# PerceptNet: A Human Visual System Inspired Neural Network for Estimating Perceptual Distance

Alexander Hepburn

Work done with Valero Laparra, Jesús Malo, Ryan McConville, Raul Santos-Rodriguez
Contact Email: alex.hepburn@bristol.ac.uk

University of BRISTOL

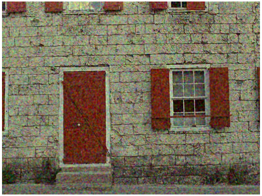
VNIVERSITAT DE VALÈNCIA

Session: SMR-05 – Perception and Quality Models for Images & Video

# Objective

Reference Image

Additive Gaussian Noise



Spatially Correlated Noise

Two-alternative forced choice (2AFC) experiment:

Which distorted image is closer to the reference image?

What is the perceptual distance between two images?

# Traditional Perceptual Systems

Most systems consist of a cascade of linear & non-linear functions that focus on different psychophysical attributes.
System $\mathbf{S}$ maps input $\mathbf{x}$ to responses $\mathbf{S}(\mathbf{x})$ where

$$\mathbf{x_0} \xrightarrow{\mathbf{S_0}} \mathbf{x_1} \ldots \xrightarrow{\mathbf{S_i}} \mathbf{x_i}$$

and each layer $\mathbf{S}_i$ is made up of a linear and nonlinear function

$$\begin{tikzcd} \mathbf{x_i} \arrow[r, "\mathbf{L_i}"] \arrow[rr, bend right] & \mathbf{y_i} \arrow[r, "\mathbf{N_i}"] & \mathbf{x_j} \end{tikzcd}$$

This can either be trained individually so each layer recreates the desired psychophysical effect, or to replicate experimental values.

# Human Phenomena

We'll look at 4 main characterstics of the human visual system; gamma adaptation, colour adaptation, LGN behaviour, V1 behaviour.

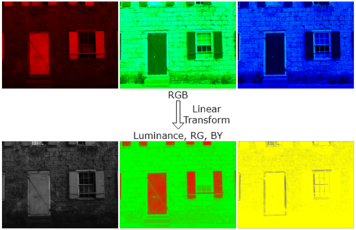
## Gamma Adaptation

Non-linear transform that aims to enhance the response in the low-luminance regions.

# Colour Adaptation

## Opponent Colour Space

More efficient for the visual system to record differences in cone responses; Long, Medium, Short wave light.




- Tune each channel based on other channel values
- Calculate opponent colour channels

## Chromatic Adaptation

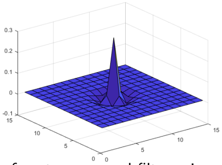
Non-linear colour adaptation that accounts for colour constancy in human visual system (similar to Von Kries).

Example of chromatic adaptation to remove a tint.

# Lateral Geniculate Nucleus (LGN) Behaviour

## Center-Surround Filters

LGN cells often modelled as center-surround, or difference of two Gaussians.




Example of center-surround filter using difference of two Gaussians.

On Off

Example of the positive (on) and negative (off) response when using a center-surround filter.

## LGN Normalisation

LGN cells have a non-linear divisive supressive field, which can take the form of local contrast masking.

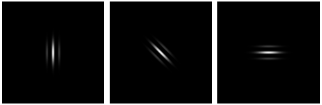
Example of center-surround $\rightarrow$ LGN normalisation.

Note the emphasis on local contrast not global contrast due to small receptive field filter.

# V1-Like behaviour

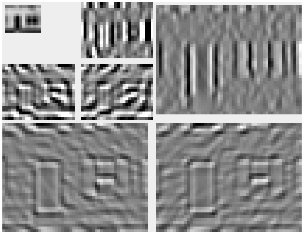
## Orientation and Multiscale

Consider different scales and orientation, like a wavelet pyramid.



