Activity Recognition Combined with Scene Context and Action Sequence

Sameera Chandimal Ramasinghe

(148060H)

Thesis submitted in partial fulfillment of the requirements for the degree
Master of Philosophy

Department of Electronic and Telecommunication Engineering

University of Moratuwa
Sri Lanka

September 2017
To my parents, for their endless love and encouragement. Also to the stillness of universe, which wipes out all the boundaries.
DECLARATION

I declare that this is my own work, and this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other university or institute of higher learning, and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part, in print, electronic, or any other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The candidate, whose signature appears above, carried out research for the MPhil dissertation under my supervision.

Signature: Date
ABSTRACT

In this study, we investigate the problem of automatic action recognition and classification of videos. First, we present a convolutional neural network architecture, which takes both motion and static information as inputs in a single stream. We show the network is able to treat motion and static information as different feature maps and extract features off them, even though stacked together. By our results, we justify the use of optic flows as the raw information of motion. We demonstrate that our network is able to surpass state-of-the-art hand-engineered feature methods. Furthermore, the effect of providing static information to the network, in the task of action recognition, is also studied and compared here. Then, a novel pipeline is proposed, in order to recognize complex actions. A complex activity is a temporal composition of subevents, and a sub-event typically consists of several low level micro-actions, such as body movement, done by different actors. Extracting these micro actions explicitly is beneficial for complex activity recognition due to actor selectivity, higher discriminative power, and motion clutter suppression. Moreover, considering both static and motion features is vital for activity recognition. However, how to control the contribution from each feature domain optimally still remains uninvestigated. In this work, we extract motion features in micro level, preserving the actor identity, to later obtain a high-level motion descriptor using a probabilistic model. Furthermore, we propose two novel schemas for combining static and motion features: Cholesky transformation based and entropy based. The former allows to control the contribution ratio precisely, while the latter uses the optimal ratio mathematically. The ratio given by an entropy based method matches well with the experimental values obtained by a Choleksy transformation based method. This analysis also provides the ability to characterize a dataset, according to its richness in motion information. Finally, we study the effectiveness of modeling the temporal evolution of sub-event using an LSTM network. Experimental results demonstrate that the proposed technique outperforms state-of-the-art, when tested against two popular datasets.

Key words—Human action recognition; Convolutional Neural Networks (CNN); Recurrent Neural Networks (RNN); Long Short-Term Memory (LSTM); Dense trajectories; BoVW
ACKNOWLEDGEMENT

I would first like thank my supervisor, Dr. Ranga Rodrigo, Department of Electronic and Telecommunication Engineering, University of Moratuwa, for his continuous guidance and tremendous support throughout this program. I would also like to express my deepest gratitude to Dr. Ajith Pasqual, Department of Electronic and Telecommunication Engineering, University of Moratuwa, for his overall supervision and valuable advises.

Furthermore, I would also like to thank my progress review panel members, Dr. Lochandaka Ranathunga and Dr. Chandika Wavegedara for their continuous suggestions and comments to improve the research work.

Also, I need to thank my fellow research partners, Mr. Jathushan Rajasegaran, Mr. Vinoj Jayasundara, Mr. Kanchana Ranasinghe and Mrs. Manosha Chathuramali for their tremendous support in carrying out research and experiments.

Moreover, I express my deep gratitude to the National Research Council of Sri Lanka for funding this research under grant 12-018.

Finally, I would also like to thank my family members for their invaluable support throughout my M.Phil. journey.
Table of Contents

DECLARATION i
ABSTRACT ii
ACKNOWLEDGEMENT iii
Table of Contents vii
List of Tables ix
List of Figures xi
ABBREVIATIONS xii

1 INTRODUCTION 1
1.1 Focus of the Thesis 3
1.2 Original Contributions 4
1.3 Thesis Structure 5
1.4 Publications 7

2 LITERATURE REVIEW 9
2.1 Classification of Activity Recognition Models 9
  2.1.1 Supervised Feature Engineered Models 9
  2.1.2 Deep Learning Based Models 18
2.2 Comparison with Closely Related State-of-the-art Work. . . . . . . . 20
2.3 Review of Important Concepts. . . . . . . . . . . . . . . . . . . . . 23
   2.3.1 Convolutional Neural Nets . . . . . . . . . . . . . . . . . . 23
   2.3.2 Recurrent Neural Networks (RNN) . . . . . . . . . . . . . . 26
   2.3.3 Long Short Term Memory (LSTM) . . . . . . . . . . . . . . 27
   2.3.4 Dense Trajectories . . . . . . . . . . . . . . . . . . . . . . . 28

