

CERTIFICATE

It is certified that the work contained in the thesis titled "*Video-based Human Action Recognition in the Wild using Deep Learning*" by *Nitika Nigam* has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all requirements of Comprehensive Examination, Candidacy, and SOTA for the award of Ph.D. Degree.



Supervisor

Dr. Tanima Dutta

Assistant Professor,

Department of Computer Science and Engineering,

Indian Institute of Technology (BHU) Varanasi,

Uttar Pradesh, INDIA 221005.

DECLARATION BY THE CANDIDATE

I, *Nitika Nigam*, certify that the work embodied in this Ph.D. thesis is my own bonafide work carried out by me under the supervision of *Dr. Tanima Dutta* from *July 2018* to *August 2023* at *Department of Computer Science and Engineering*, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully copied any other's work, paragraphs, text, data, results, *etc.* reported in journals, books, magazines, reports, dissertations, thesis, *etc.*, or available at websites and have not included them in this thesis and have not cited as my own work.

Date: 10/2/23
Place: Varanasi


(Nitika Nigam)


CERTIFICATE BY THE SUPERVISOR

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.



(Dr. Tanima Dutta)

Dept. of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi


10/02/23
Signature of Head of Department (Actg)
Dept. of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Video-based Human Action Recognition in the Wild using Deep Learning

Name of the Student: Nitika Nigam

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the *Doctor of Philosophy*.

Date: 10/2/23

Place: Varanasi

Nitika.
(Nitika Nigam)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

Dedicated to my parents,

Mrs. Varsha Nigam

and

Mr. Raj Narain Nigam

ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor, Dr. Tanimia Dutta, for her valuable asset and assistance. I regard it as tremendous happiness to express my profound sense of gratitude and sincere regard for her constant feedback and expertise during all these years. I am eternally grateful to have had the opportunity to work on my thesis under her supervision. She supported my research directions and allowed me to explore new ideas in the domain of computer vision and deep learning. My cordial thanks to all the members of the Department of Computer Science and Engineering for creating an excellent working atmosphere. I would also like to thank the other members of my Doctoral committee, Dr. Bhaskar Biswas, Department of Computer Science and Engineering, and Dr. Rajeev, Department of Mathematical Sciences, for their help and support throughout the tenure of my studies. Special thanks to Dr. Hari Prabhat Gupta for his consistent assistance in both work and life aspects. I would also like to convey my sincere gratitude to Dr. S.K Singh, Head of Department of Computer Science and Engineering and all the RPEC and DPGC members for their suggestions and endorsement to this work.

I am grateful to Deepali Verma, Rahul Mishra, Ashish Gupta, Randheer Bagi, and all my dear lab members for the long discussions and their brilliant insights that have helped me to overcome the challenges I have faced in the development of this work. I also want to thank my dear friends Suncha Verma, Pratima Yadav, Paras, Nitish, Lokendra, Nistha Dubey, Kritika Singh, Nidhi Chaubey, and Darshna for supporting me in my hard times.

Finally, I express my heartfelt gratitude to my parents Mrs. Varsha Nigam and Mr. Raj Narain Nigam, and my sister Shipra Nigam for their constant support, love, encouragement, and sacrifices. Their affectionate love and care cannot be expressed in words.

(Nitika Nigam)

Contents

List of Figures	xii
List of Tables	xv
List of Symbols	xvii
List of Abbreviations	xix
Preface	xxi
1 Introduction	1
1.1 Definition	2
1.2 Applications	3
1.3 Challenges	4
1.4 Problems	5
1.5 Contributions of the Thesis	6
1.6 Organization of Thesis	7
2 Related work	10
2.1 CNN-based for Action Recognition	10
2.2 Human action recognition based on RNN-LSTMs	16
2.3 Other deep architectures for human action recognition	17
2.4 Conclusion	19
3 Action Recognition in Presence of Representation Bias	20
3.1 Our Contribution	23
3.2 Organization of the Chapter	24
3.3 Literature Survey	24
3.4 Proposed Approach	26
3.4.1 Multi-scale Deformable Backbone (MDB) Network	27