$0^\circ$ $45^\circ$ $90^\circ$

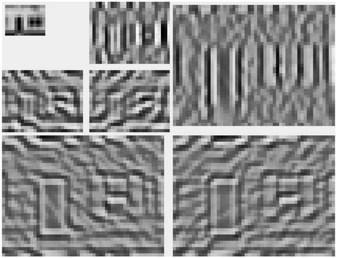
Gabor wavelets generated with different orientations.



Output of a wavelet pyramid with 2 stages and 3 orientated wavelets

## Divisive Normalisation

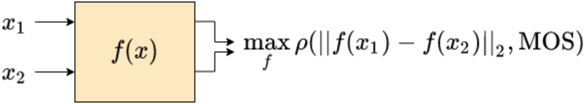
Non-linearity to account for orientation and frequency masking.



Example of frequency masking on the output of a wavelet pyramid.

# Using a Neural Network

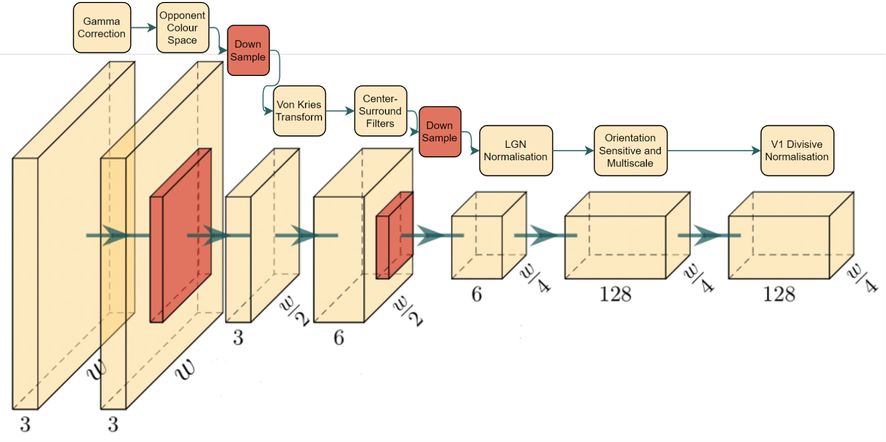
Replacing the system **S** with a neural network **f**(**x**) and optimise for some objective function.




where MOS is the mean opinion score. Usually **f**(**x**) is a commonly used architecture (AlexNet, VGG, ect.)
This disregards what we know about the human perceptual system.

Why not combine them?

# PerceptNet




PerceptNet number of parameters: $36.3k$ AlexNet number of parameters: $24.7m$

# Linear and Nonlinear Functions

## 2D Convolution Operation

Linear transform that has efficient implementations with images.
Considers spatial information.

$$y(x) = x \circledast h$$

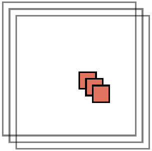
We can vary the number of filters used and how they combine.

## Generalised Divisive Normalisation (GDN)

Non-linear transform inspired by the divisive normalisation found in vision cells.
Tensors are normalised for each pixel across channels.
Considers spatial information if the network has downsampling.

$$y_i = \frac{x_i}{\sqrt{\beta_i + \sum_j \gamma_{j,i} \cdot x_j^2}}$$

where $i$ and $j$ run over channels.



Red colour is what is normalised together.

# Datasets

| Dataset | Number of Samples | Number of Distortions | Target |
|---|---|---|---|
| TID2008[1] Train | 1428 | 17 | MOS |
| TID2008 Test | 272 | 17 | MOS |
| TID2013[2] | 3000 | 24 | MOS |
| CSIQ[3] | 899 | 6 | MOS |
| LIVE[4] | 982 | 5 | MOS |
| BAPPS[5] Train | 151.4k | 425 | 2AFC Proportion |
| BAPPS Test | 36.3k | 425+ | 2AFC Proportion |

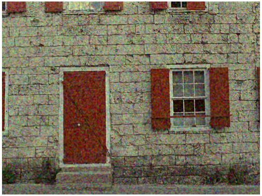
[1] Ponomarenko et al. 2009.
[2] Ponomarenko et al. 2013.
[3] Larson and Chandler 2010.
[4] Sheikh et al. 2005.
[5] Zhang et al. 2018.

