RESEARCH PROJECT IN MECHATRONICS ENGINEERING

A CONVOLUTIONAL NEURAL NETWORK FOR DETECTING VISUAL TEXTURE

Conrad Scherb

Project Report ME110-2022

Co-worker: Callan Loomes

Supervisor: Dr Luke Hallum

Department of Mechanical and Mechatronics Engineering
The University of Auckland

15 April 2022

ME110-2022

A CONVOLUTIONAL NEURAL NETWORK FOR DETECTING VISUAL TEXTURE

Conrad Scherb

ABSTRACT

Abstract goes here.

DECLARATION

Student

I hereby declare that:

1. This report is the result of the final year project work carried out by my project partner (see cover page) and I under the guidance of our supervisor (see cover page) in the 2022 academic year at the Department of Mechanical and Mechatronics Engineering, Faculty of Engineering, University of Auckland.
2. This report is not the outcome of work done previously.
3. This report is not the outcome of work done in collaboration, except that with a potential project sponsor (if any) as stated in the text.
4. This report is not the same as any report, thesis, conference article or journal paper, or any other publication or unpublished work in any format.

In the case of a continuing project, please state clearly what has been developed during the project and what was available from previous year(s):

Signature:

Date:

Supervisor

I confirm that the project work undertaken by this student in the 2022 academic year is / is not (strikethrough as appropriate) part of a continuing project, components of which have been completed previously. Comments, if any:

Signature:

Date:

Table of Contents

Acknowledgements

Thank important people here.

Glossary of Terms

Term	Definition

Abbreviations

V1	Visual cortex layer 1
2D	Two dimensional
CNN	Convolutional neural network
FRF	Filter-rectify-filter [model]

1. Scope and Objectives

Visual percpetion is an extremely complex task, requiring the integration of various parts of an image, such as edges, contours and textures let alone more complex features such as faces [1]. Computer vision, the use of computers to analyse and process images often takes inspiration from actual physiological processes in the brain. One issue with this is a significant amount of the visual processing pipeline is not well understood - effectively parts are a "black box" where we know what the input and output are but not the internal workings. As such, developing models that replicate the input-output behavior can help us both further understand how physiological processes might work as well as give clues on how to improve computer vision models.

One key technology well suited to modelling how the brain might work are neural networks - after all, they are based on the underlying biology. They are powerful machine learning tools that model artificial "neurons" and there connections. "Learning" is achieved by changing the mathematical parameters that control the connections between neurons, such as weights (whether a connection is inhibitory or excitatory and by how much so) and biases (what threshold is required for an activation). Passing in a large set of known outputs for a given inputs causes slight adjustments to these parameters over time, and eventually the network algorithm will be trained such that it can achieve a high degree of classification accuracy for a given unknown input. For certain applications such as audio or visual processing, more advanced neural network types can be used which have further steps that can improve their accuracy. CNNs in particular are widely used for image classification, and feature an additional convolution step that effectively allows them to segment images into smaller components, and are the most useful type for modelling brain activity.

We are particularly interested in how texture detection works in the brain, as detecting texture is something that is done almost instantaneously in day-to-day life but can be difficult for computer vision systems, and the physiological basis is not well understood. We are therefore interested in developing a CNN that can be used to detect texture in an image and see what that might imply in terms of the brain itself, linking it back to the known underlying biology. We are also interested in simplifying our model as much as possible to avoid overfitting and in order to make sure that it models the simple biology. We wll also compare our generated network against any pre-existing models and investigate any similarities or differences. We will be defining a visual "texture" as a 2D array of Gabor patches, either containing a signal or not based on each column sharing an orientation. Noise will be added into this texture to ensure our model is noise-resistant in order to more closely model the brain's response.

2. Literature Review

This literature review outlines the background context of biological texture detection, currnet models for discriminating between visual textures, and the current state of research into CNNs and how they could be applied to model physiological neural systems.

2.1 Biological background of texture detection

The visual cortex is a highly organised system comprised of layers of neurons that each has a specific task in visual processing, with more complex tasks such as motion or face recognition performed in higher layers [1]. Different areas on V1 correspond to specific parts of the visual field, effectively projecting the retina onto V1 forming a "retinotopic map" [2]. V1 cells are very dense and have small receptive fields as they perform the first step of integration of visual information into the brain. They are highly selective for orientation and contrast, which are key for broad detection of texture, and can be classified as either simple or complex cells. Both types are orientation specific, but as complex cells respond to large receptive fields instead of an exact mapping and have additional responses to object movement [3], we will focus on simple cells in this review. The receptive field of orientation-specific V1 simple cells is longer in one direction than another, and as such if an input edge aligns within a certain threshold to the receptive field a response will be produced. Each grouping of simple cells that forms part of a visual receptive field is only activated by one texture - e.g. a simple cell may respond to a white line on a black background but not vice versa [1].

