

UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

**Hierarchical Convolutional Neural Networks for
Vision-Based Feature Detection**

by

Qiuchen Zhu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2022

Certificate of Authorship/Originality

I, Qiuchen Zhu declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering/Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
Signature: prior to publication.

Date: 04/05/2022

ABSTRACT

Hierarchical Convolutional Neural Networks for Vision-Based Feature Detection

by

Qiuchen Zhu

This thesis is devoted to the problem of feature detection, an essential prerequisite to machine vision applications. The key to feature detection rests with the development of effective algorithms, which could incorporate machine intelligence to achieve such attributes as accuracy in pixel-wise terms and robustness against structural and stochastic uncertainty. To this end, a hierarchical convolutional neural network (HCNN) with feature preservation, is proposed to present the probability map of feature candidates. Specifically, the abstraction of features in consideration is enhanced by bidirectional branch nets. The outputs of previous convolutional blocks are unified and concatenated to the current ones to reduce the visual impairment in the up/down-sampling stage and the overall information loss. Besides, an Intercontrast-based Iterative Thresholding (IIT) approach is developed for the proposed network hierarchy at the post-processing step, whereby patterns of interest are clustered within the probability map of identified features and generate a solid feature map. To effectively overcome uncertainty, network prediction is conducted by a customised variational inference. Here, deterministic weights are converted into a probability distribution with learnable hyperparameters to adapt the interference of outliers and alleviate overfitting. Furthermore, to incorporate Bayesian modelling into high-level tasks such as resource allocation, an additional module for Gaussian heatmap is developed to meticulously present the location of the geometrical target. Then, a physics-driven training scenario is designed to gradually shrink the benchmark kernel for continuous calibration to avoid local minima. In summary, the contributions of this thesis include 1. a new hierarchical network proposed for

feature detection, whereby the abstractions of the image can be bidirectionally extracted to improve prediction performance, 2. unsupervised and gradient-sharing approaches incorporating Bayesian inference in the proposed network for enhancing its uncertainty handling capability, 3. a new training strategy for representation learning via spatial indexing to link the primary geometrical features with quantitative allocation, and 4. an average F-measure proposed for evaluation of robustness along with other metrics for performance evaluation. An extensive comparison with existing techniques is conducted using various datasets and evaluation criteria for evaluation. Experimental results demonstrate the crack detection merits of the proposed architecture over existing techniques applied to numerous images. To illustrate generality of the developed network architecture, additional tests are also conducted for various applications, including salient object detection, anthropometric and facial landmark detection, and measurement retrieval. The results obtained show the scalability and robustness of the proposed model to medium and high-level image processing tasks.

Dissertation directed by Professor Quang Ha

School of Electrical and Data Engineering, University of Technology Sydney

Dedication

To my parents Jialin Zhu and Bin Jia, and my supervisor Professor Quang Ha.

Acknowledgements

First and foremost, I would like to thank my supervisor Quang Ha for his patient guidance and encouragement for this work. I respectfully acknowledge Dinh Tran Hiep, Manh Duong Phung and Ricardo P. Aguilera for the technical support, and thank my fellow group-mates for the stimulating discussions and nice team work. I would also like to pay my respect to Ngaiming Kwok for his invaluable suggestions on this thesis. I would like to acknowledge University of Technology Sydney, Chinese Scholarship Council, SmartFit and ApparelTech for the financial assistance to my research.

Finally, My special thanks to my family for the spiritual support throughout my PhD candidature especially during the hard time of the global epidemic.

Qiuchen Zhu
Sydney, Australia, 2022.

List of Publications

Journal Papers

- J-1. **Q. Zhu**, T. H. Dinh, M. D. Phung, and Q. Ha, “Hierarchical Convolutional Neural Network with Feature Preservation and Autotuned Thresholding for Crack Detection,” *IEEE Access*, vol. 9, pp. 60201-60214, 2021, DOI: 10.1109/ACCESS.2021.3073921.
- J-2. **Q. Zhu** and Q. Ha, “A Bidirectional Self-Rectifying Network with Bayesian Modelling for Vision-Based Crack Detection,” *IEEE Transactions on Industrial Informatics*, accepted on 30 APR 2022.
- J-3. T. H. Dinh, **Q. Zhu**, M. D. Phung and Q. Ha, “Summit Navigator Automatic Thresholding for Image Binarization with Application to Crack Detection,” *IEEE Transactions on Systems, Man and Cybernetics: Systems*, revised and resubmitted.

