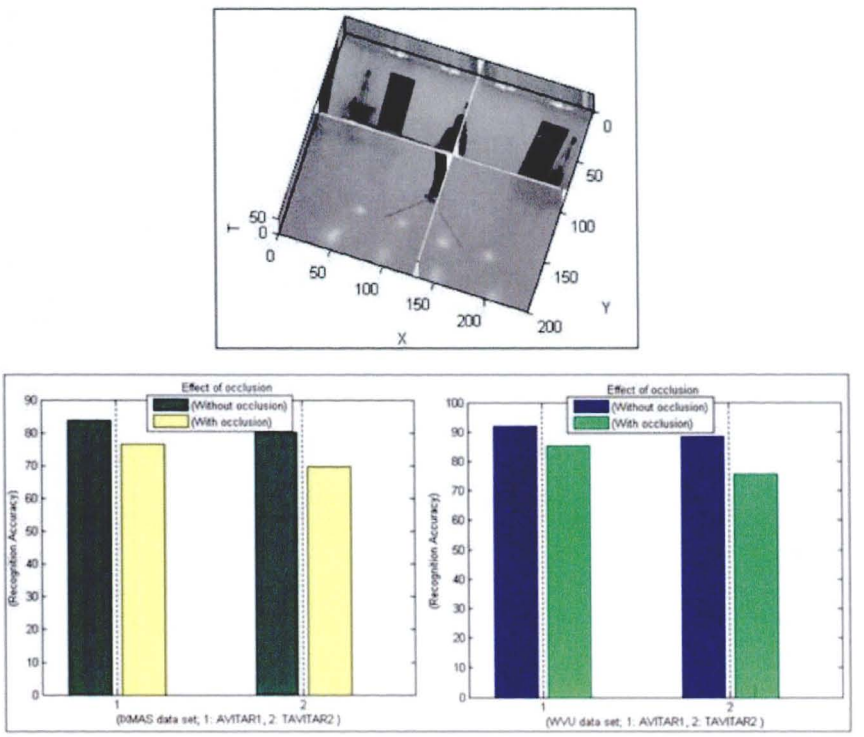


Figure 4.11: Effect of occlusion on recognition accuracy of two datasets used: IXMAS and WVU using AVITAR 1 and AVITAR 2 approaches .

V6 proved different viewpoints while actions nodd-head and wave1 showed comparatively low accuracy as other actions.

Similarly, for IXMAS dataset, we take cam1 as query and match it against all 5 different views separately. We calculate action matching scores using both algorithms, (AVITAR 1, AVITAR 2).

Finally, we calculate confusion matrices based on average recognition accuracy against all five cameras which are shown in figure 4.11. In addition, we show performance of our approach for each of five cameras separately shown in figure 4.13 and 4.14.

The cam5 proved to be the most difficult viewpoint as it is placed exactly above the actor. The average accuracy for all five viewpoints for static and two-body case is 83.69% and 79.45% respectively which is comparable to the state of the art as shown in table 4.5.5. Two-body case came with lower accuracy which is due to large variations in background scenes. Another explanation is that most frames came with less reliable matches and we had to skip these frames.

4.5.4 Robustness to noise and occlusion

Due to unavailability of action dataset with known effects of noise and occlusion, we decided to use artificial effects to introduce noise and occlusion in the dataset.

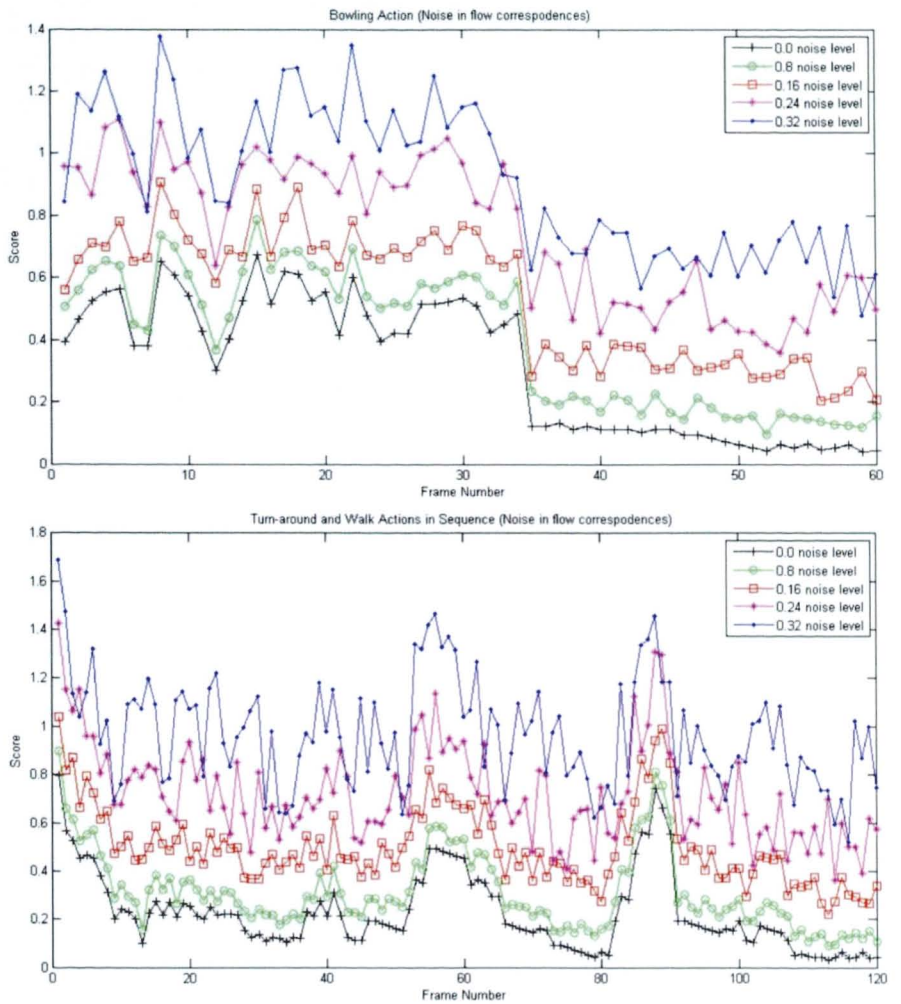


Figure 4.12: Effect of noise on recognition accuracy of two datasets used: IXMAS and WVU using AVITAR 1 and AVITAR 2 approaches .

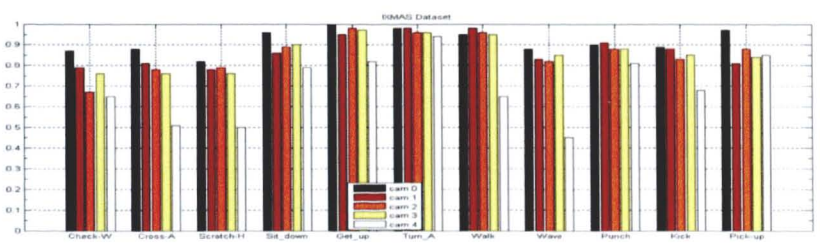


Figure 4.13: Performance for five views of IXMAS dataset against static fundamental matrix based metric.

For noise, we follow the footsteps of [43] and designed an experiment. We added noise sampled from zero mean normal distribution with σ from 0.8 to 0.32 into flow correspondences and recalculated the matching scores (AVITAR1, AVITAR2) for actions bowling and turn around respectively and similar action videos from taken from different viewpoints. Results are shown in figure 4.12 in which X-axis shows frame numbers and Y-axis shows respective matching score. It shows that shows small average divergences

of 0.14 if σ from 0.0 to 0.16 (robustness to certain level)and rise sharply afterwards (due to large number of false correspondences) for AVITAR 1. This trends is followed by AVITAR2 that shows small divergences of 0.22 if σ from 0.0 to 0.16 (robustness to certain level)and rise sharply afterwards (due to large number of false correspondences). We repeat this experiment for other dataset videos that resulted in decrease of 7.8% and 9.1% in recognition rate using AVITAR1 and AVITAR2 respectively.

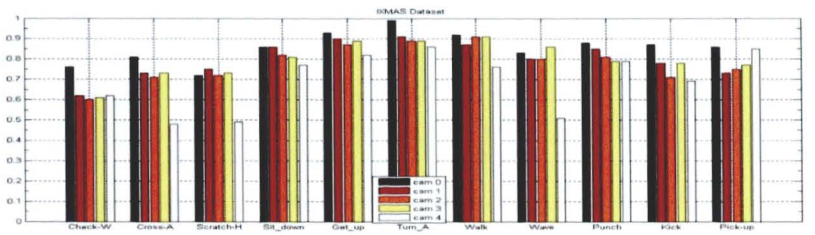


Figure 4.14: Performance for five views of IXMAS dataset against two-body fundamental matrix based metric.

For occlusion, we artificially added occlusion by introducing horizontal and vertical linea. It is done by setting middle 35 rows and columns values to zero. Figure 4.11 shows the reconstructed videos as well as their effects on recognition rate. It resulted in decrease of 4.2 and 7.9 percent in recognition rate using AVITAR1 and AVITAR2 respectively. We observed that AVITAR2 is more effects by both noise and occlusion than AVITAR 1. One possible reason is that addition and occlusion effects the number of correct flow correspondences in this case.

4.5.5 Comparison to other approaches

We used leave one out cross validation in our experimentation, therefore, we compare our approach to those approaches in literature that have used the same strategy.

Method	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
AVITAR	89.0	84.6	86.2	85.0	72.7	83.5
Ref. [52]	86.6	81.1	80.1	83.6	82.8	82.8
Ref. [51]	86.7	89.9	86.4	87.6	66.4	83.4
Ref. [71]	74.8	74.5	74.8	70.6	61.2	71.2

Table 4.1: Performance comparison with the existing techniques. Average recognition is the average performance for all five cameras.

The comparison is taken in terms of average recognition rate. The most of analysis experimental data and results are available about IXMAS action dataset and its five views. In table 4.5.5, we give a detailed comparison in terms of recognition rate.

4.5.6 Discussion and limitations

In this chapter, we explored the use of epipolar geometry for view-invariant action recognition without tracking and training and showed how we maximize its exploitation by defining different variants of geometrical models according to the available input.

We showed that dense optical flow based fundamental matrix can help in devising a tracking free solution to view invariant action recognition. Additionally, in case when

exact segmentation of actor and background is not available, constraints based on static F are no more applicable. It deals correspondences from actor body and background at the same time leading to errors in fitted epipolar geometry. This problem can be solved using two-body fundamental matrix or segmentation matrix.

During experimentation, we also observed a small trade-off between automation provided and recognition accuracy achieved. In case of AVITAR1, we need actor silhouettes but the recognition rate is quite high as we are using only interesting part from original video sequences. On the other hand, for AVITAR2, recognition is slightly lower but it provides an additional ease of fitting epipolar geometry without actor segmentation from original videos. However, the objective and theme of both AVITAR1 and AVITAR 2 is same: (i) They both are based on epipolar geometry, (ii) They both need no tracking involved and (iii) They both need no training to achieve recognition.

Some limiting aspect of our approach is its computational time which is mostly due to computation of dense optical flow but the rest of calculations and computations are very fast (steps after the calculation of flow correspondences). On Intel (R) CoreTM 2 Duo system with 4GB RAM and Matlab code, we get average run-time for testing video of 35 frame is 179.3 seconds for AVITAR1 and 285.2 seconds for AVITAR2 in processing time. Experimental results show that further improvement in recognition accuracy is possible if more accurate optical flow and robust features extraction techniques are used.

4.6 Conclusions

In this chapter, we propose a method to achieve view invariance in action recognition without any tracking. The smallest singular value of measurement matrix is sorted out in static and two-body fundamental matrices and used as action matching score. New action matching scores have been proposed based on efficient utilization of multiple view geometry constraints. The optimal utilization of different preprocessing options is investigated and ways of their minimization are sorted out theocratically and experimentally. The experimental evaluation against well known action datasets validates the fact that actions can be matched across different views without tracking and other strong assumptions.

However, inherent issues of time complexity for optical flow and addition assumption of temporal synchronization of actions are considerable obstacles using geometrical approaches. Both of these problems are dedicated research fields in computer vision. Therefore, to further exploration of view-invariance, we try to exploit those approaches which overcome limitations of optical flow measurements and feature extraction. In next chapter, we discuss 3D frequency domain filtering which tries to fix these issue by providing a better solution without feature extraction and optical flow calculations.

Chapter 5

Action Analysis using 3D Frequency-Domain Filtering

> By nature, men are nearly alike; by practice, they get to be wide apart.
>
> ~ Confucius (551 479 BC)

The research work presented in this chapter has been published as:

1. Anwaar-ul-haq, I. Gondal and M. Murshed, *Action recognition using spatio-temporal distance classifier correlation filters*, *In Proc.* DICTA, Noosa Resort, 2011.
2. Anwaar-ul-haq, I. Gondal and M. Murshed, *VIEW-DCCF: Space Time correlation Filter for View-invariant Action Recognition*, *submitted to* pattern recognition Letters, 2012.

In previous chapter, we discussed visual cues for action recognition build on geometrical modeling based on multiple view geometry. We pointed out computational complexity of optical flow and restriction of temporal synchronization associated with geometrical methods. Despite its success in matching actions across different viewpoints, practical applicability is hard to visualize due to initial assumptions and optical flow. In this chapter, we explore a global action representation based on 3D frequency domain information analysis and filtering. On one hand, it is free from complication of feature extraction and restriction of number of features and on the other hand, it is faster as compared to optical flow based geometrical modeling due to fast frequency domain phase matching.

One successful approach is the application of space-time pattern templates. Representative work includes temporal matching of periodicity information from a set of optical flow frames [48]; two component temporal template of motion energy image (MEI) and motion history image (MHI) [13], space-time shapes induced by the silhouettes in the space-time volume [49] and space-time behavior based correlation [49]. However, majority of these template based approaches suffer from high computational overhead due to spatial template matching.

To overcome the problems faced by these template based methods, the utilization of correlation filters is investigated for recognizing action instances with promising results.

The representative work in this regard is the development of Action filters [29, 121] that has generalized traditional 2D *Maximum Average Correlation Height* (MACH) filter into 3D MACH filters by including temporal dimension. However, the major gain is in terms of low computational cost as response of the filter can be analyzed in frequency domain. A similar frequency domain action matching strategy has been proposed by [50] which addresses inherent discrepancies in MACH filters. Despite promising results for action recognition, these techniques provide no support for recognition actions across different viewpoint variations. As compared to template based view invariant action matching framework like motion history volumes [15], a research gap in present for the development of space-time action filters for matching actions across different viewpoints. It motivates us to propose frequency domain action filtering strategy with robustness to view variations. In addition, we address the inherent discrepancies in traditional filters like ActionMACH filters [29, 121].

One of the weaknesses of MACH filters is their ineffectiveness to encapsulate inter-class variability. Therefore, these filters are trained only for one action class at a time and separate ActionMACH filters are needed for each action class. Secondly, ActionMACH filters overemphasize average training sample, a biased treatment of low frequency components and behave like average filter and may loose finer details of the training set. They emphasize high energy (low frequency) components and attenuate low energy (high frequency) components of the training set leading to poor intra-class discrimination. Thirdly, as action datasets are normally misaligned in space and time, they create problems in learning and testing as synthesized filters are not shift-invariant. Finally, action recognition frameworks based on these correlation filters are not view-invariant. Therefore, to fully utilize the benefits of correlation based action filtering, it is highly desirable to develop correlation filters for unconstraint action recognition. Some representative actions scenes are shown in figure 5.1.

Figure 5.1: Two representative action classes, (Lifting, Walking) show strong intra-class similarity and inter-class discrimination which should be encapsulated by a discriminative filter.

In this chapter, we address above mentioned weaknesses and propose an extended *spatio-temporal distance classifier correlation filter* (Action ST-DCCF filter) for action recognition. Our approach offers following advantages: (i) A single Action ST-DCCF filter successfully captures inter-class variability and avoids overemphasize on average training sample by empirically setting contributions of low as well as high frequency information. (ii) Secondly, it presents a different interpretation of correlation filters as method of applying a *spatio-temporal transformation* to the data and transformation matrix is restricted to being Toeplitz ensuring *shift invariance*. It measures similarity between an ideal transformed reference and testing action using a shift-invariant mean square distance measure handling misalignments and (iii) Another benefit is that resulting decision boundaries are

quadratic which are more 'selective' for choosing feature space portions for assigning to various classes and utilize entire correlation plane rather than emphasizing only single point like correlation peak. These advantages of Action ST-DCCF filter can potentially improve action recognition performance.

In addition, we address above mentioned weaknesses and propose a new view-invariant action recognition approach based on our extended *space-time distance classifier correlation filter* (VIEW DCCF filter) for view invariant action recognition. Our objective is to recognize an unknown test action category taken from arbitrary viewpoint against space-time action filters, each trained for given action categories taken from a specific viewpoint. Our approach offers following advantages: (i) It provides view-invariance, (ii) Action DCCF filter successfully captures inter-class variability and avoids overemphasize on average training sample by empirically setting contributions of low as well as high frequency information. (iii) It presents a different interpretation of correlation filters as method of applying a *spatio-temporal transformation* to the data, restricted to being Toeplitz ensuring *shift invariance*. It measures similarity between an ideal transformed reference and testing action using a shift-invariant mean square distance measure handling misalignments and (iv) another benefit is that resulting decision boundaries are quadratic which are more 'selective' for choosing feature space portions for assigning to various classes and utilize entire correlation plane rather than emphasizing only single point like correlation peak.

These advantages of VIEW DCCF filter can potentially improve performance of view-invariant action recognition. Finally, we extract an action similarity score based on class votes and within-cluster distance ratio. It helps us to recognize actions from an arbitrary viewpoint not present in training view clusters. Class votes help setting priority for class with maximum votes in all view clusters and within-cluster distance ratio highlights margin of selected class from other classes in a view cluster. All these contributions successfully fill up the research gap present in space-time filtering based action recognition.

5.1 The Action ST-DCCF filter

The problem of action recognition can be considered as multi-class discrimination problem by simultaneously including all the classes to be separated. By applying global transformations to the input data, inter-class distance can be increased while making classes as compact as possible. To achieve this objective, correlation can be visualized as a linear transformation and filtering process can be mathematically expressed as multiplication by a diagonal matrix in the frequency domain [116]. For a correlation filter to be used as transform, we require that instances of different classes become as different as possible after filtering. Then, shift-invariant mean square error distances can be computed between the filtered class instance and the transformed references of different classes and input is assigned to the class to which the distance is the smallest. Distance classifier correlation filter is a filter with the above mentioned objective.

Human action is a spatio-temporal construct and therefore, temporal information is an important attribute of action instance. To visualize a distance transform for action instances, we need to extend it in spatio-temporal sense. We name this extension as spatio-temporal distance classifier correlation filter (ST-DCCF).

Mathematically, distance of input action instance $A_{(x,y,t)}$ to a reference $R^c_{(x,y,t)}$ of class c under a linear transformation H can be described as:

$$d^c = |H^*A - H^*r^c|^2 = (A - R^c)^+ HH^*(A - R^c), \tag{5.1}$$

where A is a d-dimensional column vector constructed from a spatio-temporal volume of action instance with $d = x * y * t$ pixels with x horizontal axis, y vertical axis and t as time

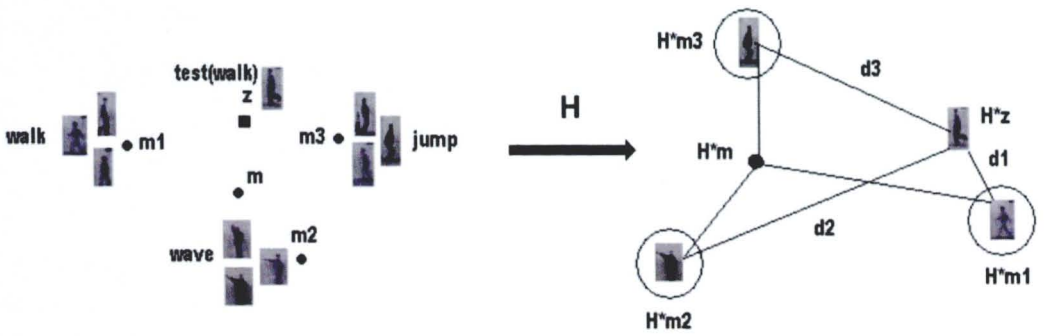


Figure 5.2: The schematic diagram of Action ST-DCCF filtering showing Transformation H which increases inter-class distance while simultaneously making each class more compact. It shows that after the transformation, distance d1 is the smallest making test action closest to walking class.

axis, H is a linear global transform to maximally separate the classes and the superscript $'+'$ represents the complex conjugate transpose. Figure 4.2 schematically depicts the basic idea of 3-class Action ST-DCCF filter.

A general C class distance classifier problem is formulated as: Let a_{ic} be the d-dimensional column vector containing 3D FFT (Fourier Transform) [117] of the ith training action volume of cth class, $1 \leq i \leq N$ and $1 \leq c \leq C$

Let m_c be the mean 3D-FFT of class c such that:

$$m_c = \frac{1}{N} \sum_{i=1}^{N} A_{ic} \tag{5.2}$$

Under a linear transformation H, the differences between the means of any two classes c_1 and c_2 can be written as:

$$v_{c_1 c_2} = H^*(m_{c_1} - m_{c_2}) \tag{5.3}$$

Taking the expectation of the elements of v_{c1} over all frequencies yields

$$\bar{v}_{c_1 c_2} = E_i\{v_{c_1}(i)\} \cong \frac{1}{d} h^+(m_{c_1} - m_{c_2}) \tag{5.4}$$

The quantity in eq. 5.4 is a measure of the *spectral separation* between classless c_1 and c_2 over all frequencies. We want $|\bar{v}_{c_1 c_2}|^2$ to be large. Taking all possible pairs of classes into consideration, we define *spectral separation* (SS) criterion as:

$$SS(h) = h^+[\frac{1}{C} \sum_{i=1}^{C} (c - c_i)(c - c_i)^+]h = h^+ T h \tag{5.5}$$

where $T = [\frac{1}{C} \sum_{i=1}^{C} (c - c_i)(c - c_i)^+]$ is a $d \times d$ non-diagonal matrix of rank $\leq (C - 1)$ and $c = \frac{1}{C} \sum_{i=1}^{C} c_i$ is the mean of the entire dataset. If SS is maximized by appropriate choice of h , the average content of the classes will differ greatly and become well separated. At the same time, we want to improve intra-class compactness by creating balance between low and high frequency components by *similarity measure* (SM) given as:

$$
\begin{aligned}
SM(h) &= \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N}\sum_{i=1}^{N} h^{+}[X_{ic}-(1-\beta)M_c][X_{ic}-(1-\beta)M_c]^{+}h \\
&= h^{+}Sh
\end{aligned}
\tag{5.6}
$$

This step is different from traditional DCCF filter [116] which overemphasizes low frequency components (average training sample) only. The value of β varies from 0 and 2 and by controlling its value, the filter can be prevented from the biased treatment of low frequency information. Other objective is to maximize spectral separation (SS) and minimize similarity measure (SM), thus maximizing the ratio $R(h)$ as:

$$
R(h) = \frac{SS(h)}{SM(h)} = \frac{h^{+}Th}{h^{+}Sh} \tag{5.7}
$$

$$
h = S^{-1}T \tag{5.8}
$$

We refer optimum h as the spatio-temporal distance classifier correlation filter (ST-DCCF). This filter deals with entire correlation space and not just one point at the origin.