This sensitivity of the simple cells allows for the development of artificial visual stimulus that will maximally activate the simple cells. Gabor patches, produced via a convolution of a Gaussian function and a sinusoid, match the shape of the receptive field and have the contrast between on and off regions required for activation [4]. Polat found that the intensity of response was increased via "lateral masking" especially under lower contrast conditions, where Gabors sharing the same orientation of are placed along a single line [5]. This increased intensity corresponds to the detection of an edge as an edge is defined by a constant region of uniform orientation.

However, in real-world conditions, the majority of edges that humans perceive are not perfectly straight. For example, a tree trunk has regions of different orientations against a background, yet humans can easily distinguish the overall orientation of the tree. Noise can be added to a model of texture to model for this, such that a signal with certain orientation has deviating areas of noise. Hussain & Bennet used this approach - they added Gaussian noise to texture patterns that allowed them to investigate improvements in subjects abilities to detect whether at texture or not exists in a given stimulus image [6]. This detection of textures is almost spontaneous, and can occur even without the subject's attention [7]. The physiological basis for this ability to integrate output from the simple cells into detecting overall orientation whilst filtering out noise is not well understood however, and is of interest for further research.

From this review of the biological components of texture detection, we observe key factors to be reproduced in models: the use of units with small receptive fields to detect orientation, which allow for fast segregation of images into groups of similar textures with significant noise filtering. However, the connection on how orientation-detecting cells work to create segmentation is not well understood and could give clues into how a artificial network to detect texture could be formed [8]. In terms of neural network architecture, this could suggest that we want to use an input layer with a large number of neurons to model simple cells, and that we may want to first develop a segmentation model to find edges before considering the overall texture within an image. Hence, a CNN would be best suited to this task as the input layer of a CNN is comprised of neurons with slightly overlapping receptive fields [9].

2.2 Non-machine learning based approaches

A significant amount of work has gone into developing models utilising the Fourier transform as well as 2D filters in order to distinguish between textures in idealised test samples. Previous work into these models will be outlined here, as well as what clues they provide into the underlying biology and what similarities they might share with a neural network based approach.

2.2.1 Fourier analysis

Spatial frequency is a key component of biological texture detection, as certain populations of visual cortex cells specifically respond to certain spatial frequencies [10]. Mayhew & Frisby suggested that V1 cells could effectively perform a 2D Fourier transform on an image and then use this output for further processing [10]. They found that subjects were able to distinguish between differing spatial frequencies much better than different orientations at the same spatial frequencies. This could not be fully explained by a Fourier based model given that all sample textures were composed of pure sinusoids and should have been easily decomposable into discriminable sections. Nevertheless, Fourier analysis could give some insight into the underlying spatial frequency discrimination observed.

As such, there has been significant research into the use of Fourier analysis to create models capable of discriminating between visual textures. Harvey & Gervais pioneered this work, where they developed a four channel model using 2D Fourier transforms, with each sensitive to a certain spatial frequency or texture [11]. Textures would be discriminated against by combining the output of each channels together. They compared output from the 4-channel Fourier-based system to results from human subjects in a task where textures were to be put into groups of similar textures. They found a strong relationship between perceived similarity of the textures from humans to that of the model. Other early approaches included the use of exponential or Butterworth filters [12], but these did not respond as well as the Fourier filter [11].

Hence, we can deduce that segmentation of spatial frequencies could be important in texture detection and that using spatial frequency specific channels and then recombining them later could be a useful approach. This is an analogue to using hidden layers in a CNN which integrate from the input neurons (i.e. receptive fields) and process out specific spatial frequencies in that layer.

2.2.2 Filter-rectify-filter models

Another approach used to simulate visual texture detection is FRF models or the back-pocket model as described by Landy [13]. In this model, the first filter is selective for a certain orientation and spatial frequency, similarly to simple cells in the brain. This filter responds maximally positive when the alignment and spatial frequency matches completely and minimally negative when they do not match at all. This is then rectified in order to make all filter results positive, which can be achieved by squaring or another nonlinearity such as full-wave rectification [14]. Finally, a second filter is applied to the rectifier output, which has a much larger receptive field and spatial frequency and provides the final textural output. In order to discriminate between which texture might be present within an image, multiple channels are used each with a linear filter tuned to different orientation or spatial frequency. Results from each channel are then combined similarly to the Fourier-based approach.