Conference Papers

- C-1. **Q. Zhu**, and Q. Ha, “A Bidirectional Self-Rectifying Network with Bayesian Modelling for Feature Detection and Keypoint Detection,” *The 21th International Conference on Machine Learning and Cybernetics*, Adelaide, SA, Australia, December 4-5, 2021, Doi: 10.1109/ICMLC54886.2021.9737243.
- C-2. T. X. Tran, T. H. Dinh, H. V. Le, **Q. Zhu**, and Q. Ha, “Defect Detection Based on Singular Value Decomposition and Histogram Thresholding,” in *Proc. 2020 IEEE/ASME International Conf. on Advanced Intelligent Mechatronics (AIM)*, Boston, MA, USA, July 6-9, 2020, pp. 149-1154.
- C-3. **Q. Zhu**, M. D. Phung, and Q. Ha, “Crack Detection Using Enhanced Hierarchical Convolutional Neural Networks,” *Proc. Australasian. Conf. on*

Robotics and Automation, Adelaide, SA, Australia, December 11-13, 2019, pp. 1-8. **Best Paper Award Winner.**

- C-4. V. T. Hoang, M. D. Phung, T. H. Dinh, **Q. Zhu** and Q. Ha, “Reconfigurable Multi-UAV Formation using Angle-Encoded PSO,” in *Proc. 2019 IEEE 15th International Conf. on Automation Science and Engineering (CASE)*, Vancouver, Canada, August 22, 2019, pp. 1670-1675.
- C-5. **Q. Zhu**, T. H. Dinh, V. T. Hoang, M. D. Phung, and Q. Ha, “Crack Detection Using Enhanced Thresholding on UAV based Collected Images,” in *Proc. Australasian. Conf. on Robotics and Automation*, Lincoln, Canterbury, New Zealand, Dec. 9-11, 2018, pp. 1-7.

Contents

Certificate	ii
Abstract	iv
Dedication	vi
Acknowledgments	vii
List of Publications	viii
List of Figures	xv
Abbreviation	xix
Notation	xx
1 Introduction	1
1.1 Thesis motivation	1
1.2 Research gaps	3
1.3 Research objectives	4
1.4 Thesis organisation	5
2 Literature Survey	7
2.1 Introduction	7
2.2 Basic deep learning frameworks for feature detection	7
2.2.1 Fully-convolutional models	9
2.2.2 Encoder-decoder models	10
2.2.3 Regional CNN models	11

2.3	Feature enhancements	12
2.3.1	Internal enhancements	13
2.3.2	External enhancements	17
2.4	Uncertainty in deep learning and probabilistic models	18
2.4.1	Impact of uncertainty	18
2.4.2	Types of uncertainty	20
2.4.3	Probabilistic models	22
2.5	Conclusion	25
3	Hierarchical neural networks with feature preservation	27
3.1	Introduction	27
3.2	Hierarchical neural networks	28
3.2.1	Basic architecture of networks	28
3.2.2	Logistic regression and probabilistic maps	30
3.2.3	Hierarchical calibration	32
3.2.4	Deterministic update of parameters	33
3.3	Feature preservation	34
3.3.1	Linear hypothesis and feature preservation	34
3.3.2	Feature preservation modules	35
3.4	Bidirectional self-rectifying abstractions	36
3.4.1	Architecture of bidirectional self-rectifying networks	36
3.4.2	Effective receipt fields	38
3.4.3	Dilated convolution blocks	41
3.5	Post-processing	42
3.5.1	Binarisation of probabilistic maps	42

3.5.2	One-off thresholding	42
3.5.3	Adaptive thresholding	44
3.5.4	Criteria for iteration control	45
3.6	Conclusion	47
4	Uncertainty handling in hierarchical neural networks	48
4.1	Introduction	48
4.2	Unsupervised approaches	49
4.2.1	Statistic activation functions	49
4.2.2	Stochastic dropout	49
4.3	Semi-supervised approaches	50
4.3.1	Stochastically distributed parameters	51
4.3.2	Bayesian inference in hierarchical networks	51
4.3.3	Optimisation target	52
4.3.4	Gaussianised feature maps	54
4.3.5	Spatial soft-argmax	58
4.3.6	Dissipation training and annealing effect	59
4.4	Conclusion	63
5	Surface crack detection using hierarchical neural networks	64
5.1	Introduction	64
5.2	Benchmarking	65
5.2.1	Datasets	65
5.2.2	Algorithms in comparison	66
5.2.3	Evaluation metrics	68