5.1.1 Action Classification

Given a test action input z, we determine its distance (shift-invariant mean square error) from other classes, say ideal reference for class c as:

$$
\begin{aligned}
d_c &= |H^{*}z - H^{*}m_c|^2 \\
&= |H^{*}z|^2 + |H^{*}m_c|^2 - 2R\{z^{+}HH^{*}m_c\} \\
&= p + b_c - 2R\{z^{+}HH^{*}m_c\}
\end{aligned}
\tag{5.9}
$$

where H is a diagonal matrix with h along its diagonal, R denotes real part, $p = |H^{*}z|^2$ is the energy of the transformed input, $b = |H^{*}m_c|^2$ is the energy of transformed class mean m_c and $HH^{*}m_c$ is the effective filter for class c.

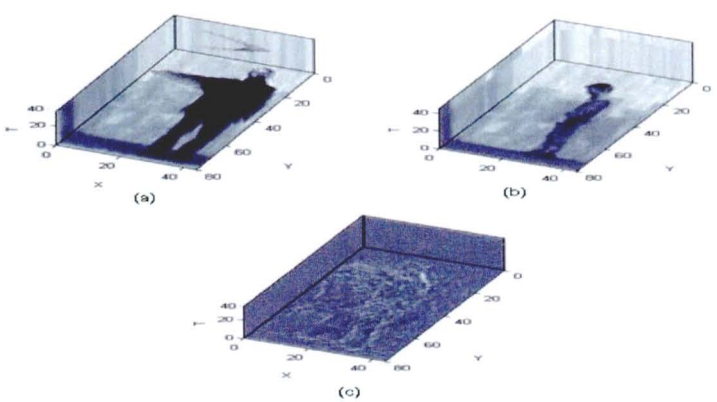


Figure 5.3: The simplest case of 2-class ST-DCCF filter. (Above (a),(b)) sample action volumes of wave and bend action (Bottom) A synthesized ST-DCCF transformation for two classes. (c) is visually ambiguous due to encapsulation of many training samples.

For shift-invariant distance calculation, we are interested in the smallest value of d_c over all possible shifts of the target with respect to the class references. For simplest case of only two classes (c_1, c_2), we get distances, (d_{c_1}, d_{c_2}) and input sequence is assigned to class c_1 if $(d_{c_1} < d_{c_2})$ and to class c_2 if $(d_{c_2} < d_{c_1})$. In this way, action class label attached to the found class (action class with the smallest distance) is assigned to the query action. A simple case of 2 class action ST-DCCF filter is presented in figure 5.4.

5.2 Action Representation

We represent action sequences with the creation of spatio-temporal volumes by concatenating the frames of a single complete cycle of an action. We begin the process of training the Action ST-DCCF filter from the training action sequences. We compute the temporal derivative of each pixel resulting in a volume for each training sequence. Following the construction of the spatio-temporal volumes for each action in the training set, we proceed to represent each volume in the frequency domain by performing a 3-D FFT operation [117] where 3D-FFT operation for action volume $a(x, y, t)$ is given by:

$$\mathbf{A}(u, v, w) = \sum_{t=0}^{T-1} \sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} a(x, y, t) exp(-j2\pi(\frac{uv}{X} + \frac{vy}{Y} + \frac{wt}{T}), \tag{5.10}$$

where $\mathbf{A}(u, v, w)$ is the resulting volume in frequency domain, X is the number of columns, Y is the number of rows and T is the number of frames of the volume.

5.2.1 ST-DCCF for Vector Value data

ST-DCCF can be used with scalar data (e.g., intensity values, temporal derivative) as well as vector value data (e.g. optical flow) but the process of synthesizing a filter on vector value data can not employ traditional Fourier Transform for scalar data. The class of Fourier transform for vector value data is refereed as "Clifford Fourier Transform". A similar approach has been used by [29]. Elements belonging to this algebra are known as multi-vectors. The Clifford Fourier Transform for multi-valued functions in 3D is defined as:

$$\mathfrak{F}F(u) = \int F(x) exp(-2\pi i_3 \langle x, u \rangle) |dx| \tag{5.11}$$

where i_3 represents a complex number in Clifford algebra, such as $i_3 = e_1 e_3$ and $i_3^2 = -1$. the inverse transform is given by:

$$\mathfrak{F}^{-1}F(x) = \int F(x) exp(-2\pi i_3 \langle x, u \rangle) |dx| \tag{5.12}$$

The rest of filter synthesis remain same as in scalar data case.

5.3 VIEW-DCCF : View-invariant Space-time distance classifier correlation filtering

Temporal information is an important attribute of action instance. To visualize a distance transform for action instances, we need to extend it in spatio-temporal sense. We name this extension as space-time distance classifier correlation filter.

The problem of action recognition can be considered as multi-class discrimination problem by simultaneously including all the classes to be separated. By applying global transformations to the input data, inter-class distance can be increased while making

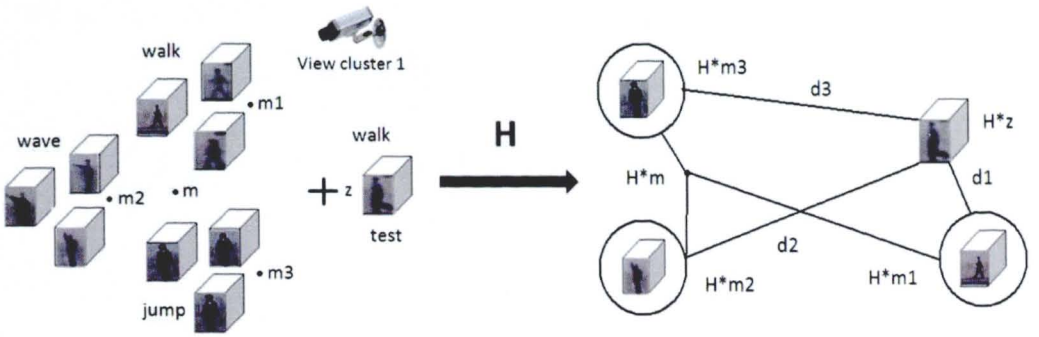


Figure 5.4: The schematic diagram of VIEW-DCCF filtering for single viewpoint showing Transformation H which increases inter-class distance while simultaneously making each class more compact. It shows that after the transformation, distance d1 is the smallest making test action closest to walking class. Similar transformation are required for each view cluster.

classes as compact as possible. To achieve this objective, correlation can be visualized as a linear transformation and filtering process can be mathematically expressed as multiplication by a diagonal matrix in the frequency domain. For a correlation filter to be used as transform, we require that instances of different classes become as different as possible after filtering. Then, shift-invariant mean square error distances can be computed between the filtered class instance and the transformed references of different classes and input is assigned to the class to which the distance is the smallest. Distance classifier correlation filter is a filter with the above mentioned objective.

5.3.1 Filter Theory

Let A_{ik} be the d-dimensional column vector containing 3D FFT (Fourier Transform of space-time action volume) of the ith training action volume of kth class, $1 \leq i \leq N$ and $1 \leq k \leq C$. Let m_k be the mean 3D-FFT of class k such that:

$$m_k = \frac{1}{N} \sum_{i=1}^{N} A_{ik}, 1 \leq k \leq C \tag{5.13}$$

Maximizing Spectral Separation : Under a linear transformation H, the differences between the means of any two classes i and k can be written as:

$$v_{ik} = H^*(m_i - m_k) \tag{5.14}$$

Taking the expectation of the elements of v_i over all frequencies yields

$$\bar{v}_{ik} = E_k\{v_i(k)\} \cong \frac{1}{d} h^+(m_i - m_k) \tag{5.15}$$

The quantity in eq. 5.15 is a measure of the *spectral separation* between classes i and k over all frequencies. We want $|\bar{v}_{ik}|$ to be large.

Taking all possible pairs of classes into consideration, we define *spectral separation* (SS) criterion in proposed filter as below:

$$
\begin{aligned}
SS(h) &= \frac{1}{C}\sum_{i=1}^{C} \mid h^{+}m - h^{+}m_i \mid^2 \\
&= h^{+}\{\frac{1}{C}\sum_{i=1}^{C}(m-m_i)(m-m_i)^{+}\}h \\
&= h^{+}Th
\end{aligned}
\tag{5.16}
$$

where $T = [\frac{1}{C}\sum_{i=1}^{C}(m-m_i)(m-m_i)^{+}$ is a $d \times d$ non-diagonal matrix of rank $\leq (C-1)$ and $m = \frac{1}{C}\sum_{i=1}^{C} m_i$ is the mean of the entire dataset. If SS is maximized by appropriate choice of h, the average content of the classes will differ greatly and become well separated.

Minimizing Within-Class Similarity : At the same time, we want to improve intra-class compactness by creating balance between low and high frequency components by *Within-class similarity measure* (WS) given as:

$$
\begin{aligned}
WS(h) &= \frac{1}{C}\sum_{k=1}^{C}\{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{A}_i h^* - (1-\alpha)\overline{\mathbf{A}}_k h^*)^{+}(\mathbf{A}_i h^* - (1-\alpha)\overline{\mathbf{A}}_k h^*)\} \\
&= \frac{1}{C}\sum_{k=1}^{C}\{h^{+}(\frac{1}{N}\sum_{i=1}^{N}(\mathbf{A}_i - (1-\alpha)\overline{\mathbf{A}}_k)^{*}(\mathbf{A}_i - (1-\alpha)\overline{\mathbf{A}}_k))h\} \\
&= h^{+}S_k h
\end{aligned}
\tag{5.17}
$$

where $S_k = (\frac{1}{N}\sum_{i=1}^{N}(\mathbf{A}_i - (1-\alpha)\overline{\mathbf{A}}_k)^{*}(\mathbf{A}_i - (1-\alpha)\overline{\mathbf{A}}_k))$, α is emphasis parameter, its value ranges from 0 to 2 and is set imperially. By controlling its value, the filter can be prevented from the biased treatment of low frequency information. Other objective is to maximize spectral separation (SS) and minimize Within-class similarity measure (WS), thus maximizing the ratio $R(h)$ as:

$$
R(h) = \frac{SS(h)}{WS(h)} = \frac{h^{+}Th}{h^{+}Sh} \tag{5.18}
$$

The solution that maximizes this ratio is given by:

$$
h = S^{-1}T \tag{5.19}
$$

In multi-class setup, the optimum solution is the dominant eignvector of $S^{-1}T$ with the largest eignvalue. We refer optimum h as the space-time distance classifier correlation filter or Action-DCCF. This filter deals with entire correlation space and not just one point at the origin.

5.3.2 View-invariant Action Classification

To deal view variations, we divide the action training set into view clusters where every view cluster contains data in range of certain view range or specific viewpoint. For every view cluster, we design Action-DCCF filter as described in the previous section. It encapsulates all action classes within a view cluster. For testing, we rely on using a shift-invariant mean square distance measure of test sequences from other action classes in a view cluster. For shift-invariant distance calculation, we are interested in the smallest

value of d_k over all possible shifts of the target with respect to the class references. The algorithmic steps has been described in Algorithm presented below:

Algorithm 3 Action-DCCF algorithm

procedure VDCCF($fftVols$, $nClass$, $nViews$)
  for $i \leftarrow 1, nView$ **do**
    for $j \leftarrow 1, nClass$ **do**
      $Cm \leftarrow$ get mean of each $fftVols$
      $Vm \leftarrow$ get variance of each $fftVols$
    end for
    $Mmean \leftarrow$ get overall mean of all classes
    $Mvar \leftarrow$ get overall variance of all classes
    $d \leftarrow$ size of single fftVolume in $fftVols$
    $H \leftarrow zeros(d+1, nClass)$
    $h \leftarrow$ get DCCF transform from $Eq.5.19$
    $H(1:d, 1) \leftarrow h$
    for $k \leftarrow 1, nClass$ **do**
      $b \leftarrow$ get class constant for class k from $Eq.5.20$
      $f \leftarrow$ get effective filter for class k from $Eq.5.20$
      $H(:, k+1) \leftarrow [f; b]$
    end for
    $Vdccf \leftarrow$ concatenate all H
  end for
end procedure

procedure DETECTA($QfftVol$, $nClass$, $nView$, $Adccf$)
  for $i \leftarrow 1, nView$ **do**
    $fptr \leftarrow$ pointer to all Action-DCCF
    $h \leftarrow Vdccf(1:d, fptr)$
    for $k \leftarrow 1, nClass$ **do**
      $H(k) \leftarrow$ get effective filter for each class
      $b(k) \leftarrow$ get class constant for each class
      $g(k) \leftarrow real(ifft3(QfftVol. * conj(H(k))))$
      $d(k) \leftarrow$ calculate distance from each class using $Eq.5.20$
      $D \leftarrow$ sort calculated distances from each class
      $d1 \leftarrow D(1), d2 \leftarrow D(2), r(i) = d1/d2$
      $S(k) \leftarrow$ calculate score for each class using $Eq.5.21$
    end for
  end for
  $DetectA \leftarrow$ classlabel with max-score in all clusters
end procedure

Given a query action input q, we determine its distance d_k (shift-invariant mean square error) for class k in a view cluster as:

$$\begin{aligned} d_k &= |H^*q - H^*m_k|^2 \\ &= |H^*q|^2 + |H^*m_k|^2 - 2R\{q^+HH^*m_k\} \\ &= p + b_k - 2R\{q^+HH^*m_k\} \end{aligned} \tag{5.20}$$

where R denotes real part, $p = |H^* q|^2$ is the energy of the transformed input, $b = |H^* m_k|^2$ is the energy of transformed class mean m_c (also known as *class constant*) and $HH^* m_k$ is the *effective filter* for class k.

Calculation of Action Similarity Score : We calculate similarity score (S) for each class. This similarity score is based on two calculations named within-cluster distance ratio (W) and class vote (V), both in range [0 - 1]. A within-cluster distance ratio (W) is computed for each Action-DCCF as ratio of the smallest distance to next (2nd) minimum. Smaller distances show better matches (ideally zero if there is an exact match with one of the classes within a view cluster) while larger ratios indicate greater ambiguity (the ratio is 1 when distances to both classes are equal). Its value may be different for different view clusters. Class vote (V) is the count of winning, vote counter for success of a class in all given view clusters. Its value is equal to 1 only if the respective class is the winner (gets the minimum distance score from query) in the respective view cluster else its value is 0 (loser). The similarity score S for each class k is calculated as:

$$S = \sum_{c=1}^{M} V(c) - W(c) \tag{5.21}$$

where c is cluster ID and M is number of view clusters. Finally, the label of the class with maximum score is assigned to the query action.

5.4 Experimental Results and Discussion

A comprehensive set of experiments are performed on two well-known human action data sets. The data sets represent actions performed both in constrained and unconstrained settings and represent different set of challenges for recognizing actions. In this chapter, we have used temporal derivatives for ST-DCCF filter synthesis but other different data representations like optical flow and spatio-temporal regularity flow can be used instead. The reason for using a simpler data representation is to get real information about improvement in performance of extracting discriminative information.

5.4.1 Dataset and Experimental Setup

The data sets used for our experimentations include KTH [27] and UCF Sports Action [29]. First data set is well known data sets and present controlled experimental settings, and therefore, can be used to benchmark our algorithm against existing algorithms. The second dataset is UCF Sports Action dataset which is relatively challenging due to its unconstraint settings.

KTH Dataset: This data set contain 600 low resolution (160 × 120, 25fps) video sequences containing six action categories: walking, running, jogging, boxing, clapping and waving. In total, there are 100 video sequences for each action performed by 25 different actors. Every actor performs each action four times in four different backgrounds. We trained ST-DCCF filter for 500 video sequences and used rest of 100 videos for testing purpose. The sample detected action instances for KTH dataset are shown in figure 5.5

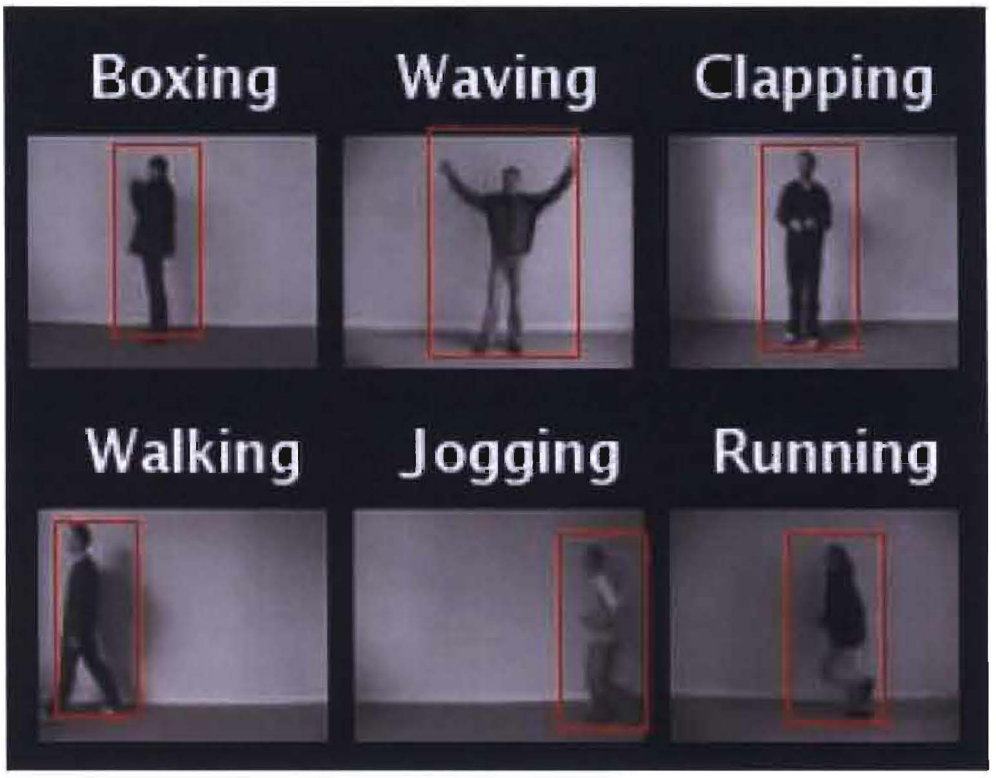


Figure 5.5: The detected action classes in KTH action dataset which include 6 action categories. This dataset is quite well known as a benchmark.

UCF Sports Action Dataset: This dataset contain 197 video sequences with resolution of 720×480. This dataset is really challenging as it contains actions performed in presence of clutter, interacting objects, shaky camera motion and captured from arbitrary viewing angles. We used off-the shelf VirtualDub Deshaker [120] for removing shaky motion. In addition, to negate background interference, we applied background substraction as pre-processing for this dataset. The actions include 10 action classes of running,walking, diving, kicking, high-bar,lifting, skating, swinging, horse riding and golf actions. We used 180 videos for training and remaining videos for testing purposes. The sample detected action instances for UCF sports action dataset are shown in figure 5.6

Figure 5.6: The detected action classes in UCF-Sports Action dataset which contain collection of broadcast sports action videos of 10 action classes including running,walking, diving, kicking, high-bar,lifting, skating, swinging, horse riding and golf actions.

5.4.2 Performance Comparison

Recognition is performed in leave one out cross validation (LOOCV) setting. Each action video is used as a query once and the best matching video is selected using the smallest value of distance described above. Action label of the best matching class is assigned to the query video. Comparison against the existing techniques is performed in terms of recognition accuracy. Recognition is performed and confusion and distance matrices are displayed.

First confusion matrix shows results obtained using ST-DCCF filter for KTH dataset. We obtained average recognition accuracy of 93.16% for KTH action dataset. The similar nature of (jogging, running) and (clapping, boxing) is one cause of their lower performance. The clear defining boundary between such actions is difficult to visualize. The rest of actions show higher recognition rate. Performance comparison with competitive techniques is shown in Table 5.1 in terms of average recognition accuracy.

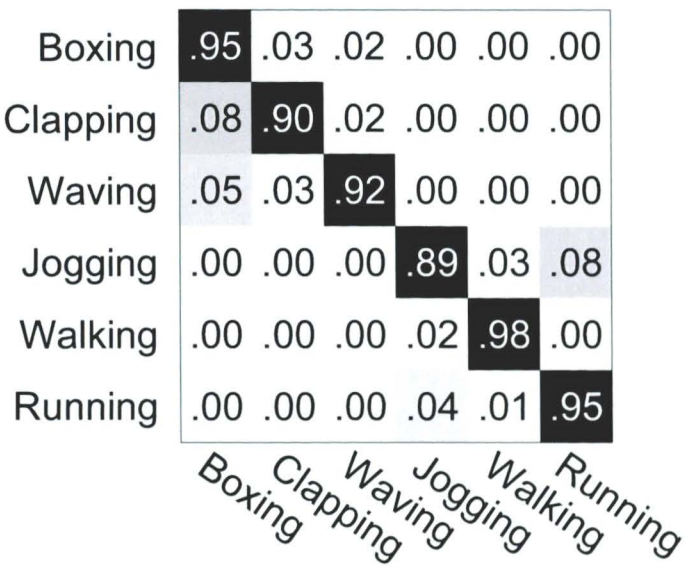


Figure 5.7: The Confusion Matrix for KTH Action dataset with recognition accuracy (93.16%) for actions, 1-boxing, 2-clapping, 3-waving, 4-jogging, 5-walking and 6-running .

The pair of recognition matrices show results for UCF Sports action dataset which shows recognition accuracy is 74.44% for UCF sports dataset which is improved compared to 69.44% presented in [29]. Most of the action classes show higher recognition results except diving, lifting and pole-vaulting. One explanation is their mix-up with each other due to similar motion patterns. Another reason is the nature of their complex dynamics which are difficult to comprehend.

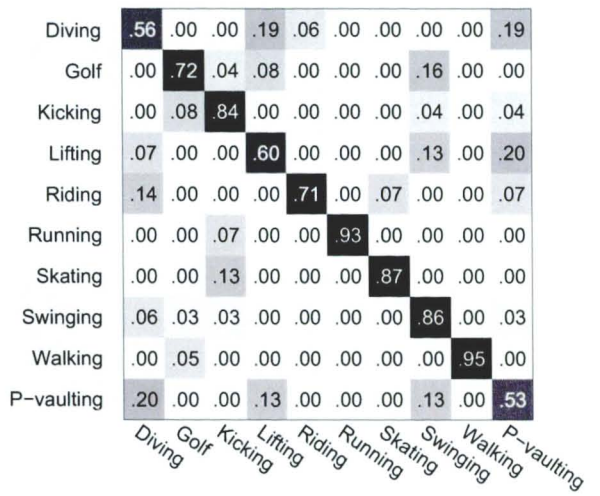


Figure 5.8: The Confusion Matrix for UCF Sports Action Dataset (accuracy 77.66%) for actions, 1-diving, 2-golf, 3-kick, 4-lifting, 5-riding, 6-running, 7-skating, 8-swinging, 9-walking and 10-pole-vaulting.

Method	Accuracy
ST-DCCF	93.16
Action MACH [29]	88.66
Cuboid Features [18]	81.17
Bag-of-words [115]	83.33

Table 5.1: Average recognition accuracy comparison for KTH dataset with other state of the art techniques.

5.4.3 Impact of Important parameters

An important parameter in action ST-DCCF filter is the value of β which effects the contribution of frequency information. The performance of action ST-DCCF filter increases as β increases within a small range. Then for a long range of β, the performance remains stable and then again shows a degradation in performance. The reason is that β controls the contribution of low-frequency components. As the value of β increases a certain range, the high frequency components would be overemphasized over the low-frequency components leading to degradation of filter performance. Table 5.2 indicates performance against three different values of β for KTH dataset. variations.

Values of β	Accuracy
0.5	92.8
0.9	93.0
1.5	90.1

Table 5.2: The effect of different values of β on average recognition accuracy comparison for KTH dataset.

5.5 VIEW DCCF Experimentation

A wide range of experiments are performed on publicly available multi-view action datasets. These are standard multi-view human action data sets and pose significant challenges to action recognition. In the next subsections, we give a brief description of datasets, experimental settings, recognition accuracy, performance comparison and discuss the effects of important factors.