3 UNSUPERVISED EXTRACTION AND FUSION OF MOTION AND STATIC
   DESCRIPTORS 30
   3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
   3.2 Architecture . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
      3.2.1 Optimization of the Network . . . . . . . . . . . . . . . . . 32
   3.3 Methodology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
      3.3.1 Enlargement of the Dataset. . . . . . . . . . . . . . . . . . . 32
      3.3.2 Stacked Motion and Static Information for Representing Video
           Segments . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
      3.3.3 Calculation of Dense Optic Flows . . . . . . . . . . . . . . . 33
      3.3.4 Stacking of Static Information. . . . . . . . . . . . . . . . . 34
      3.3.5 Data Augmentation. . . . . . . . . . . . . . . . . . . . . . . 34
      3.3.6 Initialization of Weights . . . . . . . . . . . . . . . . . . . . 35
   3.4 Results and Comparison. . . . . . . . . . . . . . . . . . . . . . . . 35
      3.4.1 Approach 1. . . . . . . . . . . . . . . . . . . . . . . . . . . . 35
      3.4.2 Approach 2 . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
      3.4.3 Conclusion and Discussion . . . . . . . . . . . . . . . . . . 36

4 EXPERIMENTS ON RICH LOCAL MOTION DESCRIPTORS 40
   4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 40
   4.2 Methodology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
      4.2.1 Manually Annotating the Temporal and Special Locations of
           Strokes . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45
      4.2.2 Creating Dense Trajectories . . . . . . . . . . . . . . . . . . 47
4.2.3 Creating Cluster Centers for the Bag-of-visual-words Model. 47
4.2.4 Training .............................................. 49

4.3 Results and Evaluation ..................................... 49
4.3.1 Evaluation Approach 1 .................................. 49
4.3.2 Evaluation Approach 2 ................................. 50
4.3.3 Interpretation of Results. ............................... 50
4.3.4 Comparison with the State-of-the-art. ................. 52

4.4 Conclusion .................................................. 53

5 SUPERVISED FUSION OF MOTION AND STATIC FEATURES 54
5.1 Introduction .................................................. 54
5.2 Methodology ................................................ 57
  5.2.1 Overview ................................................. 57
  5.2.2 Motion Features ....................................... 60
  5.2.3 Static Features ....................................... 66
  5.2.4 Fusing of Static and Motion Features ............... 69
  5.2.5 Capturing Temporal Evolution ....................... 75

5.3 Experiments and Results ................................... 76
  5.3.1 Data-sets .............................................. 77
  5.3.2 Contribution of Static and Motion Domains .......... 77
  5.3.3 Mathematical Validation of Optimum Contribution .... 80
  5.3.4 Comparison of Fusion Models ......................... 80
  5.3.5 Comparison with the state-of-the-art ................. 80
  5.3.6 Effectiveness of Capturing Time Evolution .......... 84

5.4 Conclusion .................................................. 88

6 ENHANCEMENT OF THE ACTION RECOGNITION PIPELINE 90
6.1 Introduction .................................................. 90
6.2 Improved Motion Features ................................ 93
  6.2.1 Actor Localization ................................... 93
  6.2.2 Tracking proposed candidate areas .................. 93
6.2.3 Modified K-Means for BoW ........................................ 95
6.2.4 High Level Actions from Micro Actions ..................... 97
6.3 Entropy based Fusion of Motion and Static Vectors ........... 98
6.4 Experiments and Results ........................................... 99
   6.4.1 Comparison with the State-of-the-Art .................... 101
   6.4.2 Effectiveness of Capturing Time Evolution ............... 101

7 CONCLUSION AND FUTURE WORK ................................. 106
   7.1 Summary and Conclusion ....................................... 106
   7.2 Future Work .................................................... 110

References ..................................................................... 111
List of Tables

2.1 Previous surveys on activity recognition .................................................. 10

3.1 Comparison of our network with state-of-the-art algorithms. Accuracies reported over each class are compared ........................................ 37

3.2 Comparison of results: approach 1 vs approach 2. ................................... 37

4.1 Sample set of data written to the database while manually annotating the temporal and special locations of strokes. ............................... 46

4.2 Accuracy of each stroke-class in evaluation method 1. ............................ 50

4.3 Accuracy of each stroke-class in evaluation method 2. ............................ 50

4.4 Maximum and minimum number of frames belonging to each stroke class and their respective recognition accuracy. ............................ 52

4.5 Maximum and minimum number of frames belonging to each stroke class and their respective recognition accuracy. ............................ 53

5.1 Derivation of $\rho$ values for different contribution levels of static and motion domains to the fused vector ................................................ 78

5.2 Overall accuracy of UCF-11, Hollywood2, and HMDB51 for varying ratios between static and motion components. ............................ 78

5.3 Per-class accuracy for different contribution of static and motion vectors for UCF-11 .............................................................. 79
5.4 mAP for each class for different contribution of static and motion vectors to the fused vector for Hollywood2. ............... 79
5.5 Comparison of fusion models on UCF-11 dataset. ................. 83
5.6 Comparison of fusion models on Hollywood2 dataset .............. 83
5.7 Comparison of our method with state-of-the-art methods in the literature. 84
5.8 Per-class accuracy comparison with state-of-the-art on UCF-11. . . . 85
5.9 Per-class mAP comparison with state-of-the-art on Hollywood2. . . 85