5.2.4	Implementation setup	71
5.3	Comparison for hierarchical neural networks	71
5.3.1	Results of hierarchical networks with feature preservation	71
5.3.2	Results of hierarchical networks with bidirectional abstraction	79
5.4	Comparison for frameworks with uncertainty modeling	82
5.5	Comparison of post-processing	84
5.6	Ablative study on network architecture	91
5.6.1	Ablation on feature preservation modules	91
5.6.2	Ablation on bidirectional abstraction	93
5.6.3	Ablation for the number of DCBs	96
5.6.4	Ablation for frequentist loss functions	97
5.6.5	Ablation on Bayesian inference	97
5.7	Conclusion	97
6	Abstract object detection and high-level Applications	99
6.1	Introduction	99
6.2	Medium-level and high-level tasks in image processing	99
6.3	Benchmarking	100
6.3.1	Datasets	100
6.3.2	Algorithms in comparison	102
6.3.3	Additional evaluation metrics	102
6.4	Experiment results	104
6.4.1	Results on medium-level tasks	104
6.4.2	Results on high-level tasks	106
6.5	Further extension to measurement retrieval	112

6.5.1	Anthropometric measurement retrieval	112
6.5.2	Solution and results	113
6.6	Conclusion	114
7	Conclusion	116
7.1	Thesis Summary	116
7.2	Thesis contribution	117
7.3	Thesis conclusion	119
7.4	Future work	119
	Bibliography	121

List of Figures

1.1	Diversity of homogeneous features	2
2.1	Architecture of fully-convolutional networks. Source: [63]	10
2.2	Architecture of basic auto-encoders	11
2.3	Architecture of SegNet. Source: [1]	11
2.4	Architecture of RCNN. Source: [33]	12
2.5	Skip-connection modules in ResNet	14
2.6	Staggered skip-connections in DenseNet. Source: [43]	15
2.7	Skip-connections in SRDenseNet	16
2.8	Alternative to the standard convolution	16
2.9	An adversarial sample of deep learning models. Source: [35]	19
2.10	Fake cracks to mislead DCNN	20
2.11	Homoscedastic uncertainty from blurry and colour fading. Source:[105]	21
2.12	Random shadows on the image. Source:[95]	22
2.13	Architecture of CRF-DCNN. Source: [55]	24
3.1	Architecture of HCNNFP	29
3.2	The comparison of loss calculation	33
3.3	Architecture comparison	36
3.4	Architecture of BSN	37

3.5	The marginal area of receptive fields	40
3.6	Comparison between standard and dilated convolutions	41
3.7	Difference in the effective receipt field of convolutions. Source: [64]	42
3.8	Binarization flowchart using adaptive thresholding	45
4.1	The representation of the weight W	51
4.2	Architecture of BSNBM with fully-connected layers	56
4.3	Architecture of BSNBM with heatmap indexing	57
4.4	Heatmaps from left to right: $\sigma_g = 10$ and $\sigma_g = 2$	59
4.5	From the ground truth heatmap (left) to the prediction(right): the regular training starts with a small Gaussian kernel fails to retrieve the distribution	59
4.6	From the ground truth heatmap (left) to the prediction (right): the predicted heatmap follows the shrinkage of the ground truth and represents increasingly accurate results	60
4.7	Annealing learning	61
4.8	Facial landmark detection using (a)regular training, (b)dissipation training and annealing learning; Red: predicted landmarks; Green: annotated landmarks	62
5.1	Bridge inspection	66
5.2	Comparison of frequentist's models. (Part I)	72
5.3	Comparison of frequentist's models. (Part II) From left to right on first row: original image, ground truth, results of HED, RCF, and SegNet; On second row: results of DeepCrack, FPHBN, U-Net, FCN, and HCNNFP.	73

