**Mathematisch-Naturwissenschaftliche Fakultät**

**Lernbasierte Computer Vision**

Bachelorarbeit

# Fancy Title

Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik
Autonomous Computer Vision
Marius Hobbhahn, marius.hobbhahn@student.uni-tuebingen.de, 2018/19

| | |
|---|---|
| Bearbeitungszeitraum: | von-bis |
| Betreuer/Gutachter: | Prof. Dr. Andreas Geiger, Universität Tübingen |
| Zweitgutachter: | Dr. Benjamin Coors, Universität Tübingen |

# Selbstständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Bachelorarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Bachelorarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

---

Marius Hobbhahn (Matrikelnummer 4003731), December 18, 2018

# Abstract

TODO

# Zusammenfassung

TODO

# Contents

# 1 Introduction

Object detection describes the task of detecting instances of semantic objects in visual data, i.e. images and videos in two or three dimensions. Even though the task is very easy for humans in most situations, it is very hard for computers. However, in recent years object detection algorithms have gotten significantly better for many different applications like face recognition, object tracking (e.g. the ball in a football match) and especially semantic segmentation of traffic scenes, pedestrian and car tracking.

For most state of the art (SOTA) object detection algorithm convolutional neural networks (CNNs) are used. In many implementations the filters used in those CNNs are all trained during the training period. [BM12] introduced a new technique called the Scattering Transform that uses wavelet operations on the image and performs classification tasks on those. They also show that the technique is essentially equivalent to using CNNs with fixed weights for some or all filters. It has been applied successfully in a variety of tasks. [SM13] showed that the scattering transform is applicable to texture discrimination. [OM14] have demonstrated that the scattering transform also produces results similar to other SOTA algorithms for unsupervised learning. [ACC+17] improved the classification of diseases from neuroimages considerably. Lastly, [OBZ17] shows that substituting the first layer filters of CNN approaches with the scattering transform yields equivalent results compared to these filters being trained.
The reason why the scattering transform has proven so successful are the properties it provides. It is invariant to rotation, translation and scaling. These properties are important for image classification but also necessary for object detection. For example, when detecting pedestrians in real traffic situations, the object detection algorithm must be able to identify them independent of their location, size or rotation within the image.
This work tries to harvest the useful properties of the scattering transform and combine it with already established state of the art object detection algorithms. This will be done primarily in two ways. First, the techniques are combined sequentially,

i.e. the SOTA algorithms are applied only on the outputs of the scattering transform. [OBZ17] have already shown that sequential combination is able to produce SOTA results for image recognition. This is the attempt to reproduce these findings for object detection.
Second, the techniques are combined in parallel, i.e. the information of the scattering transform are used as additional inputs for the object detection algorithm or merged at later stages. This has not been tested yet and is the primary extension of the just described related work.
If the approaches are successful that has three specific advantages. First, the scattering transform might yield information that were currently not available to the network and therefore increasing its accuracy. Even if that might only be a marginal increase, it is meaningful for application. Every little reduction of the error in object detection, especially for autonomous driving, means a reduction of risk of self driving cars. This is directly translated to lives being saved in the longterm. Second, fixed weights imply no additional training time for them. If, for example, one layer can be substituted that would reduce the length of training and save cost and energy while creating access for people who currently do not own multiple GPUs. Third, fixed weights cannot be overfit and are very maximally general. This might produce more robust algorithms and protect against black box attacks or other malicious practices applied to CNNs. This, however, will not be tested within the scope of this work but might be interesting follow-up.

# 2 Theory

## 2.1 2D Image Processing

Image processing describes the application of different algorithms on images with the purpose of gaining certain information about it or changing its representation. Most of the time images are given as two dimensional pixel arrays where each entry denotes the intensity of that pixel. In the case of grayscale images the value is between 0 and 255 representing black and white respectively. When handling color images an additional 3rd dimension is added with three channels representing a red, green, blue (rgb) encoding. Each entry, again, has values between 0 and 255 representing color intensity.
Instead of imagining an image as a flat 2D object, it can also be seen as a terrain with surface, where the height of each coordinate is determined by the intensity of its value. An example of this is shown in figure 2.1.