# Perceptual Datasets: TID, CSIQ, LIVE

Traditional two-alternative forced choice (2AFC) experiments, with mean opinion score (MOS) calculated from the results.
Experiments were performed on calibrated LCD monitors, in exact experimental conditions.

Reference Image

Additive Gaussian Noise



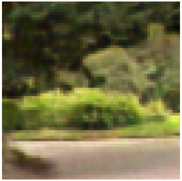
Spatially Correlated Noise

MOS calculated from these results.

# BAPPS

Berkeley Adobe Perceptual Patch Similarity (BAPPS) contains two types of judgements: **Two Alternative Forced Choice (2AFC)** and Just Noticeable Differences (JND).
Experiments were performed on Amazon Mechanical Turk (AMT).

Reference Patch



CNN Based 1

CNN Based 2

0% of people said that the CNN Based 2 was closer to the reference than CNN Based 1.

# Results

| Method | Trained On | Pearson Correlation (Spearman Correlation) with MOS | | | |
|---|---|---|---|---|---|
| | | TID2008 Test | TID2013 | CSIQ | LIVE |
| SSIM | | 0.51 (0.53) | 0.62 (0.60) | 0.77 (0.84) | 0.84 (0.95) |
| MS-SSIM | | 0.78 (0.80) | 0.78 (0.80) | 0.81 (0.91) | 0.77 (0.97) |
| $\text{FSIM}_c$ | | 0.79 (0.84) | 0.79 (0.81) | 0.82 (0.93) | 0.77 (0.92) |
| NLAPD (with GDN) | TID2008 | 0.81 (0.82) | 0.82 (0.81) | 0.90 (0.92) | 0.88 (0.96) |
| AlexNet (with ReLU) | TID2008 | 0.89 (0.89) | **0.93 (0.91)** | **0.95** (0.95) | 0.88 (0.94) |
| AlexNet (with GDN) | TID2008 | 0.91 (0.91) | 0.92 **(0.91)** | 0.94 (0.95) | 0.93 (0.95) |
| PerceptNet | TID2008 | **0.93 (0.93)** | 0.90 (0.87) | 0.94 **(0.96)** | **0.95 (0.98)** |
| LPIPS AlexNet (tune) | ImageNet + BAPPS | 0.74 (0.75) | 0.76 (0.76) | 0.88 (0.93) | 0.85 (0.96) |
| LPIPS AlexNet (scratch) | BAPPS | 0.47 (0.47) | 0.58 (0.57) | 0.72 (0.80) | 0.77 (0.89) |
| PerceptNet (tune) | TID2008 + BAPPS | 0.67 (0.72) | 0.75 (0.76) | 0.81 (0.88) | 0.85 (0.94) |
| PerceptNet (scratch) | BAPPS | 0.56 (0.67) | 0.67 (0.72) | 0.77 (0.84) | 0.80 (0.93) |

# Results

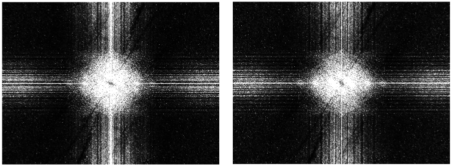
| Method | Trained On | 2AFC Accuracy (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Average | Trad-itional | CNN Based | Super Res | Video Deblur | Colour-isation | Frame Interp |
| LPIPS AlexNet (tune) | ImageNet + BAPPS | 69.7 | 77.7 | 83.5 | 69.1 | 60.5 | 64.8 | 62.9 |
| LPIPS AlexNet (scratch) | BAPPS | 70.2 | 77.6 | 82.8 | 71.1 | 61.0 | 65.6 | 63.3 |
| LPIPS PerceptNet (tune) | TID2008 + BAPPS | 67.8 | 69.4 | 81.3 | 70.6 | 60.9 | 61.9 | 62.6 |
| LPIPS PerceptNet (scratch) | BAPPS | 69.2 | 75.3 | 82.5 | 71.3 | 61.4 | 63.6 | 63.2 |
| AlexNet | TID2008 | 63.2 | 56.1 | 77.4 | 66.1 | 58.6 | 61.6 | 56.2 |
| PerceptNet | TID2008 | 64.9 | 58.1 | 80.5 | 68.3 | 59.6 | 61.6 | 58.2 |

Two-alternative forced choice (2AFC) accuracy scores for various architectures, all evaluated on the BAPPS

# Visualisations

We can see the channels in the perceptual space that contribute most to the $\ell_2$ distance when considering JPEG2000 transmission errors.