3.4.2	Actor-Object-Scene Attention (AOSA) Network	32
3.4.3	Long-Range Context (LRC) Network	38
3.4.4	Class-aware Temporal Attention Pooling (CTAP) Network	42
3.4.5	Joint Training Model	44
3.5	Experimental Results	45
3.5.1	Datasets and Metrics	46
3.5.2	Implementation Details	49
3.5.3	Ablation Study	51
3.5.4	Comparison with State-of-the-Art	57
3.6	Conclusion of the Chapter	63
3.7	Publication related to the Chapter	64
4	Human Behaviour Traits aware Action Recognition	65
4.1	Our Contribution	67
4.2	Organization of the Chapter	68
4.3	Literature Survey	69
4.4	Proposed Approach	70
4.4.1	Backbone Network	71
4.4.2	Visual Attention Network (VAN)	73
4.4.3	Long-term Attention Network (LAN)	75
4.4.4	Temporal Attention Pooling (TAP)	79
4.4.5	Joint Training Mechanism	80
4.5	Experimental Results	82
4.5.1	Datasets and Metrics	82
4.5.2	Implementation Details	85
4.5.3	Ablation Study	85
4.5.4	Comparison with SOTA on benchmark datasets	89
4.6	Conclusion of the Chapter	89
4.7	Publication related to the Chapter	91
5a	Action Recognition in the Wild: Fall Action	92
5a.1	Our Contribution	95
5a.2	Organization of the Chapter	96
5a.3	Literature Survey	96
5a.4	Proposed Approach	97
5a.4.1	Feature Extraction Module	98
5a.4.2	Uncertain-Action Supervision (US) Module	99

5a.4.3 Uncertain-Certain Action (UC) Classifier	104
5a.4.4 Joint Training Model	105
5a.5 Experimental Results	106
5a.5.1 Dataset details	106
5a.5.2 Implementation Details	108
5a.5.3 Ablation Study and Qualitative Analysis	110
5a.5.4 Comparison with State-of-the-Art	117
5a.6 Conclusion of the Chapter	118
5a.7 Publication related to the Chapter	118
5b Action Recognition in the Wild: Unusual Action	119
5b.1 Challenges in this work	122
5b.2 Our Contribution	123
5b.3 Organization of the Chapter	124
5b.4 Literature Survey	125
5b.5 Proposed Approach	126
5b.5.1 Feature Extraction Module: Light-weight ShuffleNet	127
5b.5.2 Fine-grained Human-Object Relation Module	129
5b.5.3 Federated Contrastive Learning (FCL) setup	134
5b.6 Experimental Results	136
5b.6.1 Publicly available Dataset	137
5b.6.2 Our UnusualAction Dataset	137
5b.6.3 Implementation Details	138
5b.6.4 Ablation Study and Qualitative Analysis	139
5b.6.5 Results on Habitual Actions Dataset	146
5b.7 Conclusion of the Chapter	148
5b.8 Publication related to the Chapter	149
6 Multi-label Complex Action Recognition	150
6.1 Our Contribution	153
6.2 Organization of the Chapter	154
6.3 Literature Survey	154
6.4 Proposed Approach	155
6.4.1 3D ShuffleNet Backbone Network	156
6.4.2 Kronecker-based Permutation Invariant Transformer	157
6.5 Experimental Results	163
6.5.1 Datasets and Metrics	163

6.5.2	Implementation Details	166
6.5.3	Ablation Study	167
6.5.4	Comparison with State-of-the-Art on Benchmark datasets . . .	172
6.6	Conclusion of the Chapter	172
6.7	Publication related to the Chapter	174
References		177
List of Publications		197

