Learning Structured Prediction Models in Computer Vision

by

Fayao Liu

A thesis submitted in fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Engineering, Computer and Mathematical Sciences
School of Computer Science

November 2015

Declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I also give permission for the digital version of my thesis to be made available on the web, via the Universitys digital research repository, the Library catalogue and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: ________________________

Date: ________________________

Acknowledgements

First and foremost, I would like to express my sincere gratitudes to my principle supervisor, Prof. Chunhua Shen. I would have never been able to finish this thesis without his guidance. During the course of my PhD study, he has always been an encouraging, inspiring and patient mentor, from whom I've learned not only methodologies, but the way to look at problems. His intelligence, creativity and keen perceptions in cutting-edge research topics have deeply impressed me. He has also set me a good example by his diligence and continuous efforts, as well as rigorous scientific attitudes. There are numerous other things I've learned from him, which will continue guiding me in my future career.

I would like to thank my co-supervisors, Prof. Anton van den Hengel and Prof. David Suter. They have showed generous patience and continuous support throughout my PhD candidature. As the director of the Australian Centre for Visual Technologies (ACVT), Anton has provided me a good platform as well as opportunities to communicate and collaborate with many talented researchers. He has also given me a lot of help on improving my language skills. I appreciate his generosity, encouragement and enlightenment.

I would like to thank Prof. Ian Reid, who showed me in person how to organize and write a research paper by telling a story. His innovative perspectives and insightful advices have helped me to improve my understandings and paper writing. It has always been pleasant and inspiring discussing with him.

I owe special thanks to Dr. Guosheng Lin, who taught me structured learning and many other things. He impressed me by his profound professional expertise and his extraordinary persistence. I'm deeply grateful for his unselfish sharing and constant supporting throughout my PhD life.

My sincere thanks go to ACVT researchers, especially Dr. Qinfeng (Javen) Shi, Dr. Peng Wang and Dr. Sakrapee (Paul) Paisitkriangkrai, for their kindness and valuable suggestions. Talking and discussing with them have always benefited me a lot.

Many thanks to all my current and previous lab mates, with whom I spent the most important years of my life together. Especially, I would like to mention Quoc-Huy Tran, Zhen Zhang, Yongrui Qin, Lina Yao, Lei Luo, Chao Zhang. I will always treasure those days spent with them. I also owe thanks to my friends I do not list here, with whom I share excitements and frustrations.

Finally, special gratitudes are attributed to my family, to whom this thesis is dedicated to.

Publications

This thesis is based on the content of the following peer-reviewed conference and journal papers:

1. Fayao Liu, Chunhua Shen, Guosheng Lin; *"Deep Convolutional Neural Fields for Depth Estimation from a Single Images"*; In proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
2. Fayao Liu, Chunhua Shen, Guosheng Lin, Ian D. Reid; *"Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields"*; Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2015.
3. Fayao Liu, Guosheng Lin, Chunhua Shen; *"CRF Learning with CNN Features for Image Segmentation"*; Pattern Recognition (PR), 2015.
4. Fayao Liu, Guosheng Lin, Chunhua Shen; *"Structured Learning of Tree Potentials in CRF for Image Segmentation"*; Submitted to IEEE Transactions on Neural Networks and Learning Systems (TNNLS); Major Revision.

In addition, I have published or submitted the following papers:

1. Fayao Liu, Luping Zhou, Chunhua Shen, Jianping Yin; *"Multiple Kernel Learning in the Primal for Multimodal Alzheimer's Disease Classification"*; In IEEE Journal of Biomedical and Health Informatics (JBHI), 2014.
2. Fayao Liu, Guosheng Lin, Chunhua Shen; *"Discriminative Training of Deep Fully-connected Continuous CRFs with Task-specific Loss"*; Submitted to IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
3. Fayao Liu, Ruizhi Qiao, Chunhua Shen, Lei Luo; *"From Kernel Machines to Ensemble Learning"*; Submitted to Pattern Recognition (PR).
4. Fayao Liu, Chunhua Shen, Ian Reid, Anton van den Hengel; *"Online Unsupervised Feature Learning for Visual Tracking"*; Submitted to Computer Vision and Image Understanding (CVIU).
5. Chunhua Shen, Junae Kim, Fayao Liu, Lei Wang, Anton van den Hengel; *"Efficient Dual Approach to Distance Metric Learning"*; In IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 2014.

THE UNIVERSITY OF ADELAIDE

Abstract

Faculty of Engineering, Computer and Mathematical Sciences
School of Computer Science

Doctor of Philosophy

by Fayao Liu

Most of the real world applications can be formulated as structured learning problems, in which the output domain can be arbitrary, *e.g.*, a sequence or a graph. By modelling the structures (constraints and correlations) of the output variables, structured learning provides a more general learning scheme than simple binary classification or regression models. This thesis is dedicated to learning such structured prediction models, *i.e.*, conditional random fields (CRFs) and their applications in computer vision. CRFs are popular probabilistic graphical models, which model the conditional distribution of the output variables given the observations. They play an essential role in the computer vision community and have found wide applications in various vision tasks—semantic labelling, object detection, pose estimation, to name a few. Specifically, we here focus on two challenging tasks in this thesis: image segmentation (also referred as semantic labelling) and depth estimation from single monocular images, which represent two types of CRFs models—discrete and continuous. In summary, we made three contributions in this thesis.