6.1 Per-class accuracy comparison with state-of-the-art on UCF-11 (percent accuracy values). ........................................ 97
6.2 Per-class accuracy for different contribution of static and motion vectors for UCF-11. .................................................. 100
6.3 AP for each class for different contribution of static and motion vectors to the fused vector for Hollywood2. .......................... 100
6.4 Comparison of our method with the state-of-the-art methods in the literature. ......................................................... 101
6.5 Per-class accuracy comparison with state-of-the-art on UCF-11 (percent accuracy values). ........................................ 102
6.6 Per-class mAP comparison with state-of-the-art on Hollywood2 (average precision values). ................................. 105
List of Figures

2.1 Sparse connectivity of a convolutional neural network .......................... 24
2.2 Illustration of shared weights of a convolutional neural network. ......... 24
2.3 Illustration of a convolution layer of a convolutional neural network. 
   Layer $m - 1$ consists of four feature maps and layer $m$ consists of two 
   feature maps. $W_1$ and $W_2$ are weight matrices of the two convolutional 
   windows respectively. ............................................. 25
2.4 Long short-term memory (LSTM) block cell. Source [1]. ............... 27
2.5 Illustration of the dense trajectory description. ................................. 29
3.1 CNN architecture used for generating static features. ....................... 39
4.1 Example image sequences for each class. Top row: backhand, second 
   row: forehand, third row: smash, last row: other ............................ 44
4.2 Upper diagonal camera angle .................................................. 46
4.3 Generation of trajectory aligned HOF features. ................................. 48
4.4 Visualization of dense trajectories created for a backhand stroke. .... 48
4.5 Images of mis-recognized strokes .............................................. 51
5.1 Overall methodology. .......................................................... 59
5.2 Example HOOF generation with 6 bins ......................................... 66
5.3 Generation of the final motion descriptor. Left: the dictionary is created from the HOOF pool. Right: The motion descriptor is generated using the dictionary ........................................ 67
5.4 CNN architecture used for generating static features. ............... 68
5.5 Percent standard deviation values for the first and second components of the PCA. \( n \) is the feature vector index. ......................... 74
5.6 A simple illustration of the LSTM network. .......................... 76
5.7 The process of feeding fused vectors to the LSTM network. \( c_i \) indicates the fused vector representing the \( i_{th} \) video segment. ............... 76
5.8 Accuracy distribution for different contribution levels of motion and static domains in UCF-11. ................................. 81
5.9 Accuracy distribution for different contribution levels of motion and static domains in Hollywood2. .......................... 82
5.10 Accuracy comparison between Random Forest Classifier and LSTM for UCF-11 dataset. .......................... 86
5.11 mAP comparison for Random Forest Classifier and LSTM for Hollywood2 dataset. .......................... 87
6.1 Overall methodology .................................................. 92
6.2 Initialization of candidate areas. .................................... 96
6.3 Modeling micro actions independently for each actor. ............. 96
6.4 Accuracy comparison between Random Forest Classifier and LSTM for UCF-11 dataset. .......................... 103
6.5 mAP comparison for Random Forest Classifier and LSTM for Hollywood2 dataset. .......................... 104
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SURF</td>
<td>Speeded Up Robust Features</td>
</tr>
<tr>
<td>BOVW</td>
<td>Bag-of-Visual-Words</td>
</tr>
<tr>
<td>BOW</td>
<td>Bag-of-Words</td>
</tr>
<tr>
<td>CNN</td>
<td>Convolutional Neural Net</td>
</tr>
<tr>
<td>CRF</td>
<td>Conditional Random Fields</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Models</td>
</tr>
<tr>
<td>HMM</td>
<td>Hidden Markov Models</td>
</tr>
<tr>
<td>HOF</td>
<td>Histograms of Optical Flow</td>
</tr>
<tr>
<td>HOG</td>
<td>Histograms of Oriented Gradients</td>
</tr>
<tr>
<td>KNN</td>
<td>K-Nearest Neighbour</td>
</tr>
<tr>
<td>LSTM</td>
<td>Long Short Term Memory</td>
</tr>
<tr>
<td>mAP</td>
<td>Mean Average Precision</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>STIP</td>
<td>Spatio Temporal Interest Points</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>DTF</td>
<td>Dense Trajectory based Features</td>
</tr>
<tr>
<td>SIFT</td>
<td>Scale-Invariant Feature Transforms</td>
</tr>
<tr>
<td>MBH</td>
<td>Motion Boundary Histogram</td>
</tr>
<tr>
<td>MIL</td>
<td>Multiple Instance Learning</td>
</tr>
<tr>
<td>MISL</td>
<td>Multiple Instance Single Label</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>PLSA</td>
<td>Probabilistic Latent Semantic Analysis</td>
</tr>
<tr>
<td>RNN</td>
<td>Recurrent Neural Networks</td>
</tr>
<tr>
<td>FCN</td>
<td>Fully Convolutional Neural Nets</td>
</tr>
<tr>
<td>IDT</td>
<td>Improved Dense Trajectories</td>
</tr>
<tr>
<td>ARCH</td>
<td>Adaptive Recurrent-Convolutional Hybrid network</td>
</tr>
</tbody>
</table>