5.4	Quantitative results of the first five samples on crack images: (a) $F_\beta \beta^2 = 0.25$, (b) $F_\beta \beta^2 = 0.3$, (c) AF_β , (d) $MAPE$	77
5.5	Quantitative results of the last five samples on crack images: (a) $F_\beta \beta^2 = 0.25$, (b) $F_\beta \beta^2 = 0.3$, (c) AF_β , (d) $MAPE$	78
5.6	Distribution of F_β with respect to β^2	80
5.7	Visual comparison of HCNFP and BSN. From left to right: original image, ground truth, results of HCNFP and BSN.	81
5.8	Detection results on <i>Crack500</i> against granular surfaces. From left to right on first row: original image, ground truth, results respectively of DeepCrack, FPHBN, CracknetV, and U-Net; Secound row: results respectively of FCN, PGA-Net, HDCB-Net, HAC-Net, HCNFP, and BSNBM.	85
5.9	Detection results on <i>SYDCrack</i> against regular texture. From left to right on first row: original image, ground truth, results respectively of DeepCrack, FPHBN, CracknetV, and U-Net; Secound row: results respectively of FCN, PGA-Net, HDCB-Net, HACNet, HCNFP, and BSNBM.	86
5.10	Detection results on <i>DCD</i> against confusing marks and irregular texture. From left to right on first row: original image, ground truth, results respectively of DeepCrack, FPHBN, CracknetV, and U-Net; Secound row: results respectively of FCN, PGA-Net, HDCB-Net, HACNet, HCNFP, and BSNBM.	87
5.11	Detection results on <i>GAPs</i> against poor light conditions. First column from left to right: original image, ground truth, results respectively of DeepCrack, FPHBN, CracknetV, and U-Net; Secound column from left to right: results respectively of FCN, PGA-Net, HDCB-Net, HACNet, HCNFP, and BSNBM.	88
5.12	Quantitative results of fixed thresholding and CBAT on crack images: (a) $F_\beta \beta^2 = 0.25$, (b) $F_\beta \beta^2 = 0.3$, (c) AF_β , (d) $MAPE$	90

5.13	Binarization results of the probability map.	94
6.1	Visualised results of the proposed network on MSRA10K datasets. . .	105
6.2	Visualised results of the proposed network on MSRA-B datasets. (I) .	107
6.3	Visualised results of the proposed network on MSRA-B datasets. (II)	108
6.4	Visualised results of landmark detection using the proposed model. Red: predicted landmarks; Green: annotated landmarks	110
6.5	Approximate silhouette of anthropometric circumferences	114

Abbreviation

- Adam: Adaptive moment estimation
- BSNBM: Bidirectional self-rectifying network with Bayesian modelling
- CBAT: Contrast-based autotuned thresholding
- CNN: Convolutional neural network
- CRF: Conditional random field
- DCNN: Deep convolutional neural networks
- DCB: Dilated convolutional block
- ELBO: Evidence lower bound
- FCN: Fully convolutional network
- F/REB: Forward/reverse enhancement branch
- HCNNFP: Hierarchical convolutional neural networks with feature preservation
- JI: Jaccard Index
- KL: Kullback-Leibler
- MAPE: Mean absolute error
- MAPE: Mean absolute percentage error
- MLP: Multilayer perceptron
- MSE: Mean square error
- RNN: Recurrent neural network
- RCNN: Regional convolutional neural network
- ReLU: Rectified linear units
- RF: Receptive field
- ROI: Region of interest

Nomenclature and Notation

Lower-case italic alphabets denote scalar values or vectors.

Upper-case italic letters denote matrices denotes matrices.

In a sequence, lower-case and capital italic letters respectively indicate the index and the maximum value.

Roman or calligraphic letters indicate a function.

\hat{x} represents the estimated/predicted value.

$*$ denotes the convolutional operation.

\odot is the component-wise multiplication.

$\|\cdot\|_F$ denotes the Frobenius norm.

w represents the scalar weight.

c_o denotes the coordinates of landmarks.

I is the identity diagonal matrix.

R represents the pixel region of interest.

Ω represents a sample space.

AF_β is the average F-measure.

F_β is the F-measure.

h_{smx} represents the spatial softmax function.

P denotes the probability function.

\mathbb{E} denotes the mathematical expectation.

\mathcal{F} is the Fourier transformation.

\mathcal{L} represents the loss function.

\mathcal{N} represents the Normal/Gaussian distribution.