



Faculty of Engineering and Information Technology  
School of Computing and Communications

# Action Recognition and Video Summarisation by Submodular Inference

THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF  
DOCTOR OF PHILOSOPHY

Principal Supervisor: Prof. Massimo Piccardi

Candidate: Fairouz Hussein

APRIL, 2017

### ***Certificate of Original Authorship***

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Fairouz Farouq Fayiz Hussein

16-April-2017

## Abstract

In the field of computer vision, action recognition and video summarisation are two important tasks that are useful for applications such as video indexing and retrieval, human-computer interaction, video surveillance and home intelligence. While many approaches exist in the literature for these two tasks, to date they have always been addressed separately. Instead, in this thesis we move from the assumption that action recognition can usefully drive the selection of frames for the summary and that recognising actions from a summary can prove more accurate than from the whole video, and therefore the two tasks should be tackled simultaneously as a joint objective. To this aim, we propose a novel framework based on structured max-margin algorithms and an efficient model for inferring the action and the summary based on the property of submodularity. Recently, submodularity has emerged as an area of interest in machine learning and theoretical computer science, particularly within the domains of optimisation and game theory and is therefore one of the main frameworks for this thesis. To ensure proper exploitation of the proposed method, we have conducted experiments in three different kinds of scenarios: unsupervised summaries, semi-supervised summaries and fully supervised. We also propose a novel loss function - V-JAUNE - to evaluate the quality of a predicted video summary against the summaries annotated by multiple annotators. In a last experiment, we leverage the proposed loss function not only for evaluation, but also for the training stage. The effectiveness of the proposed algorithms is proved using qualitative and quantitative tests on two challenging depth action datasets: ACE and MSR DailyActivity. The results show that the proposed approaches are capable of learning accurate action classifiers and produce informative summaries.

*Dedicated to my sweet and loving family*

## ***Acknowledgements***

Sincere feelings and strongest kind words emanating from my heart go to my supervisor Professor Massimo Piccardi. I present him with my most heartfelt thanks and gratitude together with respect and appreciation. He has given a lot to support my thesis and he is still offering his time and thoughts pro-actively and gladly, without waiting for praise or thanks. I consider myself very lucky to have had a supervisor like him who is known for his wonderful experience, creative assistance, and distinctive presence.

I would like to extend my thanks and gratitude to Tareq, my beloved husband, who is the reason for the continuation and completion of my studies, who stood by me in the toughest conditions and encouraged me to persevere and continue and to not give in to despair.

I am also thankful to my parents, my lovely kids Marah, Abdullah, Leen and Joury, and my sisters and brothers who have given me their love and care. I ask God to bless them with good health, happiness and faith.

Finally, I would like to thank all my colleagues and friends - Shaukat, Sari, Khalid, Subheih, Rana, Majeda, Dana, Ali, Raniah, Arwa, Hanadi, Hayat, and Ala'a - who filled my time at UTS with smiles and support.

# Contents

<b>Abstract</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation and Objectives . . . . .	5
1.2 Research Questions . . . . .	6
1.3 Contributions . . . . .	8
1.4 Organisation of the thesis . . . . .	9
1.5 Publications . . . . .	10
<b>2 Background and Related work</b>	<b>11</b>
2.1 Introduction . . . . .	11
2.2 Action Recognition Approaches . . . . .	11
2.3 Local Representations . . . . .	13
2.3.1 Feature detectors . . . . .	14
2.3.2 Feature descriptors . . . . .	14
2.3.3 Feature representations . . . . .	15
2.4 Types of Features . . . . .	16
2.4.1 Colour-based features . . . . .	16