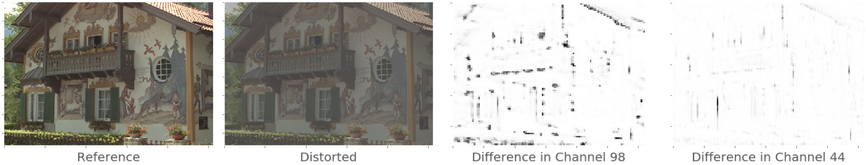

Difference in the output of the network for a reference image and distorted image.



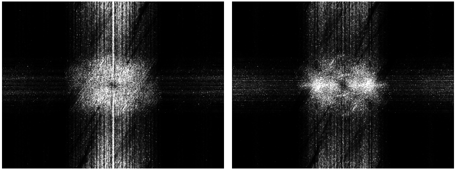
Receptive fields for channels 88 and 64.

# Visualisations

And for a different type of distortion; Contrast change

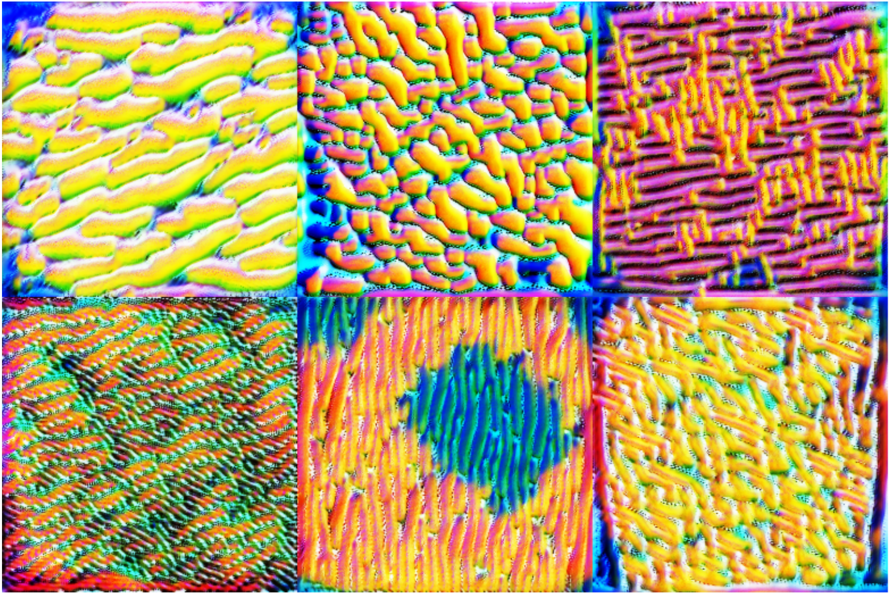



Difference in the output of the network for a reference image and distortion image.



Receptive fields for channels 98 and 44.

# Visualisations



# Conclusion

- Describe a transformation inspired by the human visual system to predict human perceived distance.
- Show that this transformation generalises well between datasets whilst having a small number of parameters to learn.
- The transformation displays a number of properties that are present in the human visual system.

# Improvements

- Spatial divisive normalisation
- Enforcing multiscale (e.g. wavelet pyramid, multiscale convolutions)
- More robust visualisations of the filters

# Thank you

Email: alex.hepburn@bristol.ac.uk

Code: https://github.com/alexhepburn/perceptnet

Additional code: https://github.com/alexhepburn/expert