Figure 5.9: An illustration of Different actions of IXMAS dataset by same actor. These actions include 1-check-watch, 2- cross-arms, 3- scratch-head, 4-sit-down, 5-get-up, 6-turn-around, 7-walk, 8-wave, 9-punch, 10-kick and 11-pick-up.

5.5.1 Multi-view Action Datasets

In this experimentation, we have used Multi-view WVU Action Dataset for initial investigation and used well-known Inria multiview IXMAS Dataset [15] for extensive validation of our approach.

WVU Multiview Action Dataset:

The dataset is collected as part of the research work on real-time human action recognition in a camera network. The multi-camera network system consists of 8 cameras that provide completely overlapping coverage of a rectangular region R (about 50 x 50 feet) from different viewing directions. It contained 11 actions, each performed by 10 actors three times and captured from eight different views (see Fig. 1 for a single action). These actions include nodding head, clapping,waving 1 hand, waving 2 hands, punching, jogging, jumping jack, kicking, picking, throwing and bowling. This dataset is available from [31].

IXMAS Action Dataset:

The Inria Xmas Motion Acquisition Sequences (IXMAS) is widely used dataset for view-invariant action recognition. It contained 11 actions, each performed by 10 actors three times and captured from five different views. We selected 1650 video sequences. These actions include check-watch, cross-arms, scratch-head, sit-down, get-up, turn-around, walk, wave, punch, kick and pick-up. The variations in viewpoints between five different cameras pose significant challenge.

Experimental Setup:

We use leave one out cross validation setting (LOOCV) to facilitate performance comparison. Action video is used as a query once and correlated with VIEW DCCF for each view cluster according to the detection algorithm described above.

The confusion matrices are calculated to demonstrate the recognition accuracy. For comparison for cross-view settings, we train ViewDCCFs for four cameras of IXMAS data and use fith for testing and train seven ViewDCCFs for WVU data and use eighth view for testing.

5.5.2 Performance Comparison

Performance comparison against the existing techniques is performed in terms of recognition accuracy. Recognition accuracy is calculated for each individual camera setting and average recognition accuracy for all cameras is displayed.

Average Recognition Accuracy:

The confusion matrix in figure 5.10 shows results for WVU action dataset while figure 5.11 shows results for IXMAS action dataset. This performance is calculated by averaging the performance accuracy of all camera viewpoints of WVU and IXMAS dataset. We achieve 89.86% for WVU and recognition accuracy of 82.9% for IXMAS. This performance for IXMAS is comparable to the existing techniques of [15], [71] and [87] which show average recognition accuracy of 72.7%, 71.2% and 82.8% respectively. These matrices show only best obtained results with vector data and optimal parameter setting. With temporal derivatives, we get average accuracy recognition of 84.92% for WVU and 78.5% for IXMAS.

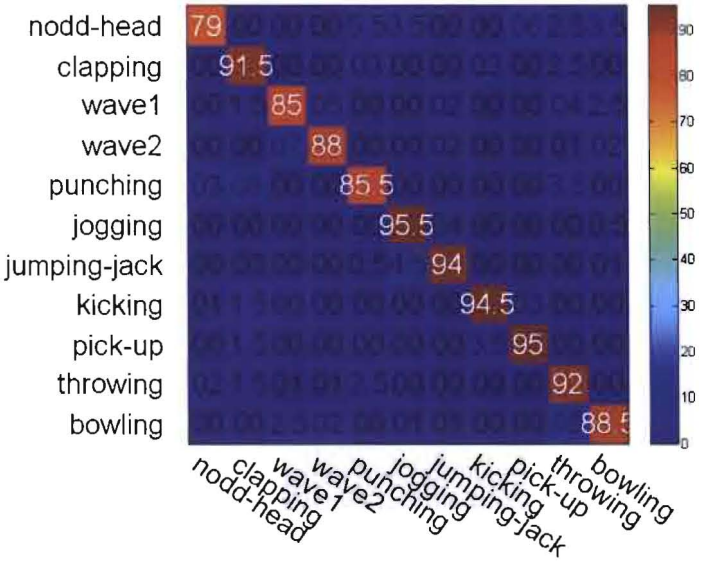


Figure 5.10: The Confusion matrix for WVU dataset which shows average recognition accuracy of all viewpoints (89.8%).

Method	Cam1	Cam2	Cam3	Cam4	Cam5	Avg
Ours	85.7	**81.6**	**84.9**	81.8	80.7	**82.9**
Ref.[15]	65.4	70.0	54.3.4	66.0	33.6	72.7
Ref.[71]	74.8	74.5	74.8	70.6	61.2	71.2
Ref.[87]	**86.6**	81.1	80.1	**83.6**	**82.8**	82.8

Table 5.3: Performance comparison with the existing techniques. Average recognition is the average performance for all five cameras.

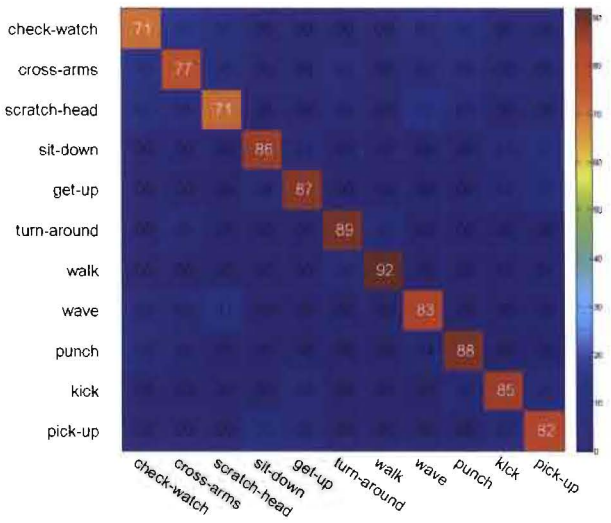


Figure 5.11: The Confusion Matrix for IXMAS Action dataset with recognition accuracy (82.9%).

Recognition Accuracy vs Camera Viewpoints:

In addition to average recognition accuracy, we have shown recognition against individual camera settings for the purpose of comparison to other exiting techniques. Table 5.3 shows performance comparison with the existing techniques.

Figure 5.11 shows recognition performance of individual actions vs five camera viewpoints. Most of the action classes show higher recognition. One notable thing is the excellent performance for actions captured from camera5 for which the camera setting is exactly above the actor and is sharp contrast to other camera settings. Majority of approaches show less performance for this camera viewpoint.

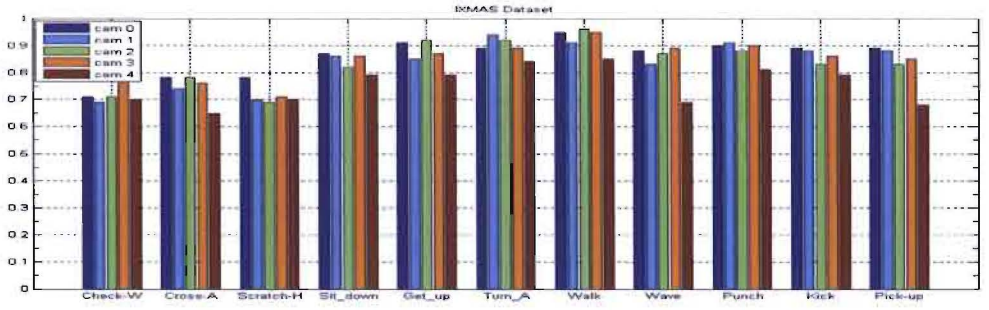


Figure 5.12: Recognition performance of IXMAS dataset from five different cameras with different viewpoints

5.5.3 Action Retrieval

To effectively test the performance of our approach, we have retrieved action instances from multi-view WVU dataset using proposed similarity score. We used seven views for training and eighth one for testing and displayed our result in figure 5.13.

ST-DCCFs (spatio-temporal distance classifier correlation filters) are trained for seven different views containing all action classes, nodding head, clapping, waving 1 hand, waving 2 hands, punching, jogging, jumping-jack, kicking, picking, throwing and bowling. Figure

5.13 shows that score is maximum for the respective query class, punch action for different view.

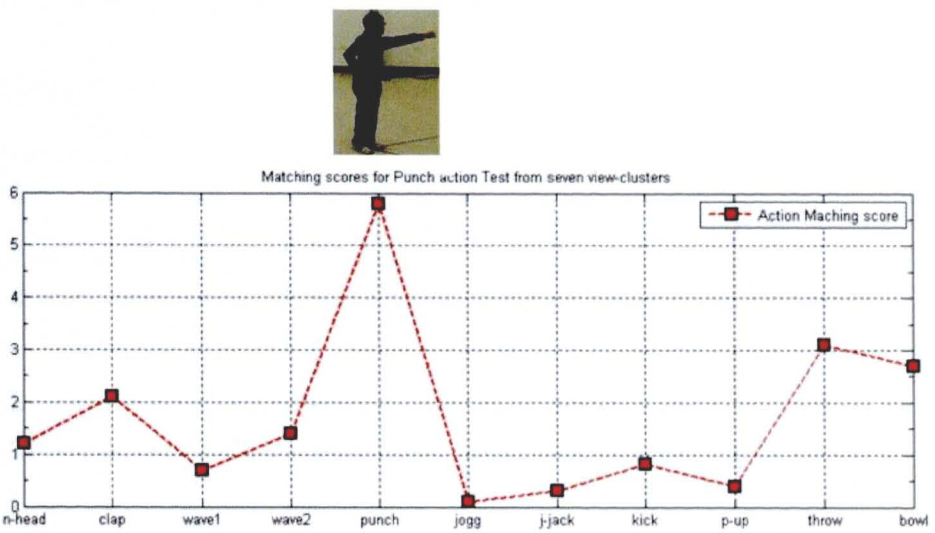


Figure 5.13: Action matching score for individual Action (Punch) in WVU dataset. View-DCCFs are trained for seven different views, each containing respective action classes, nodding head, clapping,waving 1 hand,waving 2 hands, punching, jogging, jumping-jack, kicking, picking, throwing and bowling. It shows that score is maximum for the respective query class, a punch action for different view not used in training phase.

5.5.4 Impact of Important parameters

An important parameter in VIEW-DCCF filter is the value of emphasis parameter α which effects the contribution of frequency information. The performance of VIEW-DCCF filter increases as α increases within a small range, then for a long range of α, the performance remains stable and then again shows a degradation in performance. The reason is that α controls the contribution of low-frequency components. As the value of α increases a certain range, the high frequency components would be overemphasized over the low-frequency components leading to degradation of filter performance. Table 5.4 indicates performance against three different values of α for WVU and IXMAS action datasets.

Values of α	Accuracy
0.5	86.1, 80.1
0.9	89.6, 82.8
1.5	84.2, 81.3

Table 5.4: The effect of different values of α on average recognition accuracy comparison for WVU and IXMAS datasets respectively.

5.5.5 Computational Time

On Intel (R) CoreTM 2 Duo system with 4GB RAM and un-optimized Matlab code, we get average run-time for testing video as 14.38 seconds compared to 18.65 seconds described in [29] which shows improvement in processing time.

5.6 Conclusions

In this chapter, we propose the concept of space time correlation filtering for matching human actions captured from different views. It is based on spatio-temporal correlation that is very useful for separating multiple classes. The proposed space-time frequency domain filter overcomes the weaknesses of existing correlation filters by presenting improvements which include: (i) support for view invariance action recognition framework, (ii) single space-time filter for multiple action classes in single view cluster decreasing computational overhead, (iii) shift-invariant distance providing more generalization for misaligned test sequences and (iv) improved intra-class similarity measure contributing balanced treatment of low and high frequency information. Experimentation has been performed on challenging action datasets which validates the utilization of our proposed VIEW-DCCF filter for view invariant action recognition.

However, in all previous chapters including this chapter, context of action sequence was not given enough consideration for action recognition. Although context is not always important for recognition but it becomes a valuable visual cue in unfavorable visual conditions like night vision. In next chapter, we investigate contextual action recognition by taking a challenging case study of night vision.

Chapter 6

Action Analysis using Contextual Associations

> Always design a thing by considering it in its next larger context - a chair in a room, a room in a house, a house in an environment, an environment in a city plan.
>
> ~ Eliel Saarinen (1873 - 1950)

The research work presented in this chapter has been published as:

1. Anwaar-ul-haq, I. Gondal and M. Murshed, *Contextual Action Recognition in Night-time video sequences, In Proc.* DICTA, Noosa Resort, 2011.
2. Anwaar-ul-haq, I. Gondal and M. Murshed, *Automated multi-sensor color video fusion for nightime video surveillance, In Proc.* IEEE ISCC, Riccione, Italy, 2010.
3. Anwaar-ul-haq, I. Gondal and M. Murshed, *A novel color image fusion QoS measure for multisensor night vision applications, In Proc.* IEEE ISCC, Riccione, Italy, 2010.
4. Anwaar-ul-haq, I. Gondal and M. Murshed, *SCARF: semi-automatic colorization and reliable image fusion, In Proc.* DICTA, Sydney, 2010.

Our visual world experience is captured in scenes where visual dynamics occur in rich surroundings, exhibiting in-between contextual associations. It indicates that contextual analysis and scene perception can provide powerful clues for recognizing visual events which seldom occur without any background or related objects. Human actions are spatio-temporal visual events and recognizing human actions is an important computer vision research problem. It has a large number of potential applications in the areas of visual surveillance, video retrieval, sports video analysis, human computer interfaces, and smart rooms. These applications also represent action contexts and contextual cues which can provide a priori knowledge for modeling action representations. In previous chapters, we restricted our discussion about action recognition in presence of viewpoint variations. In this chapter, we extend this investigation by including the context of actions. Due to

variety of possible contexts, we focus our investigation towards less explored and challenging context of night vision. We show how contextual association and their enhancement enhances action recognition performance.

The need for understanding actions in context is discussed by different researchers. Scene context is used for event recognition by [54] but it was only applied to static images. Recognizing actions in context is discussed by [53] which is formulated on bag-of-features framework and scene-action SVM- based classifier. It is focused on annotated actions in movies and uses script mining for visual learning. A similar approach [55] captures generic object based context by detectors and their descriptors are used as input for supervised learning.

Figure 6.1: A nighttime scenario of *waving* action captured by low light visible and infra-red sensors which presents visual information of complementary nature and lack certain visual information on individual basis.

More recently, modeling of scene and object context is discussed by [56] for Hollywood2 action dataset. All above approaches target action recognition in high-resolution action videos in movies. One typical benefit available to these approaches is the ease of finding visual interest points and detectors related to actors and their context.

In this work, we present actions in night vision scenario which offers real challenges due to extreme low light conditions. None of the above approaches discuss nighttime visual context and recognition of actions at nighttime. Mostly recently, human action activity recognition is discussed in [57, 58] which focus recognition in infra-red spectrum. However, these approaches ignore action contexts which is not properly captured by infra-red senors and can not be categorized as contextual action recognition approaches.

We argue that contextual action recognition is not possible using single sensor platform due to the limitations of individual sensor to grab all available visual information about the scene. This situation motivates the use of multiple sensors of complementary nature. A common multi-sensor night vision system uses infrared images in case of forward looking infrared cameras and low light images in case of low light visible cameras. The infrared images are maps of infra-red radiation emission which is partly governed by the temperature of the object. Therefore, such sensors prove good for perceiving hot targets in a busy background, seeing through fog, and monitoring paths through a cluttered forest. However, they are not much effective during thermal crossover periods at night or after long periods of rain, as well as capturing scenery such as trees, leaves and grass in natural scene. On the other hand, low light visible cameras are able to capture surrounding environment but most of the time fail to capture specific targets especially hot bodies like a person in camouflage. In addition, even in case when targets are not hiding, low light conditions make their observation obscure.

Figure 6.1 shows a nighttime scenarios of *waving* action captured by low light visible and infra-red sensors. While actor and his motion is quite visible in infra-red spectrum, it obscures the context. On the other hand, actor hands which represent *waving* action are not visible while scene context is relatively visible in low light visible spectrum.

The chapter is organized as follows: In next subsection, we describe the context enhancement of multi-sensor videos. Contextual action recognition of multi-sensor nighttime videos is presented in next section. Finally, experimental results and conclusion is presented.

6.1 Context Enhancement

The objective of context enhancement is to give day-like appearance to nighttime videos. Another justification behind this step is the limitation of individual sensor to present complete information about the scene. It is possible through video fusion of registered video streams from infra-red and visible or transfer of nighttime motion contents to daytime static background images of the same scene. Although the color information is not explicitly used in our method but it gives general look and feel of daylight images. Due to variability and original quality issues with different datasets, we used different ways for context enhancement. Here we discuss these approaches briefly.

6.1.1 Through Video Fusion

The video fusion is a process of visual information integration from a number of registered video sequences without loss of information and introduction of distortion. The goal of video fusion is to create a single enhanced video sequence from complementary video inputs that is more suitable for the purpose of human visual perception, action and context recognition. To achieve better quality and computation trade-off, we divided videos into low and high resolution and used different fusion approaches for their enhancement. For low visual quality videos, we used wavelet based fusion framework [145] in which approximation fusion is performed using principal component based fusion while absolute maximum rule is used for detail information. Figure 6.2 gives an illustration of video fusion results in this category.

Figure 6.2: An illustration of video fusion: (a) An infra-red video stream (b) A registered video stream from low light visible and (c) A fused video sequence from (a) and (b).

For high visual quality videos, we used automatic color transfer based video fusion [146] which enhances video context by color transfer from a source image. The illustration of this approach is given in figure 6.3.

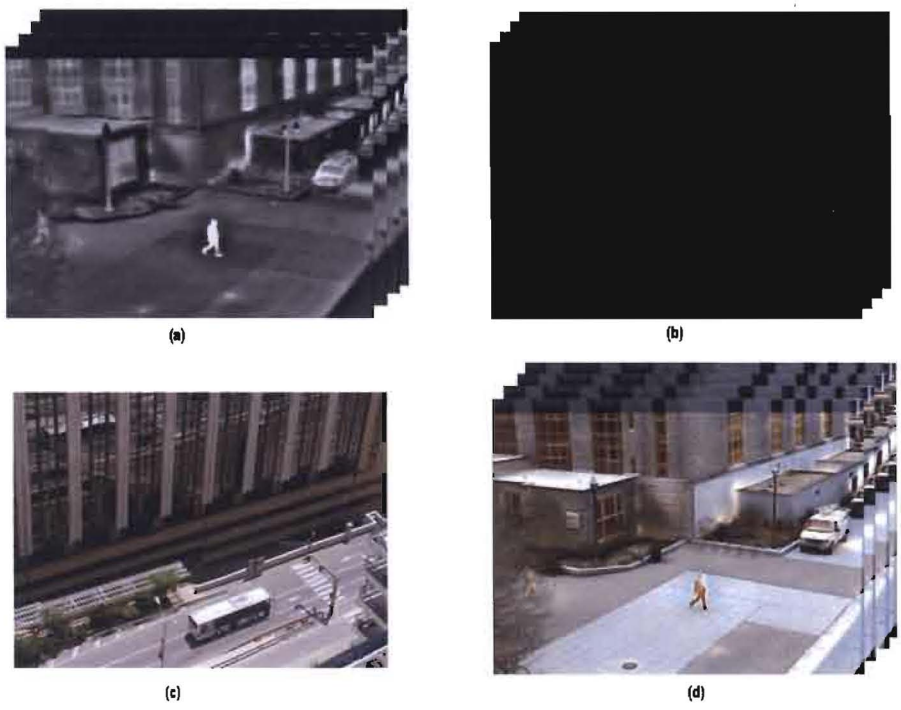


Figure 6.3: Color Transfer based Video Fusion: (a) An infra-red video stream (b)Registered video stream in low light visible spectrum (c) a source color image for color transfer and (d) a color fused video stream which contains structural fusion from (a) and (b) and color transfer from (c).

6.1.2 Through Dynamic Contents Transfer

We propose an alternative way of context enhancement for those videos for which we are not able to find nighttime visible counterpart. This is based on motion transfer from one video to another video and inspired from image blending [147]. The first video is infra-red containing actor motion and the second video is stack of static background images captured at daytime. The motion transfer method is as follows:

1- For each frame $f_1 \in F$, create a video sequence V in which the ith frame v_i is generated as given below:

2- Calculate optical flow between frame f_i and f_{i+1} to estimate temporal motion field m_i

3- Obtain the mask of moving pixels: $t_i = |m_i| > T$ where T is a threshold.

4- Treat q as background, f_i for foreground, t_i the mask of foreground, and apply Poisson blending [147] to obtain v_i.

The example of motion transfer based video generation is shown in figure 6.4. in which motion of action captured in infra-red video is transferred to a static scene video (a stack of static background images).

6.2 Contextual Action Recognition

Our contextual action recognition is based on action matching score calculated from action similarity score while penalizing contextual dissimilarities. The flow diagram of our system is shown in figure 6.5. It is based on information fusion and action similarity estimation. The detailed description of these steps is as under. Information fusion and frequency domain matching are core points of our recognition system.

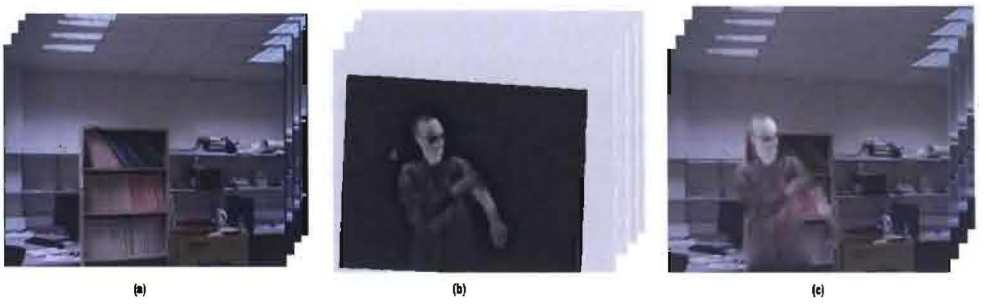


Figure 6.4: Dynamic Content Transfer: (a) A static scene video sequence (a stack of static scene frames), (b) An infra-red video stream with dynamic scene of actor performing an action and (c) Video sequence (a) after transferring motion contents from (b).

Background subtraction is used as a pre-processing step for dealing action and context. There are numerous ways to achieve background subtraction, we have used mixture of Guassians for background subtraction [148]. After building Guassian model of background, for a foreground frame, we can estimate for each pixel whether it belong to background or foreground by comparing mean and standard deviation of values at that pixel position.

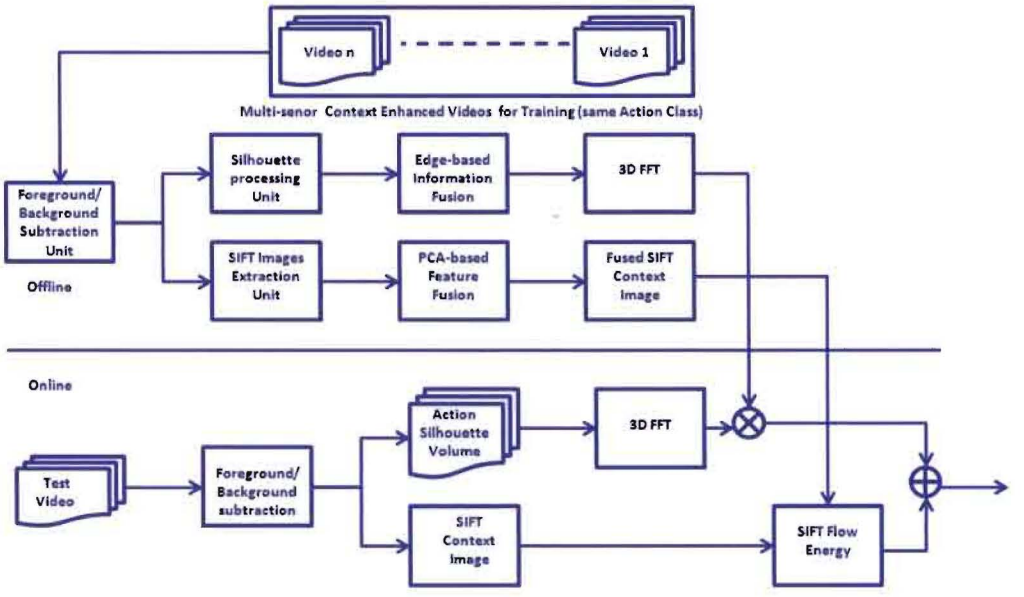


Figure 6.5: The flow diagram of our contextual action recognition system for context enhanced multi-sensor videos which includes background subtraction, information fusion, action and context similarity estimation. However, it shows offline training for only one action class and every action class needs to be trained separately in a similar manner.