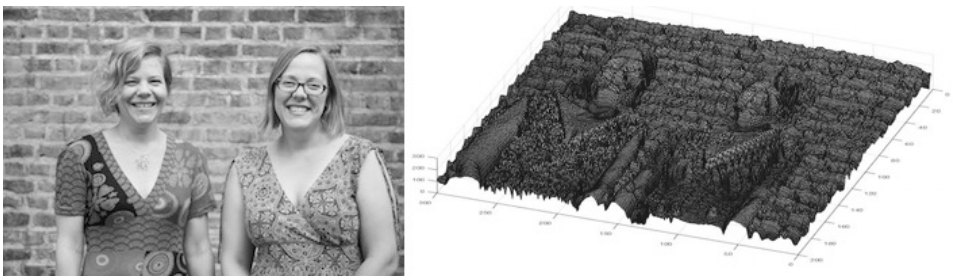

**Figure 2.1:** Left: image represented as 2D flat surface. Right: image as 3D terrain with uneven surface. [1]

Like every other surface, these images can now be approximated as the sum of many different two dimensional sine waves. 2D sine waves are defined as in equation 2.1, where f $a$ is the amplitude and $h$,$k$ are the frequencies in $x$ and $y$ direction respectively.

[1] Figure taken from `https://plus.maths.org/content/fourier-transforms-images`

$$f = a\sin(h \cdot x + k \cdot y) \tag{2.1}$$

To give an example of how this approximation looks like, figure 2.2 shows examples of three different two dimensional sine waves. It can be observes that higher amplitudes dominate the resulting wave, i.e. determine the direction of the wave stronger than the smaller amplitudes.

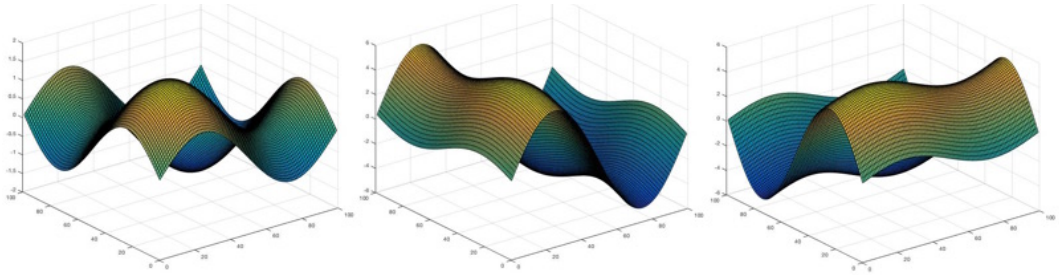

**Figure 2.2:** Left: $\sin(x)+\sin(y)$. Middle: $5\sin(x)+\sin(y)$. Right: $\sin(x)+5\sin(y)$. On the middle and right images the higher amplitudes of 5 dominate the resulting wave. [2]

## 2.2 Fourier Transformation

A Fourier Transform (FT) decomposes a signal into the frequencies that make it up.

### 2.2.1 One dimensional FT

In the case of one dimensional signals the decomposition are the coefficients of the sine waves representing the signal. A good example of this would be the decomposition of a The FT is defined by equation 2.2 for any real number $\omega$ and any integrable function $f : \mathbb{R} \to \mathbb{C}$.

$$\tilde{f}(\omega) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \omega} dx \tag{2.2}$$

To get back to the Fourier domain when the given a frequency, the inverse Fourier transform defined in equation 2.3 is used.

[2] Figure taken from `https://plus.maths.org/content/fourier-transforms-images`

$$f(x) = \int_{-\infty}^{\infty} \tilde{f}(\omega)\, e^{2\pi i x\omega} d\omega \tag{2.3}$$

When using discrete instead of continuous functions, the integrals in the definitions become sums. Then the definition of the forward FT is given in equation 2.4 and in equation 2.5 for the inverse FT.

$$\tilde{f}(\omega) = \sum_{x=1}^{n} f(x)\, e^{-2\pi i x\omega} \tag{2.4}$$

$$f(x) = \sum_{x=1}^{n} \tilde{f}(\omega)\, e^{2\pi i x\omega} \tag{2.5}$$

### 2.2.2 Two dimensional FT

Since images are two dimensional objects the Fourier transform needs to be extended. The Fourier transform then becomes a complex function of two or more real frequency variables $\omega_1, \omega_2$. Since images images are finite objects the discrete version of the two dimensional Fourier transform is given in equation 2.6 for the forward case and in equation 2.7 for the inverse case.

$$\tilde{f}(\omega_1, \omega_2) = \sum_{x=1}^{n} \sum_{y=1}^{m} f(x, y)\, e^{-2\pi i(\omega_1 \cdot x + \omega_2 \cdot y)} \tag{2.6}$$

$$f(x, y) = \sum_{x=1}^{n} \sum_{y=1}^{m} \tilde{f}(\omega_1, \omega_2)\, e^{2\pi i(\omega_1 \cdot x + \omega_2 \cdot y)} \tag{2.7}$$

