

Semi-supervised and Unsupervised Extensions to Maximum-Margin Structured Prediction



Shaukat ABIDI

Faculty of Engineering and Information Technology
University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

2016

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.

Student's Name : Shaukat ABIDI

Signature of Student :

Date : 18/July/2016

Abstract

Structured prediction is the backbone of various computer vision and machine learning applications. Inspired by the success of maximum-margin classifiers in the recent years; in this thesis, we will present novel semi-supervised and unsupervised extensions to structured prediction via maximum-margin classifiers.

For semi-supervised structured prediction, we have tackled the problem of recognizing actions from single images. Action recognition from a single image is an important task for applications such as image annotation, robotic navigation, video surveillance and several others. We propose approaching action recognition by first partitioning the entire image into “superpixels”, and then using their latent classes as attributes of the action. The action class is predicted based on a graphical model composed of measurements from each superpixel and a fully-connected graph of superpixel classes. The model is learned using a latent structural SVM approach, and an efficient, greedy algorithm is proposed to provide inference over the graph. Differently from most existing methods, the proposed approach does not require annotation of the actor (usually provided as a bounding box).

For the unsupervised extension of structured prediction, we considered the case of labeling binary sequences. This case is important in a detection scenario, where one is interested in detecting an action or an event. In particular, we address the unsupervised SVM relaxation recently proposed in (Li et al. 2013) and extend it for structured prediction by merging it with structural SVM. The main contribution of the proposed extension (named Well-SSVM) is a re-organization of the feature map and loss function of structural SVM that permits finding the violating labelings required by the relaxation. Experiments on synthetic and real datasets in a

fully unsupervised setting reveal a competitive performance as opposed to other unsupervised algorithms such as k-means and latent structural SVM.

Finally, we approached the problem of unsupervised structured prediction by M^3 Networks. M^3 Networks are an alternative formulation of maximum-margin structured prediction that can satisfy the complete set of constraints for decomposable feature and loss functions; hence, the entire set of constraints is considered during the search for the optimal margin as opposed to Structural SVM. In the thesis, we present the interpretation of M^3 Networks in Well-SSVM, thus allowing us to use in a semi-supervised and unsupervised scenario.

Acknowledgements

I would like to take this opportunity to acknowledge enormous support from my supervisors, very useful suggestions from my group and a friendly environment of our lab especially during coffees, lunches and dinners.

First of all, I would like to thank my principal supervisor Professor Mary-Anne Williams who provided me the opportunity to come to Australia for PhD. Without her continuous support and supervision, I would have never gotten a chance to explore beautiful practical applications of robotics and computer vision.

I would like to convey big thanks to my co-supervisor Professor Massimo Piccardi, who nurtured my technical knowledge for doctoral degree. His stream of ideas kept me occupied due to which I was able to explore amazing research path. His continuous technical support especially regular meetings, sometimes at coffee shops, has played a central role in the formulation of my technical abilities.

I am thankful to all of my friends, colleagues and group members of Magic Lab and Surveillance Lab: Dr. Benjamin Johnston, Dr. Xun Wang, Dr. Saleha Raza, Dr. Rony Novianto, Wei Wang, Pramod Parajulli, Jonathan Vitale, Mahya Mirzae, Ali Raza, Sylvan Rudduck, Nima Ramezani, Robert Lange, Sari Awwad, Fairouz Hussein, and Ava Bargi. I am thankful to Professor Ivor Tsang, who provided valuable insights and critical reviews for my work. I am grateful to the visiting professors of Magic lab who shared their research experience with me: Professor Pavlos Peppas, Professor Peter Gärdenfors and Associate Professor Sajjad Haider. I would like to thanks Benjamin Johnston and Xun Wang again, with whom I developed demos for robots.

In the end, I would like to thanks my parents without whom I would never be a person I am at the moment and will be in the future.

Contents

Certificate of Original Authorship	i
Abstract	ii
Acknowledgements	iv
List of Figures	viii
List of Tables	x
1 Introduction	1
1.1 Structured Prediction	2
1.2 Semi-supervised Structured Prediction	3
1.3 Unsupervised Structured Prediction	5
1.4 Contributions	6
1.5 Thesis Organization	6
2 Literature Review	8
2.1 Maximum Margin Classifiers	8
2.2 Support Vector Machines (Binary Case)	10
2.2.1 Intuitions for Margin	10
2.2.2 Hard-Margin SVM	12
2.2.3 Soft-Margin SVM	14
2.3 Multiclass Support Vector Machines	17
2.4 Multiple Kernel Learning	18
2.5 Structured Prediction	21
2.6 Still Image Action Recognition	29
2.6.1 Learning an Action Recognition Classifier	30
2.6.1.1 Action Representation	30