6.2.1 Action Silhouette Processing, Information Fusion and matching

Background subtraction provides us foreground object information. We extract action silhouettes from these foreground frames for space-time information fusion related action instance.

For action silhouette fusion, actor silhouettes are first spatially aligned to a center position and then binary OR-based-fusion rule is applied to combine binary action silhouettes from all training examples related a single action class to create a single fused action silhouette volume. OR-based-fusion rule is applied at each frame level to form a video sequence of fused silhouettes to be represented as a fused action silhouette volume. An illustration of action silhouette volume is shown in Fig. 6.6.

A 3D FFT is applied to actor silhouette volume gives a frequency domain representation, 3D-FFT operation for action volume $a(x, y, t)$ is given by:

$$A(u, v, w) = \sum_{t=0}^{T-1} \sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} a(x, y, t) exp(-j2\pi(\frac{uv}{X} + \frac{vy}{Y} + \frac{wt}{T}), \tag{6.1}$$

where $A(u, v, w)$ is the resulting volume in frequency domain, X is the number of columns, Y is the number of rows and T is the number of frames of the volume.

For matching 3D FFT volumes, we first convert 3D FFT volumes into 1D column vectors. These FFT column vectors are extracted from text sequence and fused action volume which gives us initial action matching score as.

$$S_A(T, A) = Corr(F_A, F_T) \tag{6.2}$$

where S_A is similarity score for action class between test sequence T and fused action class volume A. F_A, F_T are 1D FFT column vector for fused action class volume and test sequence.

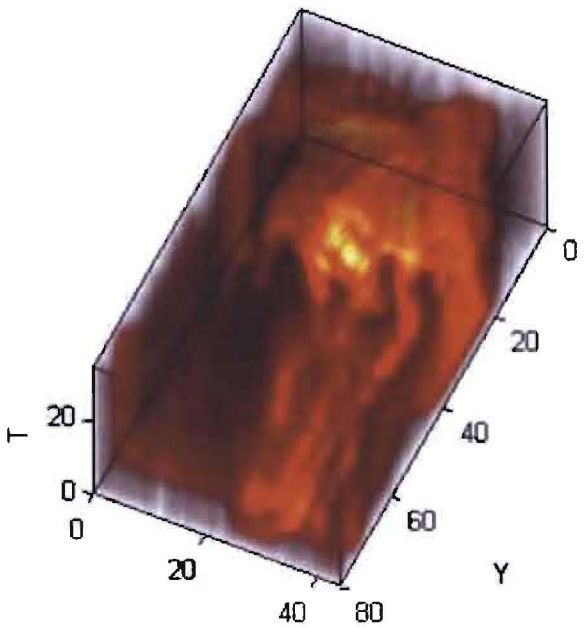


Figure 6.6: A fused action silhouette volume for wave1 action class based on OR-based-fusion rule applied at every frame instance from all training examples related wave1 action class.

6.2.2 Context Processing, Information Fusion and matching

We encapsulate contextual visual information in background images in SIFT context images. In a SIFT context image, a SIFT descriptor [100] is extracted at each pixel to characterize local image structures and encode contextual information.

We use median of five video frames from each video sequence (rather than using each frame) and call it context image. For all training videos in a single action category, we generate SIFT context images for each context image. Fig. 6.7 shows a visualization of SIFT context image. This visualization is obtained by mapping the first three principal components of each descriptor into the principal components of the RGB color space (i.e. the first component is mapped into $R+G+B$, the second is mapped into $R-G$ and the third into $R/2+G/2-B$).

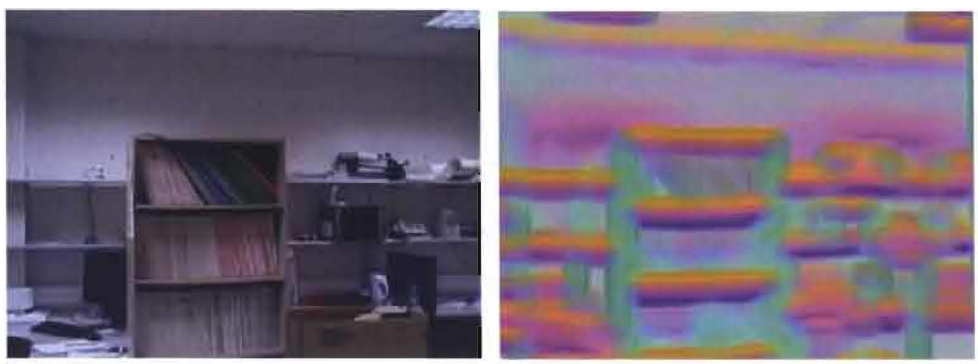


Figure 6.7: An example of SIFT context image. SIFT descriptors are computed on a regular dense grid (for each pixel in an image). This visualization is obtained by mapping the first three principal components of each descriptor into the principal components of the RGB color space .

Context in every training video is now represented by SIFT context image of dimension $h \times w \times 128$ where h, w stand for height and width respectively. For n training videos related an action category, we get n SIFT context images. This high dimensionality creates computational burden for any further processing. To deal with this problem, we use principal component analysis (PCA) based feature fusion.

For n SIFT context images, we build a feature vector at every SIFT position by concatenating their descriptors and apply principal component analysis to obtain a fused SIFT descriptor of 128 dimension. The combined result of feature fusion generates a single fused SIFT context image for all instances of single action class. We use this fused SIFT context image for matching.

We match SIFT context image from test video and fused SIFT context image using SIFT flow energy employed by [87] for SIFT flow, we rather use negative energy function as context matching score. For two SIFT context images T, F, test and fused context images, we define context matching score as:

$$
\begin{aligned}
S_w(T,F) &= -\{\sum_p \|T(p) - F(p+w)\|_1 \\
&\quad + \frac{1}{\sigma^2}\sum_p (u^2(p) + v^2(p)) \\
&\quad + \sum_{p,q\in\varepsilon} min(\alpha|u(p) - u(q)|, d) \\
&\quad + min(\alpha|v(p) - v(q)|, d))\}
\end{aligned}
\tag{6.3}
$$

where p, q are two neighboring pixels in ε which is 4×4 neighborhood, w is a flow vector at p. Other threshold parameters are $\sigma = 300, \alpha = 0.5, d = 2$. To speed up SIFT image matching, we used coarse-to-fine matching scheme described by [150].

6.2.3 Contextual Action Matching Score

The final contextual action matching score S is calculated by combining action similarity score and contextual matching score as:

$$S = S_A + \gamma S_w \tag{6.4}$$

where γ is contextual weight which controls influence of contextual cues in recognizing actions and would be discussed in the next section.

6.3 Contextual enhancement using Video Fusion

In this section, we extend the discussion about contextual enhancement using video fusion and propose automated color video fusion approach for night vision. In addition, we propose an objective quality index for objective evaluation of these approaches.

Our visual world is furnished with colors which aid in visual perception as human eye can perceive only 100 shades of gray comparative to more than 400 hues (dominant color) and about 20 saturation (degree of delusion) levels per hue [154]. The color information is badly affected at nighttime due to the absence of sunlight creating a natural obstacle for attaining color night vision as conventional camera model is based on processing of sunlight and its dispersion into different colors. Specially designed night vision devices like light intensifiers use star or moonlight to gather few photons, convert photons into electrons, amplify electrons and convert them back into photons to get visible light for capturing views of the night scenes. State of the art hardware approaches like fourth generation night vision devices can act even in very low light conditions but these systems are very expensive. On the other hand, as electrons are hurled against a phosphorus screen, a green color image is produced which is far from day like color appearance. An alternative and cost effective approach is the use of multiple sensors and fusion of captured nighttime imagery.

The operational requirement to fuse night vision imagery is due to the limitations of individual sensor to grab all available visual information about the scene [59]. A common multi-sensor night vision system uses infrared images in case of forward looking infrared cameras and low light images in case of low light visible cameras. The infrared images are maps of infra-red radiation emission which is partly governed by the temperature of the objects. Therefore, such sensors prove good for perceiving hot targets in a busy background, seeing through fog, and monitoring paths through a cluttered forest. However,

they are not much effective during thermal crossover periods at night or after long periods of rain and capturing scenery such as trees, leaves and grass in natural scene. On the other hand, low light visible cameras are able to capture surrounding environment but mostly fail to capture specific targets especially hot bodies like a person in camouflage. In addition, even in case when targets are not hiding, low light conditions make their observation obscure.

To solve this problem, image fusion is used which extracts meaningful information from complementary sensor images and combines visual information into a single output image. Over the years, several image fusion techniques are developed which vary in their complexity, robustness and quality. One major trend in image fusion research is to sacrifice complexity to gain quality. However, oppose to images, complexity criterion has more significance in video domain which is intended for real time use. Therefore, video applications do not encourage algorithmic complexity and require simple and efficient information fusion. Furthermore, to meet real time surveillance needs video representation is necessary which gives complete spatio-temporal visual information compared to limited spatial information presented by still images. The video fusion is a process of visual information integration from a number of registered video sequences without loss of information and introduction of distortion. The goal of video fusion is to create a single enhanced video sequence from complementary video inputs that is more suitable for the purpose of human visual perception, object detection and target recognition.

Color is another important requirement in addition to fusion but colorization of fused grayscale imagery is a daunting task. Most recently, various manual and semi-automatic colorization techniques have been reported in the literature to solve this difficulty. A highly cited work is colorization based on optimization [60] which needs user defined color scribbling. It proves to be an attractive method which based on the idea that neighboring pixels in space-time with similar intensities should have similar colors and requires neither precise image segmentation nor accurate region tracking. However, one shortcoming of this method lies in the requirement that input images are annotated with user defined color scribbles and thus lacks full automation.

Another popular work is colorization based on color transfer [61] using statistical analysis to impose one images color characteristics to another image. It uses a de-correlated color space $\ell\alpha\beta$ and swatches for color transfer from target color image. This technique has the same drawback that it requires manual selection of a color target image and swatches. In addition, color space conversions and swatches make additional burden in terms of complexity. Despite these shortcomings, above approaches have transformed the cumbersome work of manual colorization into semi-automatic colorization. Due to their successful application in colorization and color correction, these techniques are extended for colorizing night vision imagery [62, 181] presenting a software based approach to night vision offering a cheaper and reliable solution. Therefore, it is highly desirable that fully automated colorization should be introduced to facilitate real-time video processing for night vision applications.

In this chapter, we build upon the idea of [146], and propose a software based approach which overcomes above mentioned limitations by simultaneously fusing information from forward looking infra-red and low light visible sensors and introduce automatic colorization for context enhancement at nighttime. In addition, we restrict our colorization in RGB color space avoiding different color space conversions. At First, corresponding frames from complementary video streams are fused and pseudo-colorized using RGB color channel integration. Then, efficient color morphing technique is used in RGB color space avoiding any color space conversion. Automation is introduced by integrating source color image selection with contextual features and colorfulness characteristics. A prototype night vision

system named *SCENT* is developed based on proposed approach. The abstract visualization of our proposed system for color exploitation at night-time (*SCENT*) is presented in figure 6.9 which describes how original grayscale video sequence from two different modalities are simultaneously integrated and colorized into a colorful representation. Extensive experimentation is performed on different nightime datasets which comprise registered video streams from forward looking infer-red and low light visible sensors and performance is compared to state of the art approaches in terms of objective quality measures. Quality evaluation shows that our approach not only gives promising fusion and color quality but also proves to be efficient in terms of execution time.

6.4 System Architecture

In this section, we present the system architecture of our proposed night vision system, *SCENT* with a flowchart and briefly describe its functional components.The flowchart is shown in figure 6.9. The inputs to our system are registered video streams captured from infra-red and low light visible sensors. We have used already registered video streams filtered with median filter for noise removal.

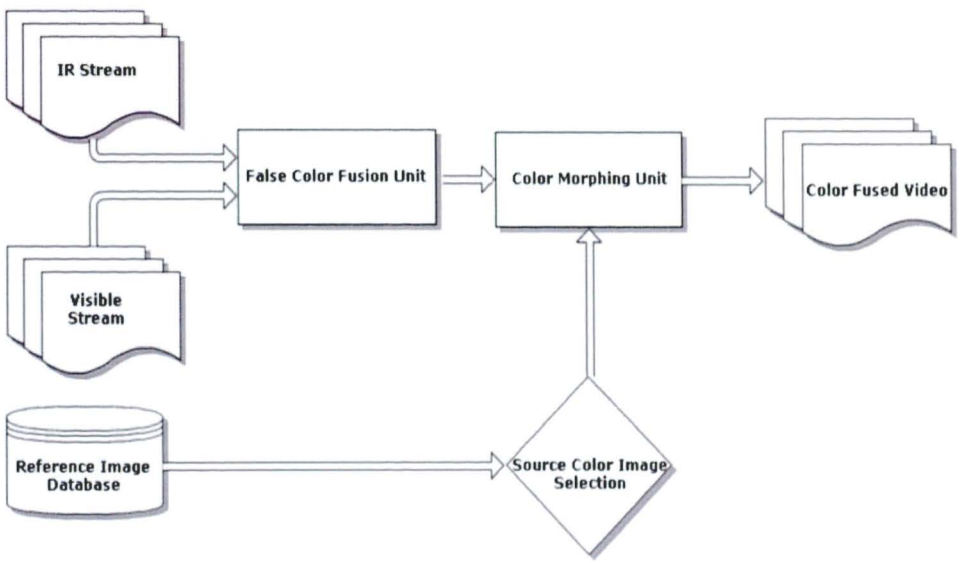


Figure 6.9: Flow chart of the color morphing based video fusion and colorization system: SCENT (system for color exploitation at nighttime).

These video streams are fed into false color fusion unit which is responsible for efficient fusion and false-colorization using *RGB* color channel fusion. This unit produces a fused and false colored video stream which is then fed into color morphing unit. The color morphing unit is the backbone of our system which transforms color distribution of false video streams according to a reference (source) color image. This source color image is selected from source color image selection unit which efficiently selects it from color image collection based on contextual features and colorfulness. Based on selected target color image, color morphing unit generates color fused video stream as final output of our system which resembles day-like color appearance. The detailed processing involved in these functional units is described in the next section.

6.5 The Proposed Video Fusion and Colorization Approach

In this section, we present algorithmic steps of our approach to apply natural day-like color appearance to grayscale nighttime video streams with color morphing. In addition, we describe the selection of color target image based on contextual association and colorfulness.

6.5.1 Fusion and colorization in RGB color space

Color transfer methods [61, 62, 181] use multiple color space conversions e.g. RGB to $\ell\alpha\beta$ color space conversion to get minimum correlation between color coordinate axes. This de-correlation is required to manipulate colors to individual color channels without changes into color distribution of other channels. For instance, in RGB color space, most pixels will have large values for the red and green channel if the blue channel is large which suggests that to change the appearance of a pixel's color in a coherent way, we need to modify all color channels in aggregation. It makes any color modification a difficult process and thus an orthogonal color space is required without correlations between the axes. Color channels in RGB color space are correlated which complicates the manipulation of individual color channels. To deal this problem, a color space, called $\ell\alpha\beta$ was proposed in [163], which minimizes correlation between channels for many natural scenes and is being used for color transfer between color images. The steps include many intermediate color space conversions like RGB to to device independent XYZ tristimulus values, XYZ to LMS, LMS to $logLMS$ and $logLMS$ to $\ell\alpha\beta$ transformations which involve many matrix multiplications increasing computation complexity of the original algorithm. Another disadvantage is the color contrast loss due to logrithmic transformation of $logLMS$. Therefore, we avoid color space conversions by restricting color transformation in RGB color space. In following subsections, we describe video fusion and colorization in RGB color space.

RGB Color Channel Fusion

The objective of this step is to generate a fused and false colored video stream from two grayscale video inputs. To achieve this objective, we integrate infrared and low light visible video streams in a meaningful way. We generate single RGB fused representation from sensor outputs, consisting of three channels, F_R,F_G,F_B. Frames from infra-red video stream are assigned to F_R while visible sensor output is assigned to F_G and F_B channels, respectively. This step is efficient enough to integrate visual information from two inputs and introduces false colorization as well.

$$F_R^n = F_{IR}^n, F_G^n = F_{VIS}^n, F_B^n = F_{VIS}^n \tag{6.5}$$

where n denotes frame number and IR, VIS stand for infra-red and visible inputs. False color video is generated in this step to get a color input for color transformations which are usually defined between two color inputs, the color target and color source. This step is illustrated in figure 6.10 which displays both input and output frames.

RGB Color Channel De-correlation

This step aims to attain de-correlation in RGB color channels for additional color processing without color space conversions. It can be achieved by eigen value decomposition of covariance matrices between RGB components of source and target farmes. We calculate mean and covariance matrices along RGB axis for both target and source (false-fused) frames. We denote $(\bar{r_t}, \bar{g_t}, \bar{b_t})$,$(\bar{r_s}, \bar{g_s}, \bar{b_s})$ as mean and C_t and C_s as covariance matrices for

(a) (b) (c)

Figure 6.10: RGB color channel Fusion: (a) infrared input, (b) low light visible input ,(c) pseudo-fused color output

target (false-color) and source (reference daytime color)frames. Eigen value decomposition can be used to further decompose the covariance matrices as: $C = U\Lambda U^{-1}$ where Λ is the diagonal matrix of eigenvalues of matrix U, having dimension $M \times M$. We can represent $\Lambda = diag(\lambda_R, \lambda_G, \lambda_B)$ where $\lambda_R, \lambda_G, \lambda_B$ are eigenvalues. The eigen values and eigenvectors are ordered. The mth eigen value corresponds to mth eigenvector. These eigenvectors are orthogonal to each other, de-correlated and can be used for further processing.

Color Transfer through Color Morphing

To give day-like color appearance to target (false-fused) video sequence according to color distribution of source (reference daytime color) image, we use color transformations and call this process as color morphing. It contains ellipse fitting to original color distribution of the target and its transformation according to source color distribution which generates similar color look and feel. This transformation includes translation, rotation and scaling applied to color distribution of target frames as defined below:

$$F_{final} = (T_s.R_s.S_s.S_t.R_t.T_t)F_t \tag{6.6}$$

where notation of $F_{final} = (R, G, B, 1)^T$ and $F_t = (R_t, G_t, B_t, 1)^T$ are RGB homogeneous coordinates of final output and target (pseudo-fused) frames while T, R, S stand for translation, rotation and scaling matrices defined below:

$$T_s = \begin{bmatrix} 1 & 0 & 0 & T_s^r \\ 0 & 1 & 0 & T_s^g \\ 0 & 0 & 1 & T_s^b \\ 0 & 0 & 0 & 1 \end{bmatrix}, T_t = \begin{bmatrix} 1 & 0 & 0 & T_t^r \\ 0 & 1 & 0 & T_t^g \\ 0 & 0 & 1 & T_t^b \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6.7}$$

$$R_s = \begin{bmatrix} U_s^{11} & U_s^{12} & U_s^{13} & 0 \\ U_s^{21} & U_s^{22} & U_s^{23} & 0 \\ U_s^{31} & U_s^{32} & U_s^{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, R_t = \begin{bmatrix} V_t^{11} & V_t^{12} & V_t^{13} & 0 \\ V_t^{21} & V_t^{22} & V_t^{23} & 0 \\ V_t^{31} & V_t^{32} & V_t^{33} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{6.8}$$

$$S_s = \begin{bmatrix} S_s^r & 0 & 0 & 0 \\ 0 & S_s^r & 0 & 0 \\ 0 & 0 & S_s^r & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, S_t = \begin{bmatrix} S_t^r & 0 & 0 & 0 \\ 0 & S_t^r & 0 & 0 \\ 0 & 0 & S_t^r & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{6.9}$$

where $V_t = U_t^{-1}$, $T_s^r = \bar{r_s}, T_s^g = \bar{g_s}, T_s^b = \bar{b_s}$, $T_t^r = -\bar{r_t}, T_t^g = -\bar{g_t}, T_t^b = -\bar{b_t}$ while $S_s^r = \lambda_s^r, S_s^g = \lambda_s^g, S_s^b = \lambda_s^b$ and $S_t^r = 1/\sqrt{\lambda_t^r}, S_t^g = 1/\sqrt{\lambda_t^g}, S_t^b = 1/\sqrt{\lambda_t^b}$.

Above transformations modify the color distribution of source (false-fused) frames according to color distribution of target image. This color morphing to false colored video

sequence results in better and natural color appearance like source which indicates the importance of reference(source) color image selection which we discuss in next subsection.

Automated Color Source Image Selection

The strength of color morphing lies in the fact that irrespective of contents of source color image, color look and feel is transformed from source to target. Our visual experience tells that context plays an important role in color distribution of the scene which implies that if context of source and target finds similarity, better results can be anticipated.

A research gap is discussed earlier about the lack of automation in previous approaches [61, 62, 181] about the automated selection of suitable source color image for color transfer. A straight-forward solution is the use of image retrieval framework based on recognition of objects and similarity detection. However, it involves computational complexity related object detection and recognition. Based on the requirement of our approach, we do not focus on finding exact structural match between source and target but on overall scene context. To attain this objective, we take advantage of global contextual features which estimate the shape or structure of the scene with few perceptual dimensions e.g spatial properties of the scene made by composite set of boundaries like walls, sections, ground elevation, slant of the surfaces. Generally, three level of abstraction are required to model the scene structure:

- subordinate level: analysis of local structure e.g. objects,
- basic level: similarity in shape,
- super-ordinate level: highest level of abstraction like scene category.

In our approach, we focus on super-ordinate level of abstraction and use global scene categorization based on GIST features[161]. The GIST feature is a vector of features f, where each individual feature f_k is computed as:

$$f_k = \sum_{x,y} w(x,y) \times |I(x,y) \otimes h_k(x,y)|^2 \tag{6.10}$$

Where $\otimes$ denotes image convolution, $\times$ presents pixel wise multiplication, $I(x,y)$ denotes luminance channel of input image, $hk(x,y)$ is filter from a bank of multiscale-oriented Gabor filters (6 orientation,4 scales) and w is a spatial window that would compute the average output energy of each filter at different image locations. The window $w(x,y)$ divides the image in a grid of 4×4 non-overlapping windows resulting in a descriptor of size $4 \times 4 \times 6 \times 4 = 384$.

Figure 6.12 illustrates the amount of context information preserved by $GIST$ features. It shows original scene, the output magnitude of multi-scale oriented filters on a polar plot and abstract $GIST$ descriptor. The average response of each filter is computed locally by splitting the image into 4×4 windows. Each different scale is color coded (red for high spatial frequencies, and blue for low spatial frequencies) with intensity proportional to the energy of each filter output. This illustration shows that GIST features provide a course description of the texture present in images and their spatial organization by preserving relevant information needed for categorizing scenes into categories which can potentially be used for establishing contextual association between source color image and target frames.