List of Figures

1.1	Action Labels	2
1.2	Video-based Action	2
3.1	Ambiguous actions due to representation bias	22
3.2	Overview of FactorNet	26
3.3	Overview of Multi-scale Deformable Backbone Network	28
3.4	Attention Map w/ and w/o Deformable Convolution	32
3.5	Similar Action Pattern with different Contextual Elements	33
3.6	Architecture overview of AOSA network	36
3.7	Accuracy vs. number of layers in Fish-body and Fish-head blocks . . .	53
3.8	Accuracy vs. number of hidden states	55
3.9	Visualization of actor, object, and scene attention maps of AOSA network	58
3.10	Visualization of sigmoid, softmax, object and scene attention map . . .	59
3.11	Performance of different modules on FactNet	60
4.1	Video-based actions: multiple human traits	66
4.2	Overall architecture of our HAANet	70
4.3	Architecture of Long-Range Attention Network	75
4.4	Mean accuracy of individual modules in VALC dataset	83
4.5	The confusion matrix of VALC dataset	84
4.6	Attention maps	88
5a.1	Certain and Uncertain Human Actions	93
5a.2	Overview of FallNet	98
5a.3	Performance of different modules on FallAction dataset	107
5a.4	Class activation maps for certain and uncertain action classes	116
5b.1	Habitual vs. Unusual Action	120
5b.2	Illustration of UnusualFedNet	127

5b.3	Functional overview of UBNet	128
5b.4	Action: grinding a phone	130
5b.5	Illustration of local training in UnusualFedNet	133
5b.6	Impact of number of edge devices	142
5b.7	Impact of convergence rounds with and without edge device	143
5b.8	Visualization of activation maps	144
6.1	Challenges in Complex Action Recognition.	150
6.2	Overall architecture of TIKNet	156
6.3	The overall architecture of KP-former.	158
6.4	Representation TIKNet modules on CompositeNet	165
6.5	mAP vs. training samples	171
6.6	Attention map	171

List of Tables

3.1	Architecture of fish-tail block of backbone network	31
3.2	Statistics of the datasets used in experimentation	46
3.3	Statistics of FactNet dataset	48
3.4	Confusion matrix of FactNet dataset	49
3.5	Effect of Temporal Range	52
3.6	Effect of backbone network variants	52
3.7	Effect of Multi-scale Deformable Backbone Network	53
3.8	Effect of deformable convolution in backbone network	53
3.9	Effect of actor and object branches of AOSA network	54
3.10	Comparison of AOSA	54
3.11	Effect of actor-object and scene branch	55
3.12	Comparison of temporal pooling mechanism	56
3.13	Effect of individual modules	56
3.14	Comparison on UCF101 and HMDB51 datasets	61
3.15	Comparison on Kinetics400 dataset	62
3.16	datasets on Breakfast-Actions and ActivityNet datasets	62
3.17	Comparison on Something-Something V1 dataset	63
4.1	Architecture of proposed SBN	71
4.2	Statistics of VALC dataset	83
4.3	Comparison on VALC dataset	84
4.4	Effect of temporal length	86
4.5	Effect of 3DResNet and C3D network as SBN	86
4.6	importance of FE, gesture, and FE + gesture	87
4.7	Impact of temporal pooling mechanisms	87
4.8	Effect of different modules	88
4.9	Comparison on short-term videos	89
4.10	Comparison on Kinetics400 dataset	90

4.11 Comparison on long-term videos	90
5a.1 Comparison on FallAction dataset	109
5a.2 Confusion matrix of FallAction dataset	110
5a.3 Performance of individual stage of our network	111
5a.4 Impact on computational complexity	111
5a.5 Comparison of computational complexity	112
5a.6 Classification accuracy	112
5a.7 Comparison of dilation rate	112
5a.8 Effect of different loss functions	113
5a.9 Impact of Max-pooling	113
5a.10 Effect of supervision network	113
5a.11 Effectiveness of FallNet in balancing the data	113
5a.12 Different aggregation strategies in TCDC layers	114
5a.13 Comparison on number in TCDC layers	115
5a.14 Comparison on OOPS dataset	115
5a.15 Comparison on Kinetics-600 dataset	115
5a.16 Comparison on HMDB51 dataset	117
5b.1 Comparison of computational complexity	140
5b.2 Impact on computational complexity of UBNet	140
5b.3 Impact of fusion strategy	141
5b.4 Performance on iid and non-iid FCL setup	142
5b.5 Effectiveness of λ in local objective function	145
5b.6 Effectiveness of \mathcal{L}_{cls} , \mathcal{L}_{icon} , and \mathcal{L}_{uaf}	146
5b.7 Comparison on EPIC-Kitchens (centralized)	148
5b.8 Comparison on EPIC-Kitchens (federated)	148
5b.9 Comparison on SSV2 (centralized)	148
5b.10 Comparison on SSV2 (federated)	149
6.1 The confusion matrix of CompositeNet dataset performed by TIKNet.	164
6.2 Performance of current state-of-the-art composite-actions methods on CompositeNet dataset.	164
6.3 Effect of length of RGB frames as an input in our TIKNet on benchmark datasets, <i>i.e.</i> , Charades and CompositeNet in terms mAP and mean accuracy.	168