First, we present a new approach to exploit tree potentials in CRFs for the task of image segmentation. This method combines the advantages of both CRFs and decision trees. Different from traditional methods, in which the potential functions of CRFs are defined as a *linear* combination of some pre-defined parametric models, we formulate the unary and the pairwise potentials as nonparametric forests—ensembles of decision trees, and learn the ensemble parameters and the trees in a unified optimization problem within the large-margin framework. In this fashion, we easily achieve *nonlinear* learning of potential functions on both unary and pairwise terms in CRFs. Moreover, we learn class-wise decision trees for each object that appears in the image. We further show that this challenging optimization can be efficiently solved by combining a modified column generation and cutting-planes techniques. Experimental results on both binary and multi-class segmentation datasets demonstrate the power of the learned nonlinear nonparametric potentials.

Second, we propose to model the unary potentials of the CRFs using a convolutional neural network (CNN). The deep CNN is trained on the large-scale ImageNet dataset and transferred to image segmentation here for constructing unary potentials of superpixels. The CRFs parameters are then learned within the max-margin framework using structured support vector machines (SSVM). To fully exploit context information in inference, we construct spatially related co-occurrence pairwise potentials and incorporate them into the energy function. This prefers labellings of object pairs that frequently co-occur in a certain spatial layout and at the same time avoids implausible labellings during the inference. Extensive experiments on binary and multi-class segmentation benchmarks demonstrate the potentials of the proposed method.

Third, different from the previous two works, we address the problem of continuous CRFs learning, applied to the task of depth estimation from single images. Specifically, we formulate and learn the unary and pairwise potentials of a continuous CRFs model with CNN networks in a unified framework. We term this new method as deep convolutional neural fields, abbreviated as DCNF. It jointly explores the capacity of deep CNN and continuous CRFs. The proposed method can be used for depth estimation of general scenes with no geometric priors nor any extra information injected. Specifically, in our case, the integral of the partition function can be calculated in a closed form such that we can exactly solve the log-likelihood maximization. Moreover, solving the inference problem for predicting depths of a test image is highly efficient as closed-form solutions exist. We then further propose an equally effective model based on fully convolutional networks and a novel superpixel pooling method, which is ~ 10 times faster, to speedup the patch-wise convolutions in the deep model. With this more efficient model, we are able to design very deep networks to pursue further performance gain. Experiments on both indoor and outdoor scene datasets demonstrate that the proposed method significantly outperforms state-of-the-art depth estimation approaches. We also show experimentally that the proposed method generalizes well to depth estimations of images unrelated to the training data. This indicates the potential of our method for benefiting other vision tasks.

Dedicated to my family.

Contents

List of Figures

List of Tables

Notation

Symbol	Description
$\mathbf{1}$	Column vector with all elements being 1.
$\mathbf{0}$	Column vector with all elements being 0.
$\mathbf{I}$	Identity matrix.
$\mathbb{R}$	Domain of real numbers.
i.i.d.	Abbreviation of independent and identically distributed.
$< \cdot, \cdot >$	Inner product operation.
$\odot$	Stacking two vectors.
$\otimes$	Kronecker tensor.
$\mathrm{Tr}(\cdot)$	Trace of a matrix.
$\\|\cdot\\|_2$	L_2 norm.
Superscript $\top$	Transpose.
$\delta(\cdot)$	Indicator function which equals 1 if the input is true and 0 otherwise.
C	Trade-off parameter.
m	Number of examples.
$\boldsymbol{\xi}$	Vector of slack variables.
$\mathbf{w}$	Vector of model parameters.
$\mathbf{x}$	Input observation.
$\mathbf{y}$	Structured output label.
y	Scalar output label.
$\mathcal{X}$	Input domain.
$\mathcal{Y}$	Output domain.
$\mathcal{N}$	Set of nodes.
$\mathcal{S}$	Set of edges.
$\mathcal{W}$	Working set.
$\mathcal{H}$	Domain of weak learners/decision trees.
$g : \mathcal{X} \to \mathcal{Y}$	Structured prediction function.
$f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$	Scoring function.
$l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$	General loss function.
$\mathit{\Delta} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$	Structured loss function.

E	Energy function.
U	Unary potential function.
V	Pairwise potential function.
Ψ	Feature mapping function.
$\Psi^{(1)}$	Unary feature mapping function.
$\Psi^{(2)}$	Pairwise feature mapping function.
Pr	Probability function.
Z	Partition function.
sgn	Sign function.
$\hbar$	A weak learner.
$\hbar^{(1)}$	A unary decision tree.
$\hbar^{(2)}$	A pairwise decision tree.
$\mathbf{H}^{(1)}$	A group of unary decision trees.
$\mathbf{H}^{(2)}$	A group of pairwise decision trees.