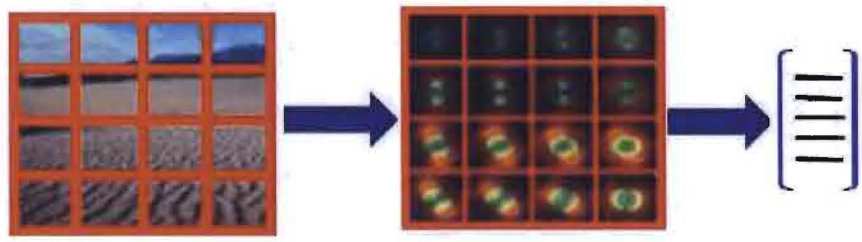


Figure 6.12: Illustration of global scene information encapsulated by respective GIST features.

The remaining challenge is to establish matching and retrieval framework efficiently. To reduce the dimension of original GIST descriptor, we use Locality Preserving Projections [134] to get a column vector of 100×1 for each descriptor. We use Euclidean-distance based nearest- neighbor approach for matching descriptors. To increase robustness, ratio of the nearest neighbor distance is utilized and any match for which the ratio of the nearest neighbor distance to the second nearest neighbor distance is greater than 0.6 is discarded. It helps in discarding many of the false matches, arising from background clutter. After matching the $GIST$ descriptors from source and target images, each source image is assigned a matching score that denotes the Euclidean distance between source and the target images. To refine our selection, we use another criterion, the colorfulness of source image which is defined as:

$$C_f = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2}, \tag{6.11}$$

where μ, σ are mean and standard deviations of the pixel cloud along two axes in opponent space, $rg = R - B, yb = 0.5(R + G) - B$, respectively. The final matching score is calculated as:

$$S_{matching} = wC_f + M_f \tag{6.12}$$

Where w is the weight parameter for colorfulness and M_f is matching score between contextual $GIST$ descriptors. The final color source image with largest matching score is selected as potential reference color image for color transfer. This selection process is embedding simplicity, efficiency and automates colorization using contextual information.

The algorithmic steps involved in video fusion and colorization can be summarized as:

Algorithm 4 Video Fusion and Colorization

1: $rgbColorChannelFusion$
2: $rgbColorChannelDe - correlation$
3: $ColorSourceImageSelection$
4: $ColorMorphing$

6.6 Objective Quality Evaluation

Visual quality assessment has subjective nature but the subjective image quality assessment requires a large number of images and human observers making it less suitable in real world applications. For this reason, objective image quality assessment techniques are being investigated and widely used with additional benefits like simplicity and fair sagacity. The goal of color fusion quality evaluation is to quantify the quality of information fusion and colorization in a precise and accurate manner. It falls in the category of blind

quality evaluation methods because of the absence of any reference image with optimal fusion and colors.

6.6.1 Color Similarity Measure (CSM)

This objective quality measure is based on structure and color similarity. Color similarity is computed by calculating similarity in hue, saturation and intensity. The original approach [167] assumes that initial inputs during information fusion are colored, so it cannot be used for quality evaluation of color transfer based techniques in its original form where initial inputs are grayscale which are colorized later on. To achieve this objective we modify this measure to serve our needs.

If two target images, color fused image and source color image are represented by A, B, F and S, the color similarity measure, CSM is defined as:

$$CSM(A, B, S|F) = \frac{CS(F, S) + SS(A, B|F)}{N}, \tag{6.13}$$

where

$$SS(A, B|F) = \frac{SSIM(A, F)}{SSIM(A, F) + SSIM(B, F)} \tag{6.14}$$

where $SSIM$ is structural similarity metric [172], N is normalizing factor with value calculated from sum of maximum values of CS (color similarity) and SS (structural similarity), while CS (range $0 - 1$) is defined as:

$$CS(a, b) = \alpha_1 * r(a, b) + \alpha_2 * IS(a, b), \tag{6.15}$$

where $r(a, b)$ represents the correlation coefficient of two color vectors with $\alpha_1 + \alpha_2 = 1, \alpha_1 > 0, \alpha_2 > 0, \alpha_1 > \alpha_2$. The similar coefficient of intensity similarity, IS is computed as:

$$IS(a, b) = 1 - \frac{|a_r + a_g + a_b - b_r - b_g - b_b|}{c}, \tag{6.16}$$

where $c = 3 * 255 = 765$. In this way, color and structural similarity is used to quantify the quality of color information fusion.

6.6.2 Color Fusion Quality Index ($CFOI$)

The proposed color fusion quality measure (CFOI), quantifies color information fusion considering structural distortion, blurring as well as color degradation and colorfulness of final fused image. In addition, it deals efficiency and reusability for use in diverse forms of color image fusion schemes, described in the introduction. Therefore, two different versions are given according to fusion framework with slight variations.

6.6.3 Case1: Color image fusion with original color sensors

First, we consider a general case when both input and output are colored as illustrated in figure 6.13. We divide our measurement into two phases: (1) color quality measurement and (2) fusion quality measurement and develop final metric by combining both measurements.

Figure 6.13: Color Image Fusion with Color Sensors

One difficulty associated with RGB color space is highly correlated color channels. It means that if we do some change in R channel, it would effect G and B channels as well. It requires de-correlation by orthogonalization of RGB color channels. For this purpose, we use eigenvalue decomposition of co-variance matrix calculated along RGB axis. Eigenvector are sorted according to eigenvalues and transformation matrix is formed by first three orthogonal eigenvectors.This transformation matrix is applied to original RGB color space to get de-correlated version. The algorithmic steps for this orthogonalization are presented in Algorithms 1. An alternative is the use of de-correlated YUV [154] or $\ell\alpha\beta$ [61] color space conversion. The result is three de-correlated channels denoted by Ch_l, Ch_{c1}, Ch_{c2} where first channel is luminance while other two are chrominance channels.

Algorithm 5 :Orthogonalization of RGB color space

1: **Input**:RGB color pixel cloud
2: Calculate co-variance matrix along RGB axis
3: Get eigen value decomposition of covariance matrix
4: Generate transform matrix from first three eigenvectors
5: Apply transform matrix to RGB correlated color space
6: **Output**:Three de-correlated channels (orthogonal to each other)

We denote two RGB color inputs as A, B, RGB fused output as F, SQ as structural quality, CQ as color quality, CF as colorfulness, FQ as fusion quality and our proposed color fusion objective index as $CFOI$. First, we calculate color quality C given as:

$$CQ(A, B, F) = w.C(A, F) + (1 - w.C(B, F)), \tag{6.17}$$

where for two image signals, x and y, the color quality C of x w.r.t y is :

$$C(x, y) = \sqrt{(SQ(x_l, y_l))^2 + (SQ(x_{c1}, y_{c1}))^2 + (SQ(x_{c2}, y_{c2}))^2} \tag{6.18}$$

and structural quality, SQ is defined as:

$$SQ(x, y) = \frac{4\sigma_{xy}\mu_x\mu_y}{(\sigma_x^2 + \sigma_y^2)(\mu_x^2 + \mu_y^2)} \tag{6.19}$$

For chrominance channels, we take original pixel values while for luminance channel we use gradient values calculated from Sobel operator [154]. It increases the robustness of quality measure against blurring effects. The relative colorfulness is calculated as:

$$CF(A, B, F) = \frac{C_n(F)}{w.C_n(A)) + (1 - w.C_n(B)}, \tag{6.20}$$

where C_n is given as:

$$C_n(A,B,F) = \frac{Cf}{Cf_{max}} \tag{6.21}$$

where $Cf_{max} = 109$ is maximum value of colorfulness defined in [169] and colorfulness Cf is given as: 7

$$Cf = \sqrt{\sigma_{rg}^2 + \sigma_{yb}^2} + 0.3\sqrt{\mu_{rg}^2 + \mu_{yb}^2} \tag{6.22}$$

where σ and μ are standard deviation and mean along two axis in opponent color space with $rg = R - B$ and $yb = 0.5(R + G) - B$, respectively. The values of colorfulness come in the range of [0 109] where 0 means no color and 109 means extreme colorful image.To make it compatible for use in our quality measure, we have normalize its values to lie in the range [0 1].

For information fusion quality evaluation, we employ mutual information which captures the common fused information.

$$FQ(A,B,F) = w.MI(A,F) + (1 - w.MI(B,F)), \tag{6.23}$$

where $MI(A,B)$ is the mutual information which is the amount of information gained about A when B is learned, and vice versa. $M(A,B) = 0$ if and only if A and B are independent. In case of two images F (fused) and B (input image), we can write it as:

$$MI(F,B) = \sum_{i_1=1}^{L}\sum_{i_2=1}^{L} h_{F,B}(i_1,i_2) log_2 \frac{h_{F,B}(i_1,i_2)}{h_F(i_1)h_B(i_2)} \tag{6.24}$$

where $h_{F,B}$ is the normalized joint gray level histogram of images F and B, h_F and h_B are the normalized marginal histogram of two images and L is the number of gray levels.

We also use local weighting procedure in color image fusion quality calculation. A local weight w is used in our measure which tells about the relative importance of one image compared to the other one. The value of w depends on the color fusion application. In case of visible difference in color significance, w is assigned a lower value. For instance, visual inspection of figure 6.13 shows that one of the input images(e.g $infra-redimage, imageA$), does not contain suitable colors. In this case local weight w would be assigned a lower value (e.g $w = 0.2$) which would automatically boost the importance of $imageB$. In case of no significant color difference, we utilize special frequencies of the images to calculate the value of w.

Spatial frequency is the measure of activity level of an image, and can be defined as:

$$sf = \sqrt{(rf)^2 + (cf)^2}, \tag{6.25}$$

where rf (row frequency) and cf (column frequency) are defined as:

$$rf(A) = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=2}^{N}[A(i,j) - A(i,j-1)]^2}, \tag{6.26}$$

and

$$cf(A) = \sqrt{\frac{1}{MN}\sum_{j=1}^{N}\sum_{i=2}^{M}[A(i,j) - A(i-1,j)]^2}, \tag{6.27}$$

If $sf(A)$, $sf(B)$ are spatial frequencies of input images, A and B then local weight λ is defined as:

$$w = \frac{sf(A)}{[sf(A) + sf(B)]} \tag{6.28}$$

Then, the quality of fused image, color fusion objective index $CFOI$ can be calculated from combining equations 6.17,6.20 and 6.23 as:

$$CFOI = \frac{CQ + CF + FQ}{3} \tag{6.29}$$

The $CFOI$ value remains in the range of [0 1]. When value approaches to 1: means better image fusion has taken place and vice versa.

6.6.4 Case2: color image fusion without color sensors

Another scenario is false color image fusion illustrated in Fig. 6.14 which is based on false colorization and color transfer from a target color image. The major difference is now inputs are grayscale while output is colored. Our proposed color fusion objective index $CFOI$ can be extended to this color fusion framework as well.

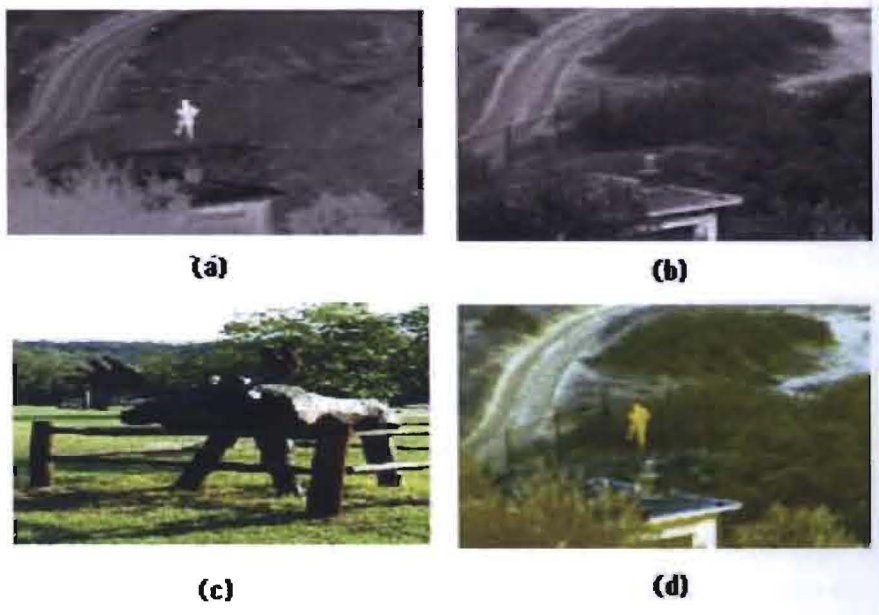


Figure 6.14: Color Image Fusion by Color Transfer

For this purpose, we consider following changes. Now T (target) and F (fused) are color images and color distribution of F should resemble that of T.Therefore, for color evaluation we consider T as reference image. Equation 6.17 is no more needed as Equation 6.18 can be used for color quality assessment of F w.r.t T as:

$$CQ(T,F) = \sqrt{(SQ(T_l,F_l))^2 + (SQ(T_{c1},F_{c1}))^2 + (SQ(T_{c2},F_{c2}))^2} \tag{6.30}$$

Similarly, equation 6.20 is modified to:

$$C(F,T) = \frac{Cf}{0.5(Ct+1)} \tag{6.31}$$

Therefore, modified color fusion objective index $CFOI$ in equation 6.29 can be written by combining equation 6.30, 6.31. The rest of calculations remain same.

Algorithm 6 :CFOI: Algorithmic Steps

1: **Input**:RGB image A,B,F,(T)
2: Calculate color quality of F (Equation: 6.17)
3: Calculated colorfulness of F (Equation: 6.20)
4: Calculate fusion quality of F (Equation: 6.23)
5: Calculate CFOI (Equation: 6.29)
6: **Output**:CFOI [Range 0-1]

6.7 Experimental Results and Discussion

6.7.1 Contextual Action recognition

A comprehensive set of experiments are performed on challenging human action data sets captured during nighttime. The video streams show actions performed both in constrained and unconstrained settings and represent different set of challenges for recognizing actions at nighttime. Infra-red and low light visible sensors are used in these video sequences and videos are registered before use.

6.7.2 Dataset and Experimental Setup

The video sequences used for our experimentations are collected from different sources. It includes 600 collected [151, 152, 153] and captured videos. The idea behind using different sources was to try different contextual setting of similar actions. This data set contains video sequences containing eight action categories: walking, wave1, wave2, stand-up, sit-down, clapping and pick-up by different actors. Five hundred videos are used for training and remaining 100 are used for testing.

The videos were recorded using two separate cameras. The IR camera is Raytheon Thermal IR-2000B and the visual camera is Panasonic WV-CP470. Alignment of the thermal and visual videos is done by manually selecting corresponding points in both views and computing a least-squared error fitting homography for each sequence. The infrared video frames are warped to align with the visual pixels. Pixels that are outside the infrared image are marked with value 255 (unknown).

The data primarily includes scenarios of short range surveillance type applications filmed under varying illumination conditions. Scenes include people (who are dressed in both civilian dress and camouflage, stationary, walking or running, or carrying various objects), vehicles, foliage and buildings/structures.

6.7.3 Action recognition

Recognition is performed in leave one out cross validation (LOOCV) setting. Each action video is used as a query once and the best matching video is selected using the contextual action matching score described above. Action label of the best matching class is assigned to the query video. Recognition is performed and confusion and distance matrices are displayed.

First confusion matrix shows results without contextual cues. We obtained average recognition accuracy of 84.87% for given action dataset. The second confusion matrix shows results with contextual cues. We obtained average recognition accuracy of 91.75% for given action dataset with performance gain of 7% percent. The clear gain in performance is visible in those actions which show interacting relationship with their context.

Most of the action classes show higher recognition results except stand-up, sit-down and pick-up. One explanation is their mix-up with each other due to similar motion

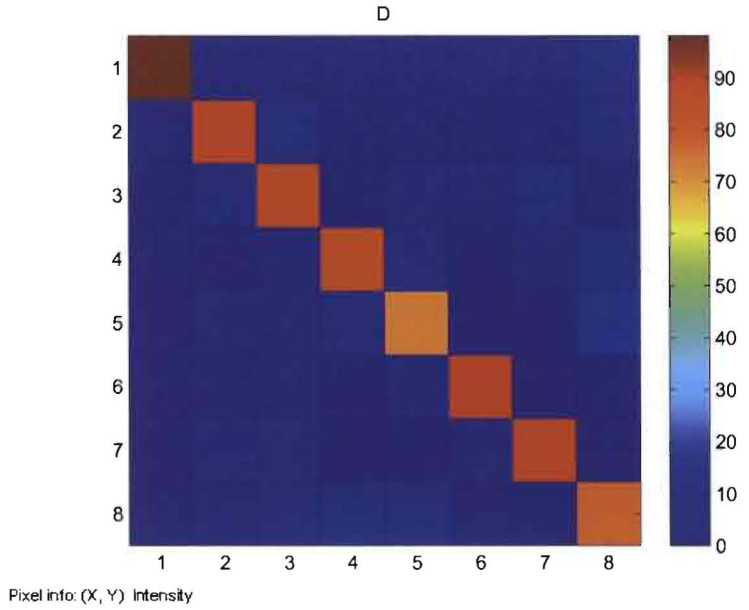


Figure 6.15: The Confusion Matrix without contextual cues (84.87%) for actions, 1-walking, 2-wave1, 3-wave2, 4-stand-up, 5-sit-down, 6-hands-up, 7-clapping and 8-pick-up
.

patterns. Another reason is the nature of their complex dynamics which are difficult to comprehend properly.

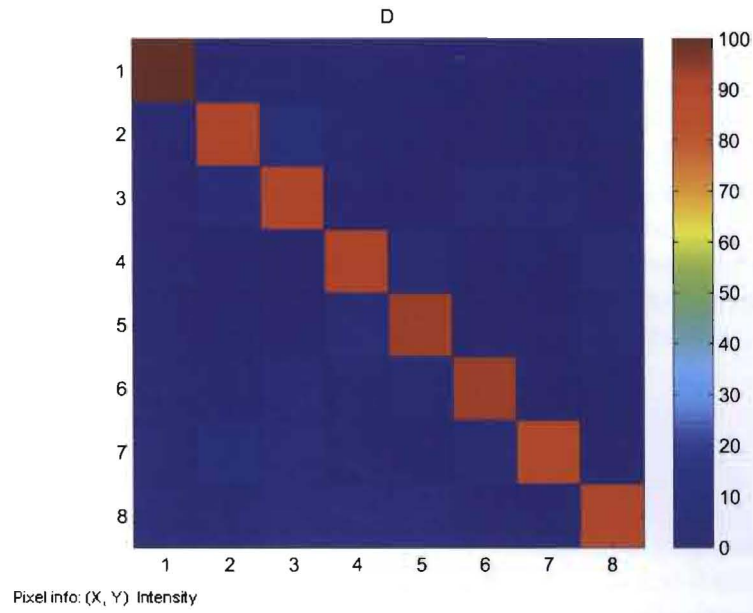


Figure 6.16: The Confusion Matrix with contextual cues (91.75%) for actions, 1-walking, 2-wave1, 3-wave2, 4-stand-up, 5-sit-down, 6-hands-up, 7-clapping and 8-pick-up

6.7.4 Automatic Contextual Action Annotation

We also present an interesting application of our work in this section. In addition to contextual action matching score, we can add context category using context matching

using same approach. It can help achieving automatic contextual action annotation of multi-sensor video data. An illustration of automatic contextual action annotation is presented below in figure 6.17.

Figure 6.17: An illustration of automatic contextual action annotation of multi-sensor video data in which action and its contextual scene is rightly recognized.

6.7.5 Contextual Enhancement Using Color Transfer: Experimentation

In this section, we present experimental results, quality evaluation, discussion and future work. We have used registered video datasets with *avi* video format. The datasets are collected from www.imagefusion.org provided by TNO Human Factor Research Institute, Netherlands, Octec Ltd. and Ohio State University, USA. The original videos are taken in Common Intermediate Format ($CIF, 352 \times 288$). We have collected arbitrary color source images from www.freefoto.com for color morphing. Un-optimized MATLAB code is used for implementation of our system. We conducted different experiments to validate our system. In this section, we give illustration of these experiments.

6.7.6 Illustrations for Visual Inspection

First, we generate video outputs by applying our approach on grayscale video inputs and present their results for visual inpection. Figure 6.18 presents first illustration which shows four frames from infrared and low light visible video streams and output video frames from our color fusion system as well. The scenario shows that a camouflage person is walking along the fence while part of building, path and trees are visible in the scene. Figure 6.18(a) presents thermal midwave $3 - 5\mu m$ version. Hot target like person is visible in frames but background imagery is cluttered and shows lesser details. The corresponding low light visible input is $(0.7 - 1\mu m)$ version in which it is difficult to distinguish a person in camouflage from the rather clear scene of background. Both inputs are incomplete and colorless. The final color fused output of our proposed system is presented in figure 6.18(c) which looks more complete, colorful and natural.

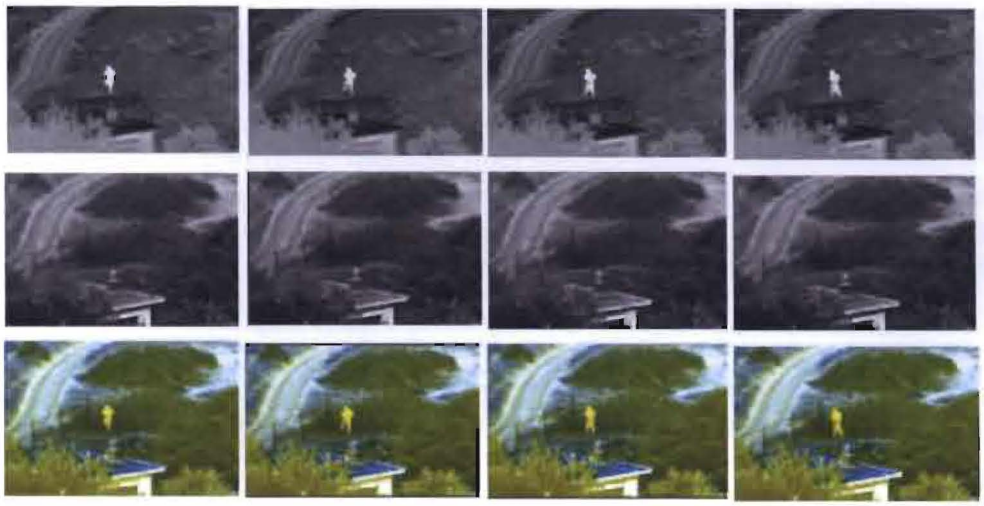


Figure 6.18: Four frames of video sequence and results (Scene A). (Above) grayscale frames from infra-red video sequence, (Middle) grayscale frames from low light visible video sequence and (Below) fused and colorized frames as result of SCENT .

