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 trajecories;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

List of Tables

List of Figures

ABBREVATIONS

SURF	=	Speeded Up Robuts Features
BOVW	=	Bag-of-Visual-Words
BOW	=	Bag-of-Words
CNN	=	Convolutional Neural Net
CRF	=	Conditional Random Fields
GMM	=	Gaussian Mixture Models
HMM	=	Hidden Markov Models
HOF	=	Histograms of Optical Flow
HOG	=	Histograms of Oriented Gradients
KNN	=	K-Nearest Neighbour
LSTM	=	Long Short Term Memory
mAP	=	Mean Average Precision
PCA	=	Principal Component Analysis
STIP	=	Spatio Temporal Interest Points
SVM	=	Support Vector Machines
DTF	=	Dense Trajectory based Features
SIFT	=	Scale-Invariant Feature Transforms
MBH	=	Motion Boundary Histogram
MIL	=	Multiple Instance Learning
MISL	=	Multiple Instance Single Label
RBF	=	Radial Basis Function
PLSA	=	Probabilistic Latent Semantic Analysis
RNN	=	Recurrent Neural Networks
FCN	=	Fully Convolusional Neural Nets
IDT	=	Imporoved Dense Trajectories
ARCH	=	Adaptive Recurrent-Convolutional Hybrid network