2.4.2	Skeleton-based features . . . . .	16
2.4.3	Depth-based features . . . . .	17
2.5	Action classification models . . . . .	17
2.5.1	Rule-based methods . . . . .	17
2.5.2	Probabilistic methods . . . . .	18
2.6	Learning . . . . .	20
2.6.1	Supervised learning . . . . .	21
2.6.2	Unsupervised learning . . . . .	23
2.6.3	Semi-supervised learning . . . . .	24
2.7	Classification Methods . . . . .	27
2.7.1	k-NN . . . . .	27
2.7.2	SVM . . . . .	27
2.7.3	Multi-class SVM . . . . .	29
2.7.4	Structural SVM . . . . .	31
2.7.5	Main applications of SSVM . . . . .	32
2.7.6	Latent structural SVM . . . . .	33
2.7.7	Formulation . . . . .	35
2.8	Submodular Functions . . . . .	37
2.8.1	Why submodularity ? . . . . .	39
2.9	Video Summarisation and Evaluation . . . . .	40
2.9.1	Video summarisation approaches . . . . .	40
2.9.2	Video summarisation evaluation . . . . .	40
2.10	Video Summarisation and Submodular Functions . . . . .	42

2.10.1 Formulation . . . . .	43
2.11 Action Recognition in Depth Videos . . . . .	43
2.12 Datasets . . . . .	44
2.12.1 ACE . . . . .	45
2.12.2 MSRDailyActivity3D . . . . .	47
2.12.3 MSR Action3D . . . . .	49
<b>3 Joint action recognition and summarisation</b>	<b>51</b>
3.1 Introduction and Related Work . . . . .	52
3.2 Recognition and summarisation by submodular functions . . . . .	53
3.3 Learning: latent variables . . . . .	56
3.4 Experimental Results . . . . .	57
<b>4 V-JAUNE: A Framework for Joint Action Recognition and Video summarisation</b>	<b>63</b>
4.1 Introduction . . . . .	63
4.2 Related Work . . . . .	65
4.3 Learning Framework . . . . .	69
4.3.1 Model Formulation . . . . .	69
4.3.2 Latent Structural SVM for Unsupervised and Semi-Supervised Learning	72
4.4 V-JAUNE: Video Summary Evaluation . . . . .	75
4.5 Experimental results . . . . .	79
4.5.1 ACE . . . . .	80
4.5.2 MSR DailyActivity3D . . . . .	85

<b>5 Minimum Risk Structured Learning of Video Summarisation</b>	<b>89</b>
5.1 Introduction and Related Work . . . . .	89
5.2 Summarisation via structured learning . . . . .	91
5.2.1 Problem Formulation . . . . .	91
5.2.2 Structural SVM for Supervised Learning . . . . .	93
5.2.3 Learning with V-JAUNE . . . . .	94
5.3 V-JAUNE for Evaluation . . . . .	95
5.4 Experimental results . . . . .	99
5.4.1 ACE . . . . .	99
5.4.2 MSR . . . . .	103
<b>6 Conclusion</b>	<b>105</b>
<b>Bibliography</b>	<b>107</b>

# List of Tables

3.1	Comparison of action recognition accuracy on the MSR Daily Activity 3D dataset.	59
3.2	Sensitivity analysis of the accuracy with different weights in (3.6) and with depth and RGB data.	59
3.3	The accuracy achieved by Latent SSVM on depth data	59
4.1	Details of the ACE dataset.	80
4.2	Comparison of the action recognition accuracy on the ACE dataset.	81
4.3	The evaluation results on the ACE dataset using various amounts of supervision.	82
4.4	Influence of the budget on the action recognition accuracy for the ACE dataset.	84
4.5	Sensitivity analysis of the action recognition accuracy at the variation of the $\lambda$ parameters for the ACE dataset (unsupervised case).	84
4.6	The evaluation results on the MSR DailyActivity3D using various flavours of learning.	86
4.7	Comparison of the action recognition accuracy on the MSR DailyActivity3D dataset (depth frames only).	87
5.1	The values of V-JAUNE measure on the ACE dataset (clipped)	100
5.2	The values of V-JAUNE measure on the ACE dataset (unclipped)	100
5.3	The values of V-JAUNE measure on the MSR DailyActivity3D dataset.	103