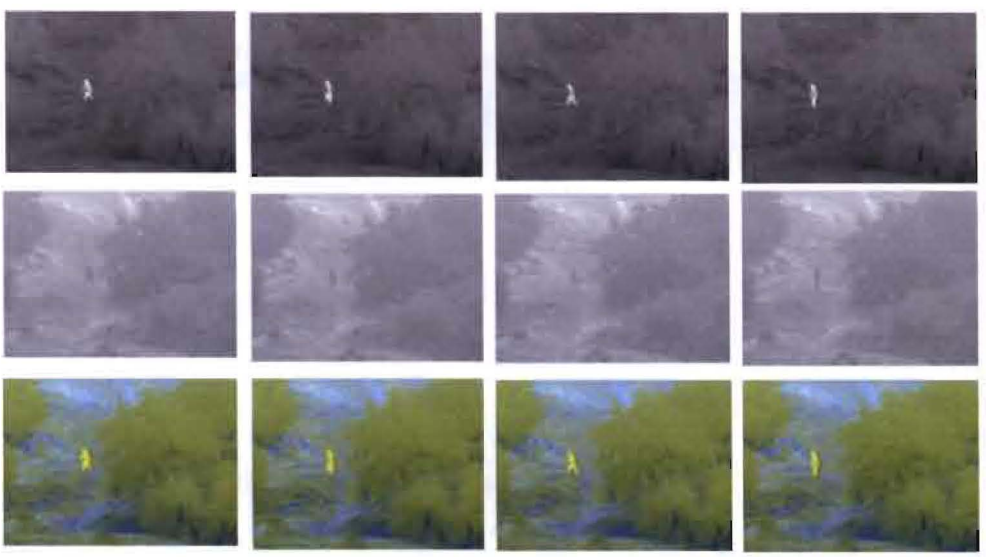


Figure 6.19: Four frames of video sequence and results (Scene B). (Above) grayscale frames from infra-red video sequence, (Middle) grayscale frames from low light visible video sequence and (Below) fused and colorized frames as result of SCENT.

A similar cluttered scene is presented in figure 6.19 in which a person is running through the jungle. Figure 6.19(a) presents thermal midwave $3-5\mu m$ version. The corresponding low light visible input is $0.7-1\mu m$ version and presented in figure 6.19(b). Although the person is present in both input videos inputs but background is much cluttered. Both inputs are noisy and colorless. The final color fused output from our proposed system is presented in figure 6.19(c) which looks complete, colorful and near natural.

6.7.7 Qualitative and Quantitative Comparison

We also present the subjective and objective comparison of our approach with state of the art color transfer based fusion methods. First, we present results for visual inspection of the reader and additionally present quantitative results for objective evaluation with color

similarity measure (CSM) and color fusion quality index(CFOI). We compare our results with Wang [182], Li [160], Toet [181], Waxman [158] which are colorization based image fusion methods. The first row in figure 6.19 shows original grayscale , infrared and visible and false color fused frames. The second row presents fused color representation from Wang [182], Li [160], Toet [181], Waxman [158] and our result. A clear difference is visible from the natural color appearance of our proposed method. Figure 6.21 and figure 6.22 show the graphical objective comparisons. Graphical results describe the performance of our proposed system against both measures (CSM and CFOI) outperforms previous approaches for different datasets.

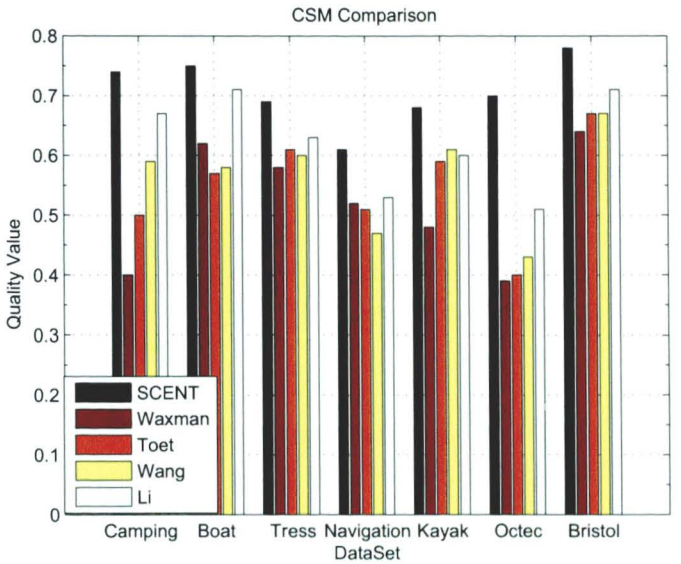


Figure 6.21: CSM comparison of our proposed system SCENT with other competitive techniques.The values of objective quality measure are in the range [0-1]. The large value are indication of better quality

6.7.8 Selection of Source Color Image

The selection of source color image is an important part of automation introduced in our approach. Therefore, we present its illustration in Figure 6.23. The query image is the false color frame from video sequence and sorted images are potential reference color images. The selection is based on contextual association and colorfulness. This experiment describes that contextual association plays an important role to select a daytime reference color image to be used for color morphing. For query image figure 6.23.(a) recision-recall curves are shown in figure 6.23.(b). For proposed contextual matching and structural matching [146] which shows that global contextual features prove better in selecting suitable source color image than mere structural similarity based selection.

6.7.9 The significance of SCENT

The above experimentation proves the significance of our developed video fusion and colorization system, *SCENT* (system for color exploitation at nighttime). With system specifications of Intel (R) Core (TM) 2 Duo CPU E8400, 3.00 GHz, 2.99 GHz and 3.43 GB RAM , our system is implemented using Matlab 7.0 un-optimized code. Table 6.1 shows CPU breakdown time for different involved sub-tasks in *SCENT* which gives us inspiration to port it in real-time using optimized C code which can provide significant

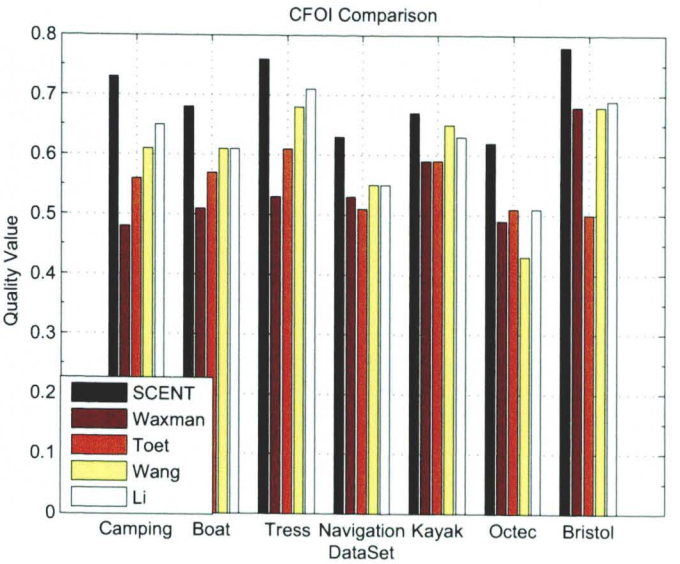


Figure 6.22: CFOI comparison of our proposed system SCENT with other competitive techniques. The values of objective quality measure are in the range [0-1]. The large value are indication of better quality

boost-up in speed and efficiency. It shows that maximum time is taken by source image selection which is performed only single time during color morphing. It does not include feature extraction time. Figure 6.24 presents the GUI (Graphical User Interface) of our proposed fusion and colorization system, *SCENT*. One limitation of our system is that it is designed to deal with stationary camera environment in which background scene does not change. In future, we want to extend this approach to moving camera scenario where scene changes with the passage of time.

Table 6.1: Breakdown of CPU Time

Task	**Time**
False Color Fusion	0.06 ms
Color Channel De-correlation	0.09 ms
Source Image Selection	0.24 ms
Color Morphing	0.40 ms

6.7.10 An interesting Application: Contextual Action Recognition at Nighttime

To introduce usability of our approach, we present an interesting application and show that enhanced situational awareness through information fusion and colorization can greatly benefit visual surveillance applications. Recognizing human actions and activities in videos is an important research problem with potential applications in area of visual surveillance and evolving from simpler constraint action data sets to challenging scenarios at daytime. We intend to experiment the contextual awareness at nighttime with combination to action recognition. This application explores importance of semantically meaningful structure and dominant color context of environment in which action is taking place. The context becomes more important due to night. To the best of our knowledge, this application is

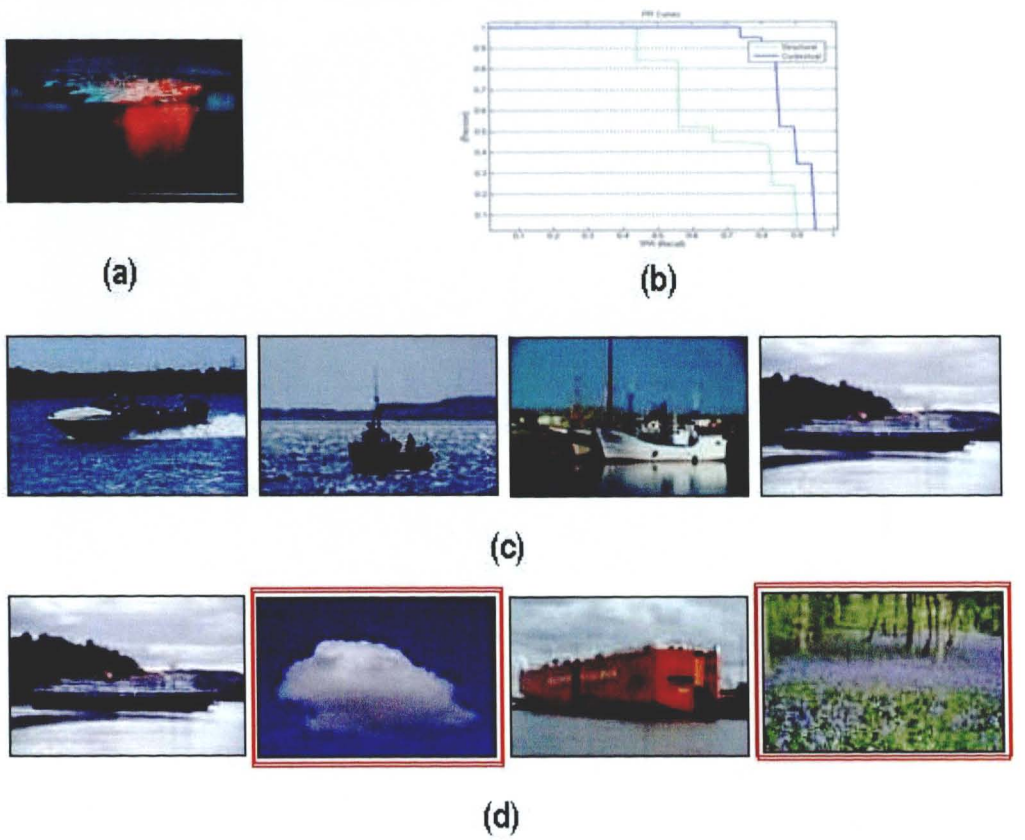


Figure 6.23: Color source Image Selection showing query image , Precision-recall curve and selected images from color image collection by (c) proposed contextual association (d) structural association [146]. False positives are shown in red bounding boxes

novel as no work exists in literature which deals action recognition at night time. Due to lack of suitable nighttime action dataset, we present results for only walking action.

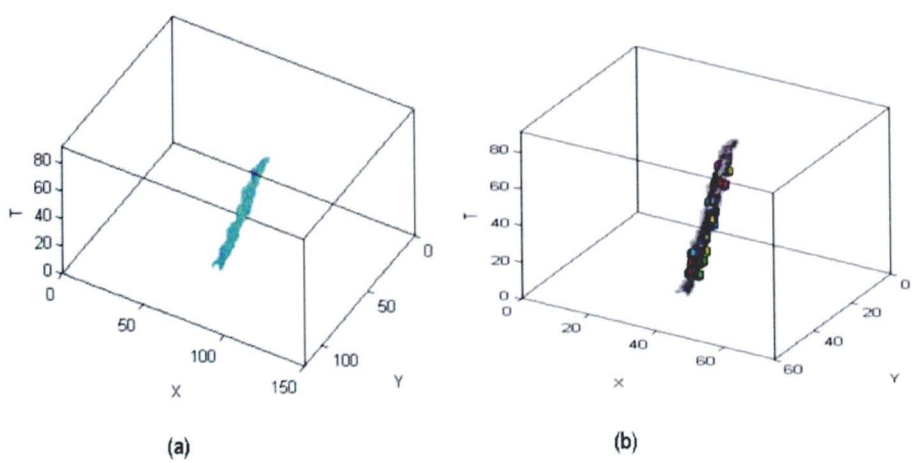


Figure 6.26: (a) 3D MSV for moving actor, (b) spatio-temporal cuboids encapsulated within 3D MSV

Figure 6.24: Snapshot of GUI (graphical user interface) developed for SCENT

Figure 6.25 illustrates such a scenario where a person is walking. Our objective is to find action and its contextual label. This experiment shows that colorization is an important clue to refine and find spatial context of presented scene. Two stage classification is involved in this application: (1) assigning action label, (2) assigning scene label. We present brief description of the implementation and detailed description of constituent techniques is out of the scope of this chapter. The first task is to segment spatio-temporal contents of an action from the scene using 3D segmentation. We use 3D MSVs(maximally stable volme)[16] extraction for 3D segmentation of action. The extracted 3D MSV of walking person in video sequence is shown in figure 6.26a. We extract spatio-temporal cuboids features [18] from given 3D MSV as shown in figure 6.26 b and action matching is performed on trained action dataset. For finding global context of an action, we utilize contextual modeling. A similar idea of contextual modeling is presented in [175]. we classify global scene content using dominant color descriptor (DCD) [174] and $GIST$ features. Dominant color is an important global features in images which describes the salient color distributions in an image. Dominant color descriptor (DCD) is one of the color descriptors proposed by MPEG-7 that has been extensively utilized for image retrieval. Dominant color descriptor is used to find clue of probable candidate classes in first stage. $GIST$ features are used for final refinement which labels the global context of scene. Finally, a combined semantic label is assigned to video sequence: $walking, building$. This labeling can be used for automated contextual action annotation.

6.7.11 CFOI: Experimentation

The performance of the proposed CFOI measure is evaluated on standard image fusion data sets.The datasets are collected from www.imagefusion.org provided by TNO Human Factor Research Institute, Netherlands, Octec Ltd. and Ohio State University, USA. Image Processing Toolbox in MATLAB R2008b is used for implementation. A set of experiments is designed to evaluate both types of color image fusion techniques described earlier in introduction using proposed CFOI with available blind (without reference) image fusion quality evaluation techniques . In all graphical representations, vertical columns in

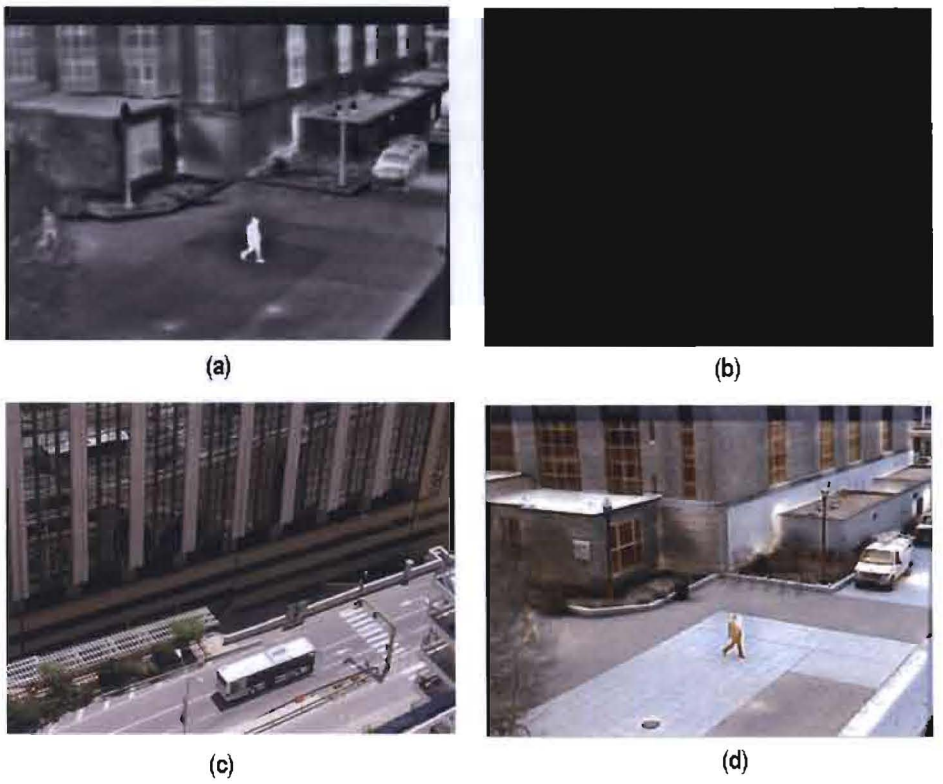


Figure 6.25: Scene Description for Contextual Action Recognition at Nighttime , (a) IR frame, (b) low light image, (c)color source (d) colorized frame

graph show color fusion quality values which is in range [0 1] while horizontal axis shows different results or different quality measures.

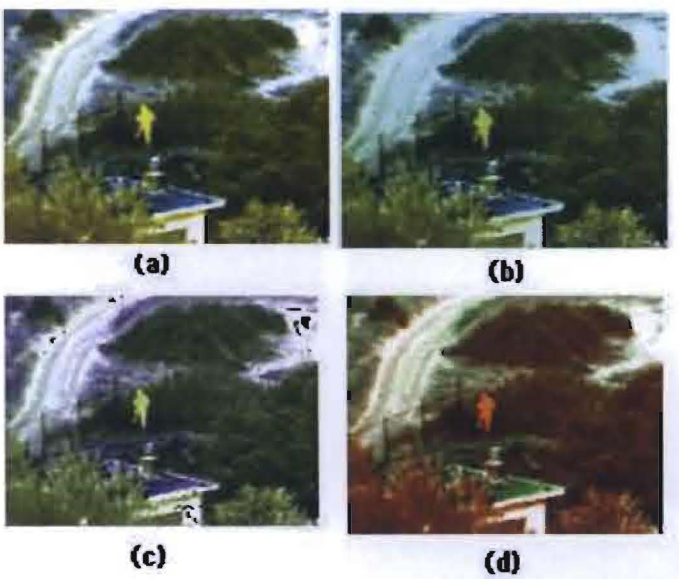


Figure 6.27: Color Image Fusion results with Color Distortions

First experiment deals with evaluation of colorization, the colors of final fused color image are modified by using different target color images (Figure 6.27a,6.27b), and also by changing their hue and saturation values (Figure 6.27c, 6.27d). First we subjectively

evaluated and assigned quality labels a,b,c,d accordingly. A good measure should capture these color variations. Figure 6.28. gives the graphical representation which compares CFOI and CSM to capture these color variations. Results show that CSM captures color variation but falsely gives more value to image (c) than image (a) which subjectively looks better with more colorfulness and natural colors unlike false distortions as in image (c). The results of CFOI are consistent with the subjective results.

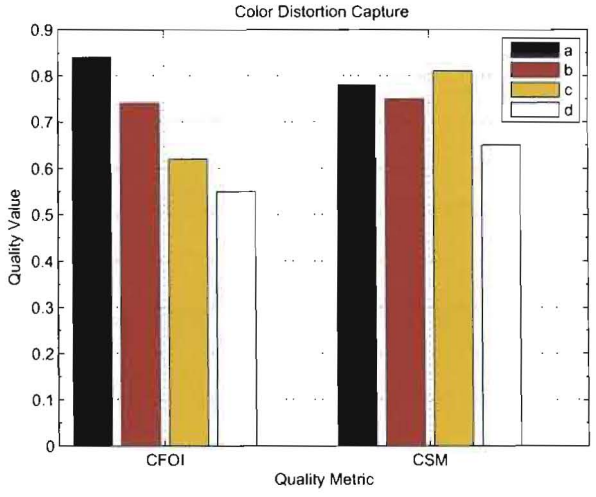


Figure 6.28: CFOI comparison for color distortions

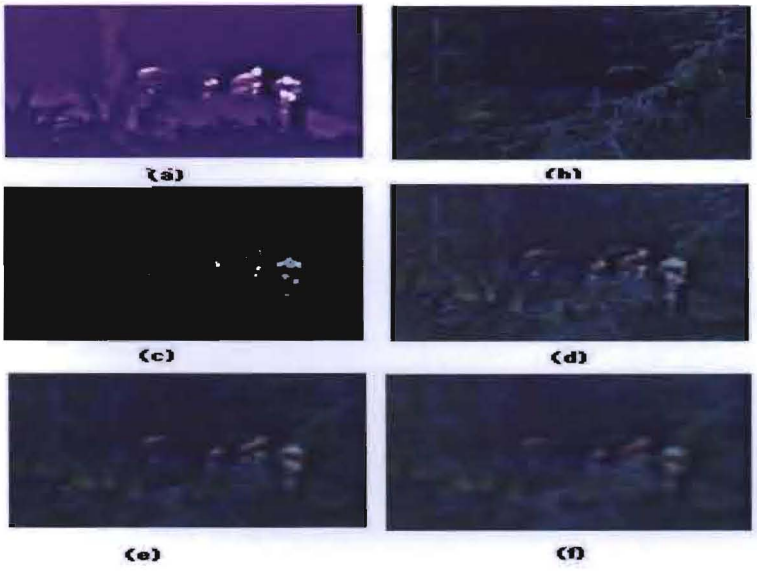


Figure 6.29: Original images (a,b), color fused image (c) and blurred images (d,e,f)

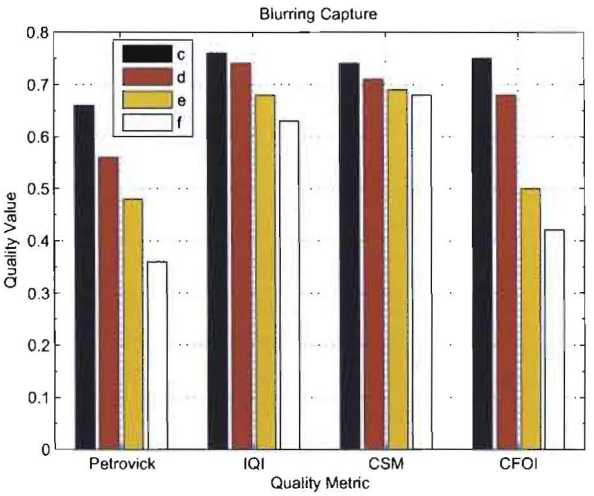


Figure 6.30: CFOI comparison for edge burring effect

Figure 6.31: Original images (a,b), color fused with averaging (c), Laplacian (d) and wavelet (e)

Second experiment is related to capturing of structure distortion and blurring. Different level of blurring by low pass filtering is introduced to final color fused image and ranked according to degree of blurring. CFOI is employed to capture the change along with other measures, CSM, Petrovick and IQI. Petrovick and IQI work for only grayscale images. Therefore, images are converted to grayscale before their calculation. Figure 6.29 describes color image fusion scenario for this experiment. First two images (Figure 6.29a,6.29b) are color infra-red and CCD images, respectively. Third image (c) is color fused image and fourth image (Figure 6.29d) is the blurred image generated for experimentation.

CFOI is applied, compared to Petrovic [164] and IQI measure [165]. Finally, graphical representation of results is presented in figure 6.30. CFOI and Petrovic measure (based

on sobel gradients), both use gradient or edge information for image fusion measurement. Therefore, blurring effect is reasonably captured by both as compared to IQI which uses SSIM which is less robust to blurring [167]. Results show the CFOI is comparable to Petrovick measure and performs better than IQI for capturing blurring effect.

Third and the last experiment, is about capturing of common fused information. CFOI is employed on different image fusion methods like simple averaging, Laplacian pyramid fusion [186] and wavelet based fusion methods [145] presented in figure 6.31. Graphical results in figure 6.32 show that Petrovic measure [164] is based on edge information only; therefore it is more biased towards contemporary information like fused high frequency edge information than common low frequency information. CFOI uses mutual information for getting common information in addition to edge information; it captures common information as well compared to Petrovick measure. Therefore, it gives more value to wavelet fusion than Laplacian fusion which was higher in case of Petrovick measure. At the same time, its results are comparable to IQI [165] for capturing common information.