## 2.3 Scattering Transform

A transformation from the Fourier to the frequency domain cannot only be performed by using the sine but in principal with any given periodic function. Wavelets are wave-like oscillation with an amplitude that begin and end at zero. In most use cases wavelets are specifically crafted to have certain properties. The Scattering Transform is based on a Morlet wavelet, which is defined in equation **??**.

$$\psi(u) = C_1(e^{iu.\xi} - C_2)e^{\frac{-|u|^2}{2\sigma^2}} \tag{2.8}$$

where $C_2$ is chosen such that $\int \psi(u)du = 0$. $u.\xi$ denotes the innerproduct of $u$ and $\xi$ and $|u|^2$ is the norm in $\mathbb{R}^2$. Figure 2.3 shows the 2 dimensional Morlet wavelet with parameters $\sigma = 0.85$ and $\xi = \frac{3\pi}{4}$. These parameters are taken from [BM12]. No additional fine tuning is done in this work.

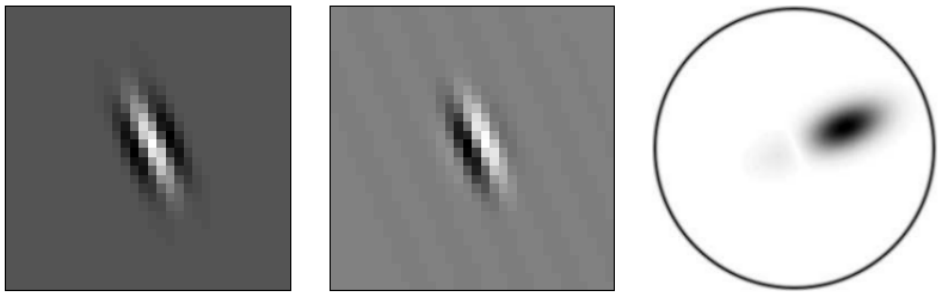


**Figure 2.3:** Complex morlet wavelet. a) Real part of $\psi$. b) Imaginary part of $\psi$. c) Fourier modulus $|\hat{\psi}|$. Image taken from [BM12].

### 2.3.1 Properties of the Scattering Transform

As already pointed out in the introduction, key properties of object detection are invariance with respect to scale, rotation, localization and deformation. The Scattering transform extends a simple 2D Fourier transform in exactly these properties. The Fourier transform is translation invariant, but unstable with respect to deformations at high frequencies. This implies that the representation is also unstable with respect to deformations. Additionally a Fourier transform loses to much information. Two different signals can have Fourier transforms with exactly the same moduli.

A wavelet, in comparison to the sinusoidal waves of the Fourier, is a localized waveform and therefore stable with respect to deformation.

## 2.4 Convolutional Neural Networks

For most image-related tasks, i.e. classification or object detection, a picture is used as a collection of pixels. However, not all pixels are equally important and subsets of the entire image form meaningfully connected subcollections. This might be a face in a photo of a family gathering. For humans the ability to detect these features and contextualize them comes naturally, for computers it does not. Therefore convolutional neural networks (CNNs) are used. Convolutions are essentially just the application of filters on an image. The filter is applied at every possible location in the image, as described in figure:

In CNNs there are multiple stages of filters in sequential order and multiple filters per layer. That means at every stage of the network different filters are applied on the outcome of an earlier step. The filters are assumed to learn different features of the images. The later the stage, the higher the level of complexity of the feature to be learned. That means, that an early filter might learn simple attributes such as edges or colors while a later filter might learn

# 3 Experiments

## 3.1 Setup

# 4 Results

TODO

# 5 Conclusion

TODO

# Bibliography

[ACC+17] Tameem Adel, Taco Cohen, Matthan Caan, Max Welling, On behalf of the AGEhIV study group Initiative, and the Alzheimer's Disease Neuroimaging. 3d scattering transforms for disease classification in neuroimaging. *NeuroImage: Clinical*, 14:506–517, 2017. Exported from https://app.dimensions.ai on 2018/10/21.

[BM12] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *CoRR*, abs/1203.1513, 2012.

[OBZ17] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. *CoRR*, abs/1703.08961, 2017.

[OM14] Edouard Oyallon and Stéphane Mallat. Deep roto-translation scattering for object classification. *CoRR*, abs/1412.8659, 2014.

[SM13] Laurent Sifre and Stephane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '13, pages 1233–1240, Washington, DC, USA, 2013. IEEE Computer Society.