# List of Figures

1.1	With Kinect games, the players are the controller.	3
1.2	Examples of actions in videos.	3
1.3	Example of a smart home (reprinted from [Simpson, 2016]).	4
1.4	Example of Input and Output of Video Summarisation from the ACE dataset.	5
1.5	Some challenges of action recognition.	6
2.1	Extraction of space-time cuboids at interest points from similar actions performed by different persons (reprinted from [Laptev et al., 2007]).	15
2.2	A generic machine learning system (reprinted from [Kadre and Konasani, 2015]).	21
2.3	Flavours of machine learning a) Fully-supervised learning; b) Unsupervised learning.	22
2.4	Binary Support Vector Machines on (a) linearly separable data and (b) non-linearly separable. Squares represent one class, circles the other one. Support vectors are laying on the margin.	30
2.5	Examples of structured problems.	34
2.6	The diminishing return property in a submodular set function.	37
2.7	A comparison between RGB channels and depth channels ( reprinted from [Wang et al., 2014a]).	45

2.8	A typical clip of ACE actions performed by five different actors (distinguishable by their clothing). . . . .	46
2.9	Some examples from the MSR DailyActivity3D (displayed as RGB and depth frames): the first column in each subfigure shows the subject standing close to the couch; the second, sitting on it. . . . .	48
2.10	Sample clips from the MSR Action3D for actions a) Draw tick and b) Tennis serve (reprinted from [Li et al., 2010]). . . . .	49
3.1	Summary examples (displayed as RGB frames) for action <i>walk</i> : a) proposed method; b) SAD. . . . .	60
3.2	Each row contains the summary of a video to represent a certain activity, the activities are: <i>drinking, eating, reading, using cell phones, writing, using computers/laptop, vacuuming, cheering up, sitting still, tossing crumpled paper, playing games, lying on the sofa, walking, playing the guitar, standing up, and sitting down</i> . . . . .	62
4.1	The graphical model for joint action classification and summarisation of a video: $y$ : action class label; $h$ : frames selected for the summary; $x$ : measurements from the video frames. . . . .	68
4.2	V-JAUNE values for the ACE test set (95 videos) with multiple annotators: blue bars: denormalised values; red bars: normalised values. . . . .	77
4.3	V-JAUNE loss for different annotators over the ACE test set (95 videos), using the first annotator as ground truth and the second as prediction. Please note that the changes in value are mainly due to the changes in magnitude of the VLAD descriptors. However, the agreement also varies with the video.	78
4.4	Examples of predicted summaries from the ACE dataset (displayed as RGB frames for the sake of visualisation). The subfigures display the following actions: a) <i>breaking</i> ; b) <i>baking (omelet)</i> ; c) <i>baking (ham)</i> ; and d) <i>turning</i> . In each subfigure, the first row is from the proposed method, the second from SAD. . . . .	83

4.5 Examples of summaries from the MSR DailyActivity3D dataset (displayed as RGB frames for ease of interpretation) for actions a) <i>Cheer</i> and d) <i>Walk</i> : in each subfigure, the first row is from the proposed method and the second from SAD. The results from the proposed method look more informative. . .	88
5.1 V-JAUNE values for the ACE test set for actions a) boiling, and b) seasoning, with multiple annotators: blue bars: denormalised values; red bars: normalised values. . . . .	97
5.2 V-JAUNE loss for different annotators for actions a) boiling, and b) seasoning, using the first annotator as ground truth and the second as prediction. Please note that the changes in value are mainly due to the changes in magnitude of the VLAD descriptors. However, the agreement also varies with the video.	98
5.3 Examples of predicted summaries from the ACE dataset (clipped). The subfigures display the actions a) <i>seasoning</i> ; and b) <i>peeling</i> . In each subfigure, the first row is from the proposed method, the second from SAD. . . . .	101
5.4 Examples of predicted summaries from the ACE dataset (unclipped). In each subfigure, the first row is from the proposed method, the second from SAD. . .	102
5.5 Examples of predicted summaries from the MSR DailyActivity3D dataset. The subfigures display the actions a) <i>using vacuum</i> ; and b) <i>playing guitar</i> . In each subfigure, the first row is from the proposed method, the second from SAD. . . . .	104