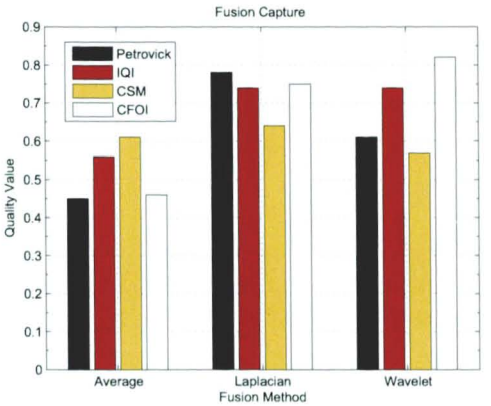


Figure 6.32: CFOI comparisons for image visual information integration

Experimental results show that CFOI captures fusion of all three factors, important for color image fusion (colors, high frequency edges and low frequency common information). In particular; it is not biased toward any of them as previous standards like Petrovic [164] and IQI [165]. In addition, it deals with colors which were not consideration of previous image fusion quality metrics.

6.8 Conclusions

In this chapter, we propose a method of recognizing actions with the help of their context. We take a case study of multi-sensor night vision consisting of infra-red and low light visible spectrum. We show how context is enhanced in such video sequences. We then use these context enhanced videos for contextual action recognition at nighttime. Our recognition scheme is based on information fusion and frequency domain matching. We show action recognition results for a large video collection. Performance comparison with the baseline shows that recognition accuracy is greatly increase for actions which show interactive relationship with their context. In addition, we propose an automated color video fusion approach for contextual enhancement at nigh-time and discuss its objective quality evaluation.

In next chapter, we conclude all the chapters by presenting an overall picture of our work. In addition, we would describe possible future work directions.

Chapter 7

Conclusion and Future Work

> Begin thus from the first act, and proceed; and in conclusion, at the ill which thou hast done, be troubled, and rejoice for the good
>
> ~Pythagoras (570 BC - 495 BC)

In this chapter, we restate the definition of our research, describe the importance of our approach and proposed solutions, take a recap of proposed approaches with a brief individual review and conclude our findings. In addition, we mention how each individual approach and solution can be investigated further and what overall future research work is possible.

7.1 A Recap of our Research Problem

The visual perception of human action is indeed a difficult phenomenon due to complex dynamics of human action, action context and scene capturing framework. One obstacle in the way of machine vision of human actions is lack of information about visual invariants. For instance, viewpoint variations cause huge problem to machine recognition of human actions because of insufficient information about viewpoint invariants. Although different visual cues and direction are helpful to devise techniques for recognizing human actions but majority of them are unable to cope with viewpoint variations which points towards a huge research gap. Can we explore and investigate salient visual cues deeply, search new viewpoint invariants and enhance the capability of their frameworks to cope with viewpoint variations?

7.2 The Significance and Impact of our Research Methodology

Due to three-dimensional nature of action video sequences, any general viewpoint invariant feature for action representation is undefined. Therefore, we exploit important visual cues that can be helpful in devising view invariant action recognition framework. The important visual cues explored for solving view-invariance include multiple view geometry, temporal order information and view clustering or information fusion. In our research work, we have investigated these visual cues and proposed view-invariant action recognition approaches,

contributing to view-restriction free action recognition. It is valuable because it can further contribute towards unrestricted action and activity recognition in computer vision.

Understanding human activity from video is one of the central problems in the field of computer vision. It is driven by a wide variety of applications in communications, entertainment, security, commerce and athletics. We have identified and focused on solution of a key problem of view-invariant action recognition to visualize practical application of action recognition in other related disciplines to broaden its impact.

7.3 A Re-cap of our proposed Approaches

We studied and explored our research problem in its deeper context and came up with the development of following new approaches:

- View-invariant action recognition framework using temporal order invariance (presented in Chapter 3)
- View-invariant action recognition using multiple view geometry (presented in Chapter 4)
- View-invariant action recognition using 3D frequency domain filtering (presented in Chapter 5)
- Context enhancement for contextual action recognition (presented in Chapter 6)

7.4 A Brief Review of our Proposed Approaches

A brief review and description of important findings is as follows:

- **View-invariant Action Recognition Framework using Temporal Order Invariance**:

  Description: This approach investigates the conjecture that temporal order of action elements (action sub-divisions) remains invariant for different viewpoints and it can help us to devise temporal order invariance constraint for view-invariant action recognition. Individual action instances as constituent action units (e.g., representation of local motion and posture variations) within an action preserve a temporal order irrespective of the camera viewpoints.

  To recognize and represent action subdivisions or local dynamics, we focus on global analysis of human actions and seek a view-invariant representation. We based our approach on the following conjecture: "*The temporal order of actions elements within an action is invariant to viewpoint variations*". We define action elements in terms of local spatio-temporal interest points and define spatio-temporal order preservation constraint in matching framework. Spatio-temporal cuboid features are taken as space-time interest points as these features are based on maximization of discrimination between behaviors.

  For each action class, we define a feature fusion table. A feature fusion table is a defined data structure to encapsulate multiple training examples against multiple viewpoints for a single action class. It is achieved through feature fusion based on principal component analysis. The fusion strategy is simple. An action video sequence contains many spatio-temporal features. (i) We arrange all video sequences of the same action class and the same view into the same group; (ii) We extract cuboid features from video sequences and sort features according to their temporal order; (iii) For all video sequences (same view, same class), we fuse features according

to their position in temporal order; (iv) Feature fusion is achieved through PCA. For instance, all features (position 1 in temporal order, 1st feature of all videos) of wave action in view 1 are concatenated into a single feature vector and principal component analysis is used to reduce its dimensionality to a single feature. (v) Finally, fused features for each class are arranged into fusion tables (to be described in the next section).

A matching score is then calculated based on global temporal order constraint and number of common features. Finally, the action label of the class with maximum value of matching score is assigned to the query action. The only bottleneck of this work is insufficient set of feature in most difficult viewpoint like a front-head camera which does not openly provide action dynamic clues. Dealing with action dynamic features in difficult viewpoints ia a possible future work direction.

Conclusion: The success of our approach validates the importance of temporal order of action instances in terms of their primitive dynamics. It concludes that if temporal order of action dynamics is ensured, better discriminative action classification can be achieved.

- **View-invariant Action Recognition using Multiple View geometry**:

  Description: This approach explores multiple view geometry and devises two incremental approaches based on exploitation of geometric constrains between action instances. In addition, these approaches address the weakness of trajectory based action recognition approaches and describe how tracking-free framework can be devised.

  We explore how dense optical flow can be employed to compensate strong assumptions of landmark point extraction and tracking in geometry based view invariant action recognition. Taking into consideration that human action is a spatio-temporal phenomenon, we apply constraints on optical flow to be spatio-temporally consistent. Spatio-temporally consistent optical flow helps us in devising spatio-temporally consistent flow fundamental matrix and by defining rank constraints on flow fundamental matrix we are able to derive a dissimilarity score for action sequences.

  We proceed incrementally by defining two variants of our approach: (1) We extract actor body silhouettes from original video sequences and calculate spatio-temporally consistent optical flow between respective frames of two videos and then fit epipolar geometry. As fundamental matrix remains same for static scenes, we can calculate action similarity score between two actions being performed in time domain, (2) In addition, we observed that silhouette extraction is not robust in all circumstances especially in case of noise and occlusion. Therefore, we remove pre-processing step of silhouette extraction theocratically by maximizing the exploitation of epipolar geometry.

  We take action representation in static camera environment as a case of dynamic scene where background is stationary and actor is dynamic. As scene is not entirely static, we get inspiration from structure and motion recovery for scenes consisting of both static and dynamic parts, also known as multi-body segmentation from perspective views without knowing which measurement belong to which part of the scene. As we consider only static background and dynamic actor, it is simplified to two-body fundamental matrix, also known as segmentation matrix. It has already been shown that such matrix can linearly be computed from image measurements after embedding all the image points in high dimensional space. Based on these investigations, we derive a new similarity measure for matching actions across different views, without prior segmentation of actors.

These proposed approaches has been named as AVITAR1 and AVITAR2 (Achieving View-invariant Tracking-free Action recognition). However, despite this success, some bottleneck issues have also been identified. These include dealing temporal un-synchronization and computational complexity associated with optical flow measurement. Both of these issue are very generic to all computer vision areas and they themselves are separate research problems away from our main research focus. However, future research work is possible in these directions.

Conclusion: The success of our approach validates the importance of multiple view geometry for achieving view invariant action recognition. It inherits the benefit that rank constraint based matching score can be calculated and used for action classification. It avoids various assumptions and simplifies the solution.

- **View-invariant Action Recognition using 3D Frequency Domain Filtering:**

  Description: This approach is based on frequency domain correlation filtering. In this regard, it proposed 3D distance classifier correlation filter named Action DCCF. This correlation filter is able to exploit intra as well as inter-class variations in 3D visual information of action sequences. This filter is further exploited for devising a view-invariant action recognition framework using view clustering mechanism. It successfully recognize actions despite viewpoint variations.

  To achieve this objective, we perform following steps: (i) We introduce space-time View-DCCF filter that can be trained for a specific viewpoint for all given action categories, it is done by establishing view clusters of action categories, (ii) View-DCCF filter successfully captures inter-class variability that is achieved by avoiding overemphasize on average training sample by empirically setting contributions of low and high frequency information, (iii) It presents a different interpretation of correlation filters as method of applying a *spatio-temporal transformation* to the data, restricted to being Toeplitz ensuring *shift invariance.* It measures similarity between an ideal transformed reference and testing action.

  In this way, it can handle linear action misalignments using a shift-invariant mean square distance measure, (iv) It utilizes entire correlation plane rather than emphasizing only single point like correlation peak as resulting decision boundaries are quadratic that are more 'selective' for choosing feature space portions for assigning to various action classes, and (v) finally, we extract an action similarity score based on class votes and within-cluster distance ratio. It helps us to recognize actions from an arbitrary viewpoint not present in training view clusters. Class votes help setting priority for class with maximum votes in all view clusters and within-cluster distance ratio highlights margin of selected class from other classes in a view cluster. All these contributions successfully fill up the research gap present in space-time filtering based action recognition.

  It also avoids bottlenecks faced by multiple view geometry based methods and spatio-temporal feature framework. It is faster in computational time and does not depend on feature extraction.

  Conclusion: The success of our approach validates the importance of frequency domain matching and its efficiency. We conclude that frequency domain signal analysis can guarantee suitable solution to action recognition even in presence of viewpoint variations.

- **Context Enhancement for Contextual Action Recognition:**

  Description: An additional but important aspect of action recognition is action context. Its importance increases in unfavorable circumstances like the challenging

case of night vision. We take this challenge and propose contextual action recognition at nighttime. To achieve this goal, we propose contextual enhancement of nighttime imagery.

We argue that contextual action recognition is not possible using single sensor platform due to the limitations of individual sensor to grab all available visual information about the scene. This situation motivates the use of multiple sensors often of complementary nature. A common multi-sensor night vision system uses infrared images in case of forward looking infrared cameras and low light images in case of low light visible cameras. The infrared images are maps of infra-red radiation emission which is partly governed by the temperature of the object. Therefore, such sensors prove good for perceiving hot targets in a busy background, seeing through fog, and monitoring paths through a cluttered forest.

However, they are not much effective during thermal crossover periods at night or after long periods of rain, as well as capturing scenery such as trees, leaves and grass in natural scene. On the other hand, low light visible cameras are able to capture surrounding environment but most of the time fail to capture specific targets especially hot bodies like a person in camouflage. In addition, even in case when targets are not hiding, low light conditions make their observation obscure. The objective of context enhancement is to give daylike appearance to nighttime videos.

We propose automated color night vision methods for context enhancement using video fusion. The video fusion is a process of visual information integration from a number of registered video sequences without loss of information and introduction of distortion. The goal of video fusion is to create a single enhanced video sequence from complementary video inputs that is more suitable for the purpose of human visual perception, action and context recognition. We also deal with the objective quality measurement of these methods which are not available in literature. We further propose contextual action recognition framework to show that how context can be a helpful visual cue for action recognition. The exploration about the importance of contextual view information to devise view-invariant action recognition framework is a possible research direction.

Conclusion: The success of our approach validates the importance of contextual information for action recognition. We conclude that contextual information is an important clue for achieving better recognition performance in action recognition especially in unfavorable visual conditions.

7.5 Future Research Directions

Many research problems still exit in computer vision and future directions of our work can be very helpful to solve these problems. A few of possible work directions are as under:

- **View-invariant Action Recognition in Restriction-free Real-time Video sequences**: The main goal of all action recognition research is to get rid of all restriction which we assume in our works. Our actions recording setting are restricted and majority of all action recognition datasets are recorded in specific requirement. On the other hand, human visual system easily recognizes any action in any circumstances.

- **View-invariant Action Recognition in Crowd Video Sequences**: A recent trend in action recognition research community is to deal situation of crowded scenarios rather than individual actions. Thus the overall objective is different that is the crowd action detection rather than individual action detection. However,

the visual cues from individual action can become helpful to devise overall crowd behavior.

- **View-invariant Complex Activity Recognition**: Another important research area is activity recognition which comprises combination of different actions and their overall interpretation. The focus of our work, however was consideration of individual actions, an important problem in this regard is the presence of scenarios which offer viewpoint variations.

7.6 Concluding Remarks

Human actions are fundamental to human existence and substance of great importance. Action analysis is a subject of primary importance, worth of scientific investigation and exploration. It is open for scientific enquiry with no defined boundaries. In other words, the analysis of human action is not restricted to some specific area of science, it is the subject of study in various scientific disciplines like neuroscience, cognitive science, agronomic, economics, psychology, praxeology and artificial intelligence.

Therefore, intelligent machines should be capable of interpreting visual scenes containing human actions. However, it is a very high level vision problem and a lot of research effort is still required to fulfill this dream. Over the years, several techniques have been developed, yet it is widely recognized that effective solutions are needed to be proposed and investigated. It is due to the nature of problem that combines the unpredictable human behavior, complex human motion dynamics, strong variations in camera environment especially viewpoint, occlusion and noise, presence of anthropometric differences and uncertainty associated with computational vision.

This thesis has mainly addressed an important problem of view-invariance, a necessary requirement for unconstrained action recognition. Multiple applications like action retrieval from video sequences and contextual action recognition at nighttime can be used for intelligent video surveillance and multimedia search. View-invariant action recognition can be used for developing interesting video games and human computer interface. Therefore, this research work has wider impact not only on computer vision research but also for other related disciplines.

Chapter 8

Appendix

Appendix A: Multi-frame (four-frame) Feature Matching

Let $I_1 = \Omega_1 \longrightarrow \mathbb{R}$, $I_2 = \Omega_2 \longrightarrow \mathbb{R}$, $I_3 = \Omega_3 \longrightarrow \mathbb{R}$ and $I_4 = \Omega_4 \longrightarrow \mathbb{R}$ be four frames of a multi-view video sequence with some common field of view on a dynamic scene. These images can be obtained from one or more unsynchronized cameras.

For each image $I_i, i \in \{1, 2, 3, 4\}$, a feature detector determines features $f_{i,k}, k \in \{1, \cdots, N_i\}$ with corresponding descriptors $s_{i,k}$ and descriptor distance function with $d(s_{i,k}, s_{j,m})$. We look for quadruple $(f_{1,k}, f_{2,m}, f_{3,n}, f_{4,o})$ such that each $(f_{i,j})$ is present in at most one quadruple.

For every quadruple, a cost $\tilde{d}$ is assigned that is the sum of the distances of all descriptors $\tilde{d}(s_{1,k}, s_{2,m}, s_{3,n}, s_{4,o}) = d(s_{1,k}, s_{2,m}) + d(s_{1,k}, s_{3,n}) + d(s_{1,k}, s_{4,o}) + d(s_{2,m}, s_{3,n}) + d(s_{2,m}, s_{4,n}) + d(s_{3,n}, s_{4,o})$. In this way, the distances between each pair of features is considered in the cost function making it independent of the ordering of the images. The four image matching algorithm can be written as :

1.(a) Match the features in I_1 and I_2, using nearest neighborhood matching [100], optionally with distance check to the second nearest neighbor. (b) Match the features in I_2 and I_1, using nearest neighborhood matching, optionally with distance check to the second nearest neighbor. (c) Accept only symmetrically matched features.

2. Remove unmatched features in I_1 and merge the remaining features on the basis of the matching in step (1) such that the new cost function between matched features in I_1 and features in I_3 is $\hat{d}(s_{1,k}, s_{3,n}) = \tilde{d}(s_{1,k}, s_{2,m}, s_{3,n}, s_{4,o})$.

3. (a)Match the features in I_1 and I_3, using nearest neighborhood matching, optionally with distance check to the second nearest neighbor. (b) Match the features in I_3 and I_1, using nearest neighborhood matching, optionally with distance check to the second nearest neighbor. (c) Accept only symmetrically matched features.

4. Remove unmatched features in I_1 and merge the remaining features on the basis of the matching in step (3) such that the new cost function between matched features in I_1 and features in I_4 is $\hat{d}(s_{1,k}, s_{4,0}) = \tilde{d}(s_{1,k}, s_{2,m}, s_{3,n}, s_{4,o})$.

5. (a) Match the features in I_1 and I_4, using nearest neighborhood matching, optionally with distance check to the second nearest neighbor. (b) Match the features in I_4 and I_1, using nearest neighborhood matching, optionally with distance check to the second nearest neighbor. (c) Accept only symmetrically matched features.

6. Interchange the role of I_1, I_2, I_3,I_4 and restart at step (1).

7. Merge the four matchings and return only those matches that are assigned in all four matching directions.

Appendix B: Tables of Abbreviations

Table 8.1: List of Abbreviations for Chapter 3

Abbreviation	Denoting
C	Set of spatio-temporal cuboid features
P	Set of STOP features
ρ	Stability criterion of MSV
L	Geometric location of features
I	Geometric inconsistency of features
S	Matching score between two videos
σ	Overall Matching score from all videos
K	Number of action classes
F	Fused features from one class
T	Feature fusion table
γ	Weighting parameter

Table 8.2: List of Abbreviations for Chapter 4

Abbreviation	Denoting
A	3D point on human body
T	Translation vector connecting two cameras
R	Rotation vector between two cameras
E	Essential Matrix between two views
F	Fundamental matrix between two views
M	Intrinsic camera parameters
O	Observation or measurement matrix
υ	veronese mapping
$\otimes$	The Knonecker product
Ω	Matrix to vector conversion
$TV-L2$	Total variation, L2 Norm
$\bigtriangledown$	smoothness term
θ	Auxiliary variable for flow
d	Flow update
ψ	Flow update weighting parameter
E	Quadratic Energy function

Table 8.3: List of Abbreviations for Chapter 5

Abbreviation	Denoting
a	Action instance
r	Reference action
H	linear transformation
A	3D FFT of original action volume
SS	Spectral seperation
SM	Similarity Measure
d	3D distance between classes
W	Within cluster distance ratio
V	Class vote
b	Energy of the transformed class mean
p	Energy of the transformed input

Table 8.4: List of Abbreviations for Chapter 6

Abbreviation	Denoting
C	Covariance matrix
U	Data matrix
Λ	diagonal matrix of eigen values of Covariance matrix
λ	Eigen values
f	GISt feature vector
CSM	Structural Similarity Measure
CS	Color Similarity
IS	Intensity Similarity
SQ	Structural Quality
MI	Mutual Information
Cf	Colorfulness of an image

References

[1] L. Von Mises. Human action: A treatise on economics. In Chicago: Henry Regnery, 1966.

[2] M. Pakaluk, Aristotles Nicomachean Ethics: An Introduction. Chicago: University of Chicago Press, 2005.

[3] A. R. Mele, The Philosophy of Action, Oxford University Press, Oxford, 1997.

[4] B.F. Skinner, The operational analysis of psychological terms, Behavioral and brain sciences, 7 (4): 54781, 1984.

[5] T. Parsons, The Present Status of "Structural-Functional" Theory in Sociology, In Talcott Parsons, Social Systems and The Evolution of Action Theory, New York: The Free Press, 1975.

[6] R. K. Merton, The Unanticipated Consequences of Purposive Social Action. American Sociological Review 1 (6): 894904, 1936.

[7] A. Goldman. A theory of human action. In Englewood Cliffs, Prentice Hall, 1970.

[8] G. Rizzolatti, L. Craighero, The mirror-neuron system, Annual Review of Neuroscience, 27:169-192,2004.

[9] M. Shah, Guest Introduction: The Changing Shape of Computer Vision in the Twenty-First Century, IJCV, 2002.

[10] T. B. Moeslund, A. Hilton, V. Krger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding, 104 (23), 90126, 2006.

[11] A. P. Brandao, E. A. do Valle Jr., J. M. Almeida1, A. A. de Araujo, Action Recognition in Videos:from Motion Capture Labs to the Web, Computer Vision and Image Understanding, 2010.

[12] R. Poppe, A survey on vision-based human action recognition, Computer Vision and Image Understanding, 28, 976990, 2010.

[13] A. F. Bobick and J. Davis, The Recognition of Human Movement using Temporal Templates, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 23, No. 3, 2001.

[14] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR05), vol. 1, pp. 886893, San Diego, CA, June 2005.

[15] D. Weinland, R. Ronfard, and E. Boyer. Free Viewpoint Action Recognition Using Motion History Volumes, *CVIU* 2006.

[16] M. Donoser and H.Bischof, 3d Segmentation by Maximally Stable Volumes, In Proc. ICPR 2006.

[17] I. Laptev. On Space-Time Interest Points, *IJCV* 2005.

[18] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, Behavior Recognition Via Sparse Spatio-temporal Features, *In Proc.* VS-PETS, 2005.

[19] D. Weinland, R. Ronfard, E. Boyer, A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition, Computer Vision and Image Understanding, 2010.

[20] R. O. Duda, P. E. Hart, D. G. Stork, Pattern classification (2nd edition), Wiley, New York, 2001.

[21] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in timesequential images using hidden Markov model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR92), pp. 379385, Champaign, IL, June 1992.

[22] L. Wang, D. Suter, Visual learning and recognition of sequential data manifolds with applications to human movement analysis, Computer Vision and Image Understanding (CVIU), 110 (2), 153172, 2008.

[23] A. F. Bobick, S. S. Intille, J. W. Davis, F. Baird, C. S. Pinhanez, L. W. Campbell, Y. A. Ivanov, A. Schutte, A. Wilson, The kidsroom: A perceptually-based interactive and immersive story environment, Presence: Teleoper. Virtual Environ. 8 (4), 369393, 1999.

[24] F. Tsalakanidou, S. Malassiotis, Robust facial action recognition from real-time 3d streams, Proceedings of IEEE CVPRW 09, 0 , 411, 2009.

[25] A. Branzan Albu, T. Beugeling, N. Virji Babul, C. Beach, Analysis of irregularities in human actions with volumetric motion history images, in: Motion 07, 2007.

[26] L. Gorelick, M. Blank , E. Shechtman, M. Irani and R. Basri. Actions as Space-Time Shapes, *TPAMI* 2007.

[27] C. Schüldt, I. Laptev, and B. Caputo. Recognizing Human Actions: A Local SVM Approach, In Proc. International Conference on Pattern Recognition, 2004.

[28] D. Weinland, E. Boyer, and R. Ronfard, Action recognition from arbitrary views using 3D exemplars, IEEE ICCV, Rio de Janeiro, pp. 17, Oct. 2007.

[29] M. Rodriguez, J. Ahmad, and M. Shah, Action MACH: A Spatio-temporal Maximum Average Correlatio Height filter for Action recognition, In Proc. Computer Vison and Pattern Recognition, 2008.

[30] J. Liu, J. Luo and M. Shah, Recognizing Realistic Actions from Videos "in the Wild", In Proc. Computer Vison and Pattern Recognition, 2009.