6.4	Performance of TIKNet with different backbone networks (3DResNet, I3D (RGB only), and C3D network) on Breakfast-Actions and CompositeNet datasets in terms of mean accuracy.	168
6.5	Comparison of different decomposition methods on Charades, Breakfast-Actions, and CompositeNet datasets in terms of mAP and mean accuracy.	169
6.6	Comparison of parameters and FLOPs of different decomposition methods on Charades and Breakfast-Actions datasets in terms of mAP and mean accuracy.	169
6.7	Comparison of number of heads on Charades, Breakfast-Actions, and CompositeNet datasets in terms of mAP and mean accuracy.	170
6.8	Impact of standalone modules in TIKNet on Charades, Breakfast-Actions, and CompositeNet datasets (mAP and mean accuracy).	170
6.9	Comparison on Charades dataset in terms of mAP.	173
6.10	Comparison on Breakfast-Actions and Multi-THUMOS datasets for RGB modality in terms of mean accuracy and mAP.	173

List of Symbols

Symbol	Description
H or \mathcal{H}	Height of feature map
W or \mathcal{W}	Width of feature map
C or \mathcal{C}	Channels of feature map
t	Timestep
T or \mathcal{T}	Temporal length
i, j, k	Temporary Variables
epr	Facial expression
gst	Limb Gesture
max	Maximum value
min	Minimum value
tanh	tanh function
att	attention
ReLU	ReLU function
W	Dilated kernel
\mathfrak{W}	Kernel
nC_r	Combination
wce	Weighted Cross-entropy Loss
wfl	weighted focal-loss
λ	hyperparameters
η	learning rate
softmax or $\delta(\cdot)$	Softmax function
sigmoid or $\sigma(\cdot)$	Sigmoid function
m	Input modulation gate
i	Input gate
o	Output gate
h	Hidden state
o	Hadamard product

Symbol	Description
*	Convolution operation
\otimes	Kronecker product

Abbreviations

Abbreviation	Description
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DL	Deep Learning
LSTM	Long short-term memory
SVM	Support Vector Machine
AR	Action Recognition
RPN	Region Proposal Network
ReLU	Rectified Linear Unit
GRU	Gated recurrent unit
NMS	Non-maximal suppression
IoU	Intersection-over-union
maxpool	Maxpool Operation
FCN	Fully Convolutional Network
GT	Ground Truth
AP	average pooling
MDB	Multi-scale Deformable Backbone
AOSA	Actor-Object-Scene Attention
LRC	Long-Range Context
CTAP	Class-aware Temporal Attention Pooling
SBN	Shallow Backbone Network
VAN	Visual Attention Network
LAN	long-term attention network
TAP	temporal attention pooling
HAANet	Human Action Attention Network
GMP	Global Max-Pool
TDC	Temporal Deformable Convolution Layer
Spatial AP	Spatial Average Pooling

Abbreviation	Description
US	Uncertain-Action Supervision
TCDC	Temporal cascaded deformable convolution
SN	Supervision network
SSV2	Something-Something V2
SOTA	State-Of-The-Art
BN	Batch Normalization
FCL	Federated contrastive learning
HRCAs	Human-robot collaboration assembly
MHA	Multi-head attention
FFN	Feed-forward network
FDL	Federated deep learning
GNN	Graph neural network
IoT	Internet of Things
Acc.	Mean accuracy
mAP	Mean average precision
OFB	Object filtering branch
FC	Fully-connected