This filter-rectify-filter model is thought to have some parallels with actual physiology as cells exhibiting similar input-output behavior to the second filter in the FRF model have been isolated in the monkey visual cortex [15]. Kingdom et al. sought to investigate the validity of FRF models by using them on texture gratings either contrast modulated, orientation modulated or spatial-frequency modulated as a model for the types of textures that could be found in an image [14]. Using a masking approach where types of modulation were used together and the response measured, they found that the addition of another type of modulation in the texture did not change the model output, with the exception of contrast modulated masks. This implies that for spatial frequency and orientation modulated textures, that independent FRF models are involved for the detection of these elements within a texture. Thus, we learn that our developed CNN might have separate pathways for detecting orientation modulated and spatial-frequency modulated textures.

2.3 Machine learning based approaches

While non-machine learning based models can have similar input-output characteristics to what we expect in physiology, a neural network would still be the best approach to gain a better understanding of how neurons are connected. CNNs in particular model this behavior well as the input layer is comprised of receptive fields, with small regions of overlap - effectively, a retinotopic map. The very design of CNNs is based upon our knowledge of underlying biology [16], hence they are the perfect tool for this application.

2.3.1 CNN layers

Broadly, CNN hidden layers will be one of three types. Convolutional layers apply a convolution to their input, which is limited to a certain receptive field. This reduces the computational complexity as each receptive field region has a smaller number of weights. Convolutional layers generate feature maps, which represent actual features in an image such as edges or contours [16]. Pooling layers act to reduce the resolution of the output of previous layers to reduce noise while keeping key information from feature maps. Small kernels are used as part of pooling which condense for example 2x2 region of neurons into a single output through taking the average or maximum of values in that kernel [17]. Finally, in fully connected layers each neuron is connected to every other neuron. While computationally expensive, fully connected layers allow the condensation of information from each neuron into a single vector, allowing for overall classification of an input image. Hence, fully connected layers generally are the last layers in a CNN architecture.

2.3.2 Machine learning algorithms

Another key component of CNNs is the machine learning component - that is, how the model is trained. The most common machine learning algorithm is the back-propagation algorithm [18]. This algorithm is based upon the idea that the error in the output of a neuron is the difference between the expected output and the actual output. This error is then propagated back through the network, which is then used to update the weights of the neurons. This process is repeated until the error is sufficiently small [18]. Thanks to libraries such as TensorFlow and Keras, the learning process can be automated on modern GPU hardware [19]. Given that we will be collecting or otherwise generating a test dataset with input images with texture presence or absence classified by humans, our task falls under the realm of supervised learning [20]. One issue with supervised learning approaches is bias in the training set, which we will need to be careful to avoid when creating it.

2.3.3 CNN architectures

Significant work has gone into refining and developing different types of CNNs. Broadly, they can either be classified as shallow or deep neural networks, which depends on the amount of hidden layers in the network. Deep CNNs are more computationally expensive due to the number of layers, suggests that they can fit complex functions better than shallow networks [21]. Deep neural networks are specifically well suited to object recognition tasks, such as object classification, such as the AlexNet architecture which was used to win the ImageNet Large Scale Vision Recognition challenge in 2012 [22]. Zhuang et al. suggested that deep neural networks can be used to model the entire ventral stream of the brain, responsible for object recognition [23]. However, given that we are only interested in texture detection in an idealised image of Gabors, a model that could fit the entire ventral visual processing pathway would not be appropriate. While segmentation of an image is also important in texture detection, and networks exist specialising in biomedical texture segmentation such as U-Net, we believe that a simpler approach using a basic shallow CNN will allow us to more accurately model underlying biology.

2.3.4 Previous use of CNNs for modelling visual systems

By design, CNNs provide a good model for the visual system, with architectures such as AlexNet providing extremely high accuracy at classifying images into one of thousands of categories. The major issue when using CNNs for this approach was the computations and size of the training set required - over a million images were used in the training of AlexNet in it's debut in the ImageNet classification challenge [9]. However, texture detection is a much simpler task than classifying an entire image - the response to whether a texture or not exists in a given dataset of Gabors is binary, and the number of images used in the training set will be comparatively small. The majority of the literature on using CNNs to model the visual systems has been focused on the whole cortex level rather than focusing on a specific element such as texture detection. Tripp used a deep system to investigate the overall architecture from a macro level, but this does not give much information as to how texture specifically is detected and modelled [24], whereas Kociolek et al did use CNNs for modelling texture directionality but did not consider what their generated network might suggest about V1 architecture [25]. As such, using a simple CNN architecture to model the texture-discriminating ability alone is a novel research interest.

References

[1] D. Purves, "Central visual pathways," in *Neuroscience*, 2019, ch. 12.

[2] G. Perry, P. Adjamian, N. J. Thai, I. E. Holliday, A. Hillebrand, and G. R. Barnes,

"Retinotopic mapping of the primary visual cortex - a challenge for MEG imaging of the human cortex," *Eur J Neurosci*, vol. 34, no. 4, pp. 652–661, Aug 2011.