[31] http://www.csee.wvu.edu/ vkkulathumani/wvu-action.html

[32] R. Hartley, A. Zisserman., A., Multiple View Geometry in Computer Vision. Cambridge University Press, 2004.

[33] M. Han and T. Kanade. Multiple motion scene reconstruction from un-calibrated views, *In Proc.* ICCV 2001.

[34] E. Trucco and A. Verri, *Introductory techniques for 3d computer vision.* Prentice Hall,1998.

[35] H. Longuet-Higgins, A computer algorithm for reconstructing a scene from two projections. *Nature* 1981.

[36] T. S-Mahmood, A. Vasilescu, S. Sethi, I. Center, and C. San Jose. Recognizing Action Events from Multiple Viewpoints. *in Proc.* WDREV 2001.

[37] V. Parameswaran and R. Chellappa. View Invariance for Human Action Recognition. *IJCV* 2006.

[38] C. Rao, A. Yilmaz and M. Shah. View Invariant Representation and Recognition of Actions. *IJCV* 2002.

[39] A. Yilmaz and M. Shah. Actions as objects: A novel action representation. IEEE Proc. *CVPR*, 2005.

[40] Shen, Y. and H. Foroosh. View-Invariant Action Recognition from Point Triplets. *TPAMI* 2009.

[41] Yuping Shen and Hassan Foroosh. View Invariant Action Recognition Using Fundamental Ratios. *In Proc.* CVPR 2008.

[42] M. Shah A. Gritai, Y. Sheikh. On the use of anthropometry in the invariant analysis of human actions. In International Conference on Pattern Recognition, 2004.

[43] A. Gritai, Y. Sheikh, C. Rao, and M. Shah. Matching trajectories of anatomical landmarks under view-point, anthropometric and temporal transforms. *IJCV* 2009.

[44] A. Yilmaz and M. Shah. Matching actions in presence of camera motion. Computer Vision and Image Understanding, 104(2-3):221231, 2006.

[45] M. Mainberger, A. Bruhn and J. Weickert. Is dense optical flow useful to compute the fundamental matrix? *LNCS* 2008.

[46] L. Wolf and A. Shashua. Two-body Segmentation from Two Perspective Views. *In Proc* CVPR 2001.

[47] R. Vidal,Y. Ma,S. Soatto and S. Sastry. Two-view Multibody Structure from Motion. *IJCV* 2002.

[48] R. Polana and R. Nelson, Low level recognition of human motion, In Proc. IEEE Workshop on Motion of Non-rigid and Artculated Objects, pp. 77-82, 1994.

[49] E. Shechtman and M. Irani; Space-time behaviour based Coorelations, In Proc. Computer Vison and Pattern Recognition, 2005.

[50] S. Ali , and S, Lucey Are correlation filters useful for human action recognition, In Proc. International Conference on Pattern Recognition, 2010.

[51] D. Weinland, O. Mustafa , and P. Fua, Making Action Recognition Robust to Occlusions and Viewpoint Changes,*In Proc.* ECCV 2010.

[52] J. Liu, M. Shah, B. Kuipers and S. Savarese. Cross-view Action Recognition via View Knowledge Transfer. *In Proc.* CVPR 2011.

[53] M. Marszalek, I. Laptev, and C. Schmid, Actions in Context, In Proc. Computer Vison and Pattern Recognition, 2009.

[54] L. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. *In Proc.* ICCV, 2007.

[55] D. Han, L. Bo, and C. Sminchisescu, Selection and context for action recognition, *In proc.* CVPR 2009.

[56] Y. Jiang, Z. Li, and S. Chang, Modeling Scene and Object Contexts for Human Action Retrieval with Few Examples, *IEEE Transactions on circuits and systems for video technology,* vol. 21, NO. 5, 2011.

[57] Jian F. Li, Wei G. Gong, Application of Thermal Infrared Imagery in Human Action Recognition, *Advanced Materials Research*, 121-122, 368, 2010.

[58] J. Han; B. Bhanu, Human Activity Recognition in Thermal Infrared Imagery, *In Proc.* CVPR Workshops 2005.

[59] L. A. Klein, *Sensor and Data Fusion: A Tool for Information Assessment and Decision making*, SPIE publishers, 2004.

[60] A. Levin, D. Lischinski and Y. Weiss, "Colorization using optimization", *ACM Trans. on Graph.*, vol.23, no.3, 2004.

[61] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley, "Color Transfer between Images", *IEEE Comp. Graph. and Appl.*, vol. 21, pp. 34-41, 2001.

[62] T. Welsh, M. Ashikhmin, and K. Mueller, "Transferring color to greyscale images", *ACM Trans. on Graph.*, vol. 21, pp. 277-280, 2002.

[63] V. Zatsiorsky, Kinematics of Human Motion, In AHuman Kinetics, 2002.

[64] B. farnell, Moving bodies, Acting selves, In Annual review of Anthropometry, 1999.

[65] K. Verfaillie. Variant points of view on viewpoint invariance. In Canadian Journal of Psychology, 1992.

[66] L. Fogassi, V. Gallese, L. Fadiga and G. Rizzolatti. Action recognition in the premotor cortex. In Brain, 1996.

[67] G. Johansson. Visual perception of biological motion and a model for its analysis. Perception and Psychophysics, 14:201211, 1973.

[68] J. Decety and J. Grezes. Neural mechanisms subserving the perception of human actions. In Trends in Cognitive Sciences, 1999.

[69] J. Aggarwal and Q. Cai. Human motion analysis: A review. In Computer Vision and Image Understanding, 1999.

[70] J. Aggarwal and S. Park. Human motion: Modeling and recognition of actions and interactions. In Second International Symposium on 3D Data Processing, Visualization and Transmission, 2004.

[71] I. N. Junejo, E. Dexter, I. Laptev and P. Prez. View-Independent Action Recognition from Temporal Self-Similarities. *TPAMI 2010.*

[72] A. Bartoli. The geometry of dynamic scenes-On coplanar and convergent linear motions embedded in 3D static scenes. *CVIU* 2004.

[73] M. Herman. Understanding body postures of human stick figures. In PhD Thesis, University of Maryland, 1979.

[74] D.C. Hogg. Interpreting Images of a Known Moving Object. PhD thesis, University of Sussex, 1984.

[75] C. Cedras and M. Shah. Motion-based recognition: A survey. In Image and Vision Computing, 1995.

[76] J. Davis and M. Shah. Three-dimensional gesture recognition. In Proc. of Asilomar Conference on Signals, Systems and computers, 1994.

[77] B. Farnell. Moving bodies, acting selves. In Annual Review of Anthropology, 1999.

[78] D. M. Gavrila. The visual analysis of human movement: A survey. CVIU, 73(1):8298, 1999.

[79] W. Liao, J. Aggarwal, Q. Cai and B. Sabata. Articulated and elastic non-rigid motion: A review. In Workshop on Motion of Non-Rigid and Articulated Objects, 1994.

[80] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pages 3844, 1996.

[81] W. Hu L. Wang and T. Tan. Recent development in human motion analysis. In Pattern Recognition, 2003.

[82] S. Avidan and A. Shashua. Trajectory Triangulation of Lines: Reconstruction of a 3D point Moving along a Line from a Monocular Image Sequence. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), June, 1999.

[83] S. Avidan and A. Shashua. Trajectory Triangulation: 3D Reconstruction of Moving Points from a Monocular Image Sequence. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 22(4), pp. 348–357, 2000.

[84] P. Lenz, J. Ziegler, A. Geiger and M.Roser, Sparse Scene Flow Segmentation for Moving Object Detection in Urban Environments, IEEE Intelligent Vehicles Symposium (IV), 2011

[85] Y. Sheikh, A. Gritai, M. Shah, On the Spacetime Geometry of Galilean Cameras, CVPR 2007

[86] I. Essa and A. Pentland, A Vision System for Observing and Extracting Facial Action Parameters, *In CVPR* 1994.

[87] J. Liu, M. Shah. Learning human actions via information maximization. *In Proc.* CVPR 2008.

[88] A. Sellent, C. Linz, M. Magnor. Consistent Optical Flow for Stereo Video. *In Proc.* ICIP 2010.

[89] A. Sellent, M. Eisemann, M. Magnor. Robust Feature Matching in General Multi-Image Setups. *In Proc.* WSCG, Plzen, Czech Republic,2011.

[90] M. Black and A. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. pages 6384, *in Proc.* ECCV 1998.

[91] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In ICCV, 2007.

[92] A. Oikonomopoulos, I. Patras, and M. Pantic. Spatio-temporal salient points for visual recognition of human actions. IEEE Trans. Systems, Man, and Cybernetics, Part B, 36 (3):710719, 2006.

[93] S. F. Wong and R. Cipolla. Extracting spatio-temporal interest points using global information. In ICCV, 2007.

[94] S. Venkatesh, N. Nguyen, D. Phung and H. H. Bui. Learning and detecting activities from movement trajectories using the hierarchical hidden markov models. IEEE Proc. *CVPR*, San Diego, CA, 2005.

[95] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. on ASSPR, Vol. 26, No.1, 1978.

[96] P. Scovanner, S. Ali and Mubarak Shah, A 3-Dimensional SIFT Descriptor and its Application to Action Recognition, ACM Multimedia, 2007.

[97] P.Yan, S aad M. Khan and Mubarak Shah, Learning 4D Action Feature Models for Arbitrary View Action Recognition, Proc. CVPR , Alaska, 2008.

[98] Y. Ukrainitz, M. Irani: Aligning Sequences and Actions by Maximizing Space-Time Correlations. Proc. *ECCV*, 2006.

[99] F. Lv and R. Nevatia. Single View Human Action Recognition using Key Pose Matching and Viterbi Path Searching. Proc. *CVPR*, pages 18, 2007.

[100] D. G. Lowe, Object recognition from local scale-invariant features. In: Proc. *ICCV* , Kerkyra, Greece, pp. 11501157, 1999.

[101] G. Zhu, C. Xu, W. Gao, and Q. Huang. Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine. Lecture Notes in Computer Science, 3979:89, 2006.

[102] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. Proc. IEEE *ICCV*, pages 726733, 2003.

[103] L. Wang. Abnormal Walking Gait Analysis Using Silhouette-Masked Flow Histograms. Proc. *ICPR*, pages 473476, 2006.

[104] C. Liu, J. Yuen and A. Torralba. SIFT flow: dense correspondence across different scenes and its applications.IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), Vol 33, No.5, 2011.

[105] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In Proc. ECCV, pages 629644, 2002.

[106] Anwaar-ul-Haq, I. Gondal, and M. Murshed, On Dynamic scene Geometry for View-invariant Action Matching , *In Proc.* CVPR, 2011.

[107] J. Little et al., Recognizing People by Their Gait: The Shape of Motion, Journal of Computer Vision Research, 1998.

[108] J. Choi, W. Jeon, and S.-C. Lee, Spatio-temporal Pyramid Matching for Sports Videos, ACM Multimedia, 2008.

[109] J. Sun, X. Wu, S. Yan, L.-F. Cheong, T.-S. Chua, and J. Li. Hierarchical Spatio-temporal ContextModeling for Action Recognition, In Proc. Computer Vison and Pattern Recognition, 2009.

[110] A. Kovashka and K. Grauman, Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition, In Proc. Computer Vison and Pattern Recognition, 2010.

[111] S. Ali and M. Shah, Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010.

[112] S. Belongie, J. Malik, and J. Puzicha, Shape Matching and Object Recognition Using Shape Context, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(4), 2002.

[113] A. Yilmaz and M. Shah, Actions Sketch: A Novel Action Representation, In Proc. Computer Vison and Pattern Recognition, 2005.

[114] J. Sivic, A. Zisserman, Video Google: A Text Retrieval Approach to Object Matching in Videos, In Proc. International Conference on Computer Vision, 2003.

[115] J. C. Niebles, H.Wang, and L. Fei-Fei, Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words, In Proc. British Machine Vision Conference, 2006.

[116] B. V. Kumar et al., *Correlation Pattern recognition*, Cambridge University Press, 2005.

[117] A. V. Oppenheim, *Signal and Systems*, Prentice Hall, 1996.

[118] Q. Hu et al., Action Recognition Using Spatial-Temporal Context, In Proc. International Conference on Pattern Recognition, 2010.

[119] L. Cao; Zicheng Liu; Huang, T.S.; Cross-dataset action detection, In Proc. Computer Vison and Pattern Recognition, 2010.

[120] http://guthspot.se/video/deshaker.htm

[121] M. Rodriguez, CRAM: Compact representation of actions in movies, *In Proc.* CVPR, 2010.

[122] Y. O. Alatas and M. Shah, Spatio-temporal Regularity Flow (STRF): Its estimation and applications, *IEEE Trans. on circuits and systems for video technology,* 2007.

[123] I. Laptev, M.Marszalek, C.Schmid, and B. Rozenfeld, Learning Realistic Human Actions from Movies. In Proc. CVPR 2008.

[124] Z. Wu, Q. Ke, M. Isard, and J. Sun, Bundling Features for Large Scale Partial-duplicate Web Image Search, CVPR 2009.

[125] Z. Wu, Q. Xu, S. Jiang, Q.Huang, P.Qui and L. Li, Adding Affine Geometric Constraint for Partial-duplicate Image retreival, ICPR 2010.

[126] A. Fathi and G. Mori, Action Recognition by Learning Mid-Level Motion Features, In Proc. CVPR, 2008.

[127] A. Gilbert, J. Illingworth, and R. Bowden, Fast Realistic Multi-Action Recognition Using Mined Dense Spatio-Temporal Features, In Proc. ICCV, 2009.

[128] J. Liu, S. Ali, and M. Shah, Recognizing Human Actions Using Multiple Features, In Proc. CVPR, 2008.

[129] Y. Yacoob and M. Black, Parameterized Modeling and Recognition of Activities, Computer Vision and Image Understanding, 1999.

[130] J. Matas, O. Chum, M. Urba, and T. Pajdla, Robust wide baseline stereo from maximally stable external regions. In Proc BMCV, 2002.

[131] K. Schindler, L. Van Gool, Action Snippets: How Many Frames Does Human Action Recognition Require? In Proc CVPR 2008.

[132] H. Riemenschneider, Donoser, M. and Bischof, H., Bag of Optical Flow Volumes for Image Sequence Recognition" , BMVC 2009.

[133] G. Willems et al. An efficient dense and scale-invariant spatio-temporal interest point detector, ECCV, 2008.

[134] X. He, P. Niyogi . Locality Preserving Projections, In Advances in Neural Information processing Systems, Cambridge, M.A. MIT Press, 2000.

[135] A. Klser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3Dgradients. In BMVC, 2008.

[136] I. Laptev and T. Lindeberg. Local descriptors for spatio-temporal recognition. In First International Workshop on Spatial Coherence for Visual Motion Analysis, LNCS. Springer, 2004.

[137] C. Harris and M.J. Stephens. A combined corner and edge detector. In Alvey Vision Conference, 1988.

[138] S. Nowozin, G. Bakir, K. Tsuda, Discriminative Subsequence Mining for Action Classification, *In Proc.* ICCV, 2007.

[139] Y. Cao, C. Wang, Z. Li, L. Zhang, Spatial-bag-of-features, *In Proc.* Computer Vision and Pattern Recognition, 2010.

[140] J. C. Niebles, C.-W. Chen and L. Fei-Fei, Modeling temporal structure of decomposable motion segments for activity classification, *In Proc.* ECCV, 2010.

[141] A. Gaidon, Z. harchaoui, C. Schmid, Actom Sequence Models for Efficient Action Detection, *In Proc.* Computer Vision and Pattern Recognition, 2011.

[142] A. Gupta and L. Davis, Objects in action: An approach for combining action understanding and object perception, *in Proc.* Conf. Comput. Vision Patt. Recog., 2007.

[143] S. Tran and L. S. Davis, Visual event modeling and recognition using Markov logic networks, *in Proc.* ECCV, 2008.

[144] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg, A scalable approach to action recognition based on object use, *in Proc.* Int. Conf. Comput. Vision, 2007, pp. 361368. *CVIU* 2006.

[145] Anwaar-ul-Haq, I.Gondal and M. Murshed, A Novel Image Fusion Algorithm based on Kernel-PCA, DWT and Structural Similarity, *In Proc.* VIIP Benidorm, Spain, 2005.

[146] Anwaar-ul-Haq, I.Gondal and M. Murshed, Automated multi-sensor color video fusion for nighttime video surveillance, *in: Proc.* of IEEE international symposium on computers and communications (ISCC), Riccione, Italy 2010.

[147] P. Perez, M. Gangnet, and A. Blake. Poisson image editing. *ACM SIGGRAPH*, 22(3):313318, 2003.

[148] I. Pavlidis, V. Morellas, P. Tsiamyrtzis, and S. Harp, Urban surveillance systems: from the laboratory to the commercial world, *In Proc. of the IEEE*, vol. 89, no. 10, pp. 1478 -1497, 2002.

[149] D. Lowe, Distinctive image features from scale-invariant keypoints, *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91110, 2004.

[150] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman. SIFT flow: dense correspondence across different scenes. *In proc.* European Conference on Computer Vision (ECCV), 2008.

[151] C. Conaire, N. E. O'Connor, A. Smeaton. Thermo-Visual Feature Fusion for Object Tracking Using Multiple Spatiogram Trackers. *Journal of Machine Vision and Applications*, 2007.

[152] J. Davis and V. Sharma, Background-Subtraction using Contour-based Fusion of Thermal and Visible Imagery, *Computer Vision and Image Understanding*, Vol 106, No. 2-3, 2007.

[153] J. J. Lewis, S. G. Nikolov, A. Loza, E. Fernandez Canga, N. Cvejic, J. Li, A. Cardinali, C. N. Canagarajah, D. R. Bull, T. Riley, D. Hickman, M. I. Smith, The Eden Project multi-sensor data set, Technical report TR-UoB-WS-Eden-Project-Data-Set, University of Bristol and Waterfall Solutions Ltd, UK, 2006.

[154] R.C.Gonzalez and R.E. Woods, *Digital Image Processing*, Prentice Hall,2nd ed. 2002.

[155] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, *Information Fusion*, vol.4, no.4, 2003.

[156] H. Li, B. S. Manjunath and S. K. Mitra, Multi-sensor image fusion using the wavelet transform, *Graph. Mod. and Img. Proc.,*. vol.57, no.3, pp.235-245, 1995.

[157] Anwaar-ul-Haq, I. Gondal and M. Murshed, Scarf: Semi-automatic Colorization and Reliable Image Fusion, in: *Proc. of IEEE DICTA*, Sydney, Australia, 2010.

[158] A. Waxman, M. Aguilar, D. Fay, A. N. Gove, M. Seibert, J. P. Racamato, J. E. Carrick and E. D. Savoye , Color Night vision:Fusion of intensified visible and thermal IR imagery, in: *Proc. of SPIE*, vol. 2463, pp.58-68, 1995.

[159] D.A.Fay, A.M. Waxman, M.Aguilar,D.B.Ireland,J.P.Racamato,W.W.Streilien, and M.I.Braun, Fusion of multi-sensor imagery for night vision: color visualization,target learning and search, in: *Proc. of 3rd Int. Conf. on Inform. Fus.*, Paris, 2000.

[160] G. Li and K. Wang, Applying daytime colors to nighttime imagery with an efficient color transfer method, in: *Proc. of Enh. and Synth. Vis.*, pp. 65590L-12, Orlando, FL, USA, 2007.

[161] A. Oliva and A. Torralba, Building the gist of a scene, the role of global image features in recognition, *Prog. in Br. Res.*, Vol.155,2006.

[162] A.Oliva and A.Torralba, The role of context in object recognition, *Tren. in Cogn. Sc.*, vol.11, no.12, 2007.

[163] D. L. Ruderman, T. W. Cronin, and C.C. Chiao, Statistics of cone responses to natural images: implications for visual coding, *J. Opt. Soc. of America*, vol.15, no.8, pp. 2036-2045, 1998.

[164] V. Petrovic , C. Xydas, Objective evaluation of signal level image fusion performance. *Opt. Eng.*, vol.44. 2005.

[165] G. Piella, and H. Heijmans. A new quality metric for image fusion, *in:Proc. of Int. Conf. of Img Proc.*, 2003.

[166] N. Cvejic, A. Loza, D. Bull, and N.Canagarajah, A similarity metric for assessment of image fusion, *Int. J. of Sig. Proc.,* vol.2, pp.178-182, 2005.

[167] X. Zhang, A novel quality metric for image fusion based on color and structural similarity, in: P*roc. of Int. Conf. on Sig. Proc. Sys.*, Singapore, 2009.

[168] Anwaar-ul-Haq, I. Gondal and M. Murshed, A novel color image fusion QoS measure for multisensor night vision application, *in: Proc. of IEEE Int. symp. on Comput. and commu.*, Italy, 2010.

[169] D. Hasler, and S. E. Suesstrunk, Measuring colorfulness in natural images, in: *Proc. of Hum. Vis. and Elec. Imag. VIII.* Santa Clara, CA, USA, 2003.

[170] G. H. Chen., Y. Chun-Ling, and X. Sheng-Li, Gradient-based structural similarity for image quality assessment, *in: Proc. of IEEE Int. Conf. on Img. Proc.*,2006.

[171] D. B. Russakoff , C.Tomasi, T. Rohlfing and C. R. Maurer Jr., Image similarity using mutual information of regions, *Lec. Notes in Comp. Sc.*, vol.3023, pp. 596-607, Springer Berlin-Heidelberg, 2004.

[172] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli , Image quality assessment: from error visibility to structural similarity", *IEEE Trans. on Img Proc.*, vol.13, no.4, pp.600-612, 2004.

[173] A. Eskicioglu, P. Fisher, Image quality measure and their performance, *IEEE Trans. on Comm.*, vol.43, no.12, pp. 2959-2965, 1995.

[174] N. Yang et al."A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval *J. Vis. Commun. Image R.*, vol.19. pp.92105, 2008.

[175] J. Vogelbernt, B. Schille, Semantic Modeling of Natural Scenes for Content-Based Image Retrieval, *Int. J. of Comput. Vis.* vol. 72, no.2, pp.133157, 2007.

[176] M. Choi, A New Intensity-Hue-Saturation Fusion Approach to Image Fusion With a Tradeoff Parameter, IEEE Transactions on geoscience and remote sensing, 44(2006).

[177] L. Bogoni, M. Hansen, Pattern-selective color image fusion, Pattern Recognition, 34(2001), 1515-1526.

[178] N. Mitianoudis,T. Stathaki, Optimal Contrast for Color Image Fusion using ICA bases,Proceedings of international conference on imformation fusion, 2008.

[179] L. Shen and Y. Niu, Blind Color Image Fusion Based on the Optimal Multi-objective Particle Swarm Optimization, International Journal of Multimedia and Ubiquitous Engineering, 2007

[180] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded up robust features. In ECCV, 2006.

[181] A. Toet, Natural colour mapping for multiband night vision imagery, Information Fusion, 4(2003)155-166.

[182] L. Wang, et al. Color fusion algorithm for visible and infra-red images based on color transfer in YUV color space. in: Proceedings of Multispectral Image Processing. Wuhan, China, 2007.

[183] Y. Chen, Z. Xue, R. S. Blum, Theoretical analysis of an information-based quality measure for image fusion, Information Fusion, 9(2)(2008), 161-175.

[184] T. D. Dixon et al. Selection of image fusion quality measures:objective, subjective, and metric assessment, Journal of Optical Society of America, 2007.

[185] Y. Zheng, Z. Qin, Objective Image Fusion Quality Evaluation Using Structural Similarity, Tsinghua Science and Technology, 703-709, 2009.

[186] P. J. Burt and E. H. Adelson, Merging images through pattern decomposition, *In Proc.* SPIE, 173-182, 1985.