2.6.1.2	Global Features	30
2.6.1.3	Local Features	31
2.6.1.4	Learning a Classifier	31
2.6.2	Advantages of Still Image Action Recognition	32
3	Semi-Supervised Structured Prediction SVM and its Application for Static Action Recognition	33
3.1	Introduction	33
3.2	Action recognition by superpixel classification	36
3.2.1	The graphical model	37
3.2.2	Object detectors	38
3.3	Semi-supervised Latent structural SVM	40
3.3.1	Feature and score functions	41
3.3.2	Loss-augmented inference	42
3.3.3	Latent variables' initialization	43
3.3.4	Inference by efficient greedy algorithms	43
3.4	Experimental results	46
3.5	Conclusion	50
4	Unsupervised Structured Prediction SVM	53
4.1	Introduction	53
4.1.1	Well-SVM	55
4.1.2	Structural SVM	57
4.2	Weakly Labeled Structural SVM	58
4.2.1	Feature Maps	60
4.2.2	Finding a Violating Labeling	62
4.2.3	Optimized Matrix-Vector Multiplication (Hy)	65
4.2.4	Balanced Sequential Labeling	67
4.3	Experimental Results	68
4.3.1	Dataset Description	69
4.3.1.1	Synthetic Dataset	69
4.3.1.2	Gesture Phase Segmentation Dataset	70
4.3.2	Initialization	71
4.3.3	Performance Comparison	71
4.4	Conclusion	72
5	Unsupervised Structured Prediction Maximum Margin Markov Networks	74
5.1	Introduction	74
5.2	Notations	75
5.3	Factorized Dual	76

5.4	Decomposition of Feature Function over Edges	80
5.5	Solution of WellSSVM via M^3N	81
5.5.1	Learning as an Instance of MKL	81
5.5.2	Finding a Violating Labeling	83
5.5.3	Update μ	84
5.6	Experiments	84
5.6.1	Datasets Description	85
5.6.1.1	Synthetic Dataset	85
5.6.1.2	Gesture Phase Segmentation Dataset	86
5.6.2	Initialisation	86
5.6.3	Performance Comparison	86
5.7	Conclusion	87
6	Conclusion	88
A	Lagrange Duality	90
B	Well-SSVM: from primal (4.13) to dual (4.15)	93
C	WellSSVM via M^3N	95
C.1	Factorized Dual	96
C.2	Implementation of \mathcal{H} , τ and Δ	97
C.3	Pictorial Representation of h and \mathcal{H}	103
	Bibliography	105

List of Figures

1.1	Examples of complex objects: a) graph with 7 nodes; b) tree with 7 nodes where A is the root node and {E,D,F,G} are leaf nodes; c) a sequence with 4 nodes.	2
1.2	a) Action class: using a computer (Yao et al. (2011a)) b) Scene class: street (Xiao et al. (2010)).	5
1.3	An example of unsupervised training for sequence prediction with 5 output (shaded) nodes.	5
2.1	An example of linear discriminant function separating distinct regions of points in 2D-space	9
2.2	a) Functional margin of points A, B and C; b) Geometric margin define as a distance between point A and B.	10
2.3	Hard margin SVM: When data is separable by linear decision boundary.	13
2.4	Soft margin SVM: When data is non-separable by linear decision boundary, the case of overlapping classes.	15
2.5	Hinge loss (blue) as a convex surrogate of zero-one loss (red).	16
2.6	Multiclass SVM: (+, - and o) represent separate class and the task is to find decision boundaries that classify test point into 3-classes.	18
2.7	Structural SVM: An example of sequential labeling.	22
2.8	Scores for all possible combinations of a given sequence. There are few assumptions made for the sake of simplicity.	24
2.9	Graphical model with latent variables (Piccardi (2013))	26
2.10	a) Action class: Playing a guitar (Yao et al. (2011a)) b) Action class: Riding a bike (Yao et al. (2011a)).	29
3.1	a) The proposed action recognition approach: bottom layer: superpixel segmentation and feature extraction; intermediate layer: superpixel classification; top variable: action class. b) The graphical model: x : superpixel measurements; h : superpixel classes, or states; y : action class. c) Factor graph representation: φ, ϕ and θ are the feature functions in (3.5).	34
3.2	Example of superpixel segmentation: a) original image; b) superpixel segmentation; c) superpixel boundaries highlighted.	39

3.3	Examples of the top 20% superpixels contributing to the action score for classes <i>applauding</i> , <i>brushing teeth</i> , <i>gardening</i> and <i>waving hands</i> of Stanford 40 Actions. The figure shows triplets of {original image, superpixel decomposition, top-20 pixels highlighted} as a continuous sequence. This figure should be viewed in color.	48
3.4	(continued)	49
3.5	Average precision achieved by the proposed method in each class of Stanford 40 Actions.	51
4.1	Feature map h . The two binary variables for node t are noted jointly as y_t ; the four binary variables for edge e are noted jointly with a double index as $y_{s,d=e}$	73
5.1	An example of a sequence with 5 output (shaded) nodes.	76

List of Tables

3.1	Main notations (notations valid for this chapter only).	38
3.2	Time in seconds for the various greedy inference algorithms (loop at lines 7-9 in Algorithm 1).	45
3.3	Comparison of mean average precision on Stanford-40.	47
4.1	Comparison of clustering accuracy over the synthetic and Gesture Phase Segmentation datasets. Accuracy is reported as F_1 score (\pm standard deviation) over 10 runs of each technique.	70
5.1	Comparison of clustering accuracy over the Synthetic and Gesture Phase Segmentation datasets. Accuracy is reported as F_1 score (\pm standard deviation).	85
C.1	Summary of notations.	95

Dedicated to my Parents