[3] Y. Lian, A. Almasi, D. B. Grayden, T. Kameneva, A. N. Burkitt, and H. Meffin, "Learning receptive field properties of complex cells in V1," *PLOS Computational Biology*, vol. 17, no. 3, pp. 1–27, 03 2021.

[4] D. Durrie and P. S. McMinn, "Computer-based primary visual cortex training for treatment of low myopia and early presbyopia," *Trans Am Ophthalmol Soc*, vol. 105, pp. 132–138, 2007.

[5] U. Polat, "Functional architecture of long-range perceptual interactions," *Spat Vis*, vol. 12, no. 2, pp. 143–162, 1999.

[6] Z. Hussain and P. J. Bennett, "Perceptual learning of detection of textures in noise," *Journal of Vision*, vol. 20, p. 22, 2020.

[7] S. P. Heinrich, M. Andrï¿œs, and M. Bach, "Attention and visual texture segregation," *Jouranl of Vision*, vol. 7, p. 6, 2007.

[8] J. Bergen and M. Landy, "Computational modeling of visual texture segregation," in *Computational Models of Visual Processing*, 1991.

[9] G. W. Lindsay, "Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future," *J Cogn Neurosci*, vol. 33, no. 10, pp. 2017–2031, 09 2021.

[10] J. E. W. Mayhew and J. P. Frisby, "Texture discrimination and fourier analysis in human vision," *Nature*, vol. 275, no. 5679, pp. 438–439, Oct. 1978.

[11] L. O. Harvey and M. J. Gervais, "Visual texture perception and fourier analysis," *Perception and Psychophysics*, vol. 24, pp. 534–542, 1978.

[12] H. Mostafavi and D. Sakrison, "Structure and properties of a single channel in the human visual system," *Vision Research*, vol. 16, no. 9, pp. 957–IN4, 1976.

[13] M. S. Landy, "Texture analysis and perception," in *The New Visual Sciences*, 2013, pp. 639–652.

[14] K. F. A. A., P. Nicolaas, and H. Anthony, "Mechanism independence for texture-modulation detection is consistent with a filter-rectify-filter mechanism," 2003.

[15] E. Peterhans and R. von der Heydt, "Subjective contours–bridging the gap between psychophysics and physiology," *Trends Neurosci*, vol. 14, no. 3, pp. 112–119, Mar 1991.

[16] G. Franchini, , V. Ruggiero, F. Porta, and L. Z. and, "Neural architecture search via standard machine learning methodologies," *Mathematics in Engineering*, vol. 5, no. 1, pp. 1–21, 2022.

[17] R. Nirthika, S. Manivannan, A. Ramanan, and R. Wang, "Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study," *Neural Comput. Appl.*, vol. 34, no. 7, pp. 5321–5347, 2022.

[18] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," vol. 86, pp. 2278–2324, 1998.

[19] M. A. et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org.

[20] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 1310–1315.

[21] S. Chauhan, L. Vig, M. De Filippo De Grazia, M. Corbetta, S. Ahmad, and M. Zorzi, "A comparison of shallow and deep learning methods for predicting cognitive performance of stroke patients from mri lesion images," *Frontiers in Neuroinformatics*, vol. 13, 2019.

[22] "Large scale visual recognition challenge 2012 (ilsvrc2012)."

[23] C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, and D. L. K. Yamins, "Unsupervised neural network models of the ventral visual stream," *Proc. Natl. Acad. Sci. USA*, vol. 118, no. 3, p. e2014196118, 2021.

[24] B. Tripp, "Approximating the architecture of visual cortex in a convolutional network," *Neural Comput*, vol. 31, no. 8, pp. 1551–1591, Jul. 2019.

[25] M. Kociolek, M. Kozlowski, and A. Cardone, "A convolutional neural networks-based approach for texture directionality detection," *Sensors*, vol. 22, no. 562, p. 562, 01 2022.

Appendix A The First Appendix

Program A1 Some MATLAB script

```
% SaveExperiment.m: This file prompts the user to save the data
% and plots the results.
%
% This file is meant to be run autoamtically after lab experiment is
% finished.
%
% Hazim Namik                             Date created: 14/4/2019

clc;

% prompting the user to specify a file name and a location
[fileName,filePath] = uiputfile('*.mat','Save file name');
% Checking if the user clicked cancel
if(~(ischar(fileName)&& ischar(filePath)))
    disp('Canceled. No data was saved.');
    return
end
% Saving the file at the specified location
save ([filePath,fileName],'ActualPumpUsage', 'ActualError', '
   TankHeightAll', 'SimPumpUsage', 'SimError');
```

Appendix B Second Appendix