

# Perception of Pictures

Anil Kokaram

# Outline

- Motivating a study of picture quality assessment
- The eye and Psychophysical measurements
- Deriving realistic metrics
- Standards

# Why is visual perception important?

- Video processing algorithms are not 100% bulletproof. Denoising goes wrong, compression goes wrong.
- “Golden eyes” are experts who are acknowledged as the “quality controllers” for high end video quality. Manual review is very important in big budget productions. It does not scale.
- Scale of video processing (streaming, consumer usage) too large to use manual review all the time
- Human visual perception is not the same as a mathematical measure of Error.

# What are our questions?

Do the output pictures look better than the input?

Do the output pictures look like the input?



Do the output pictures look good?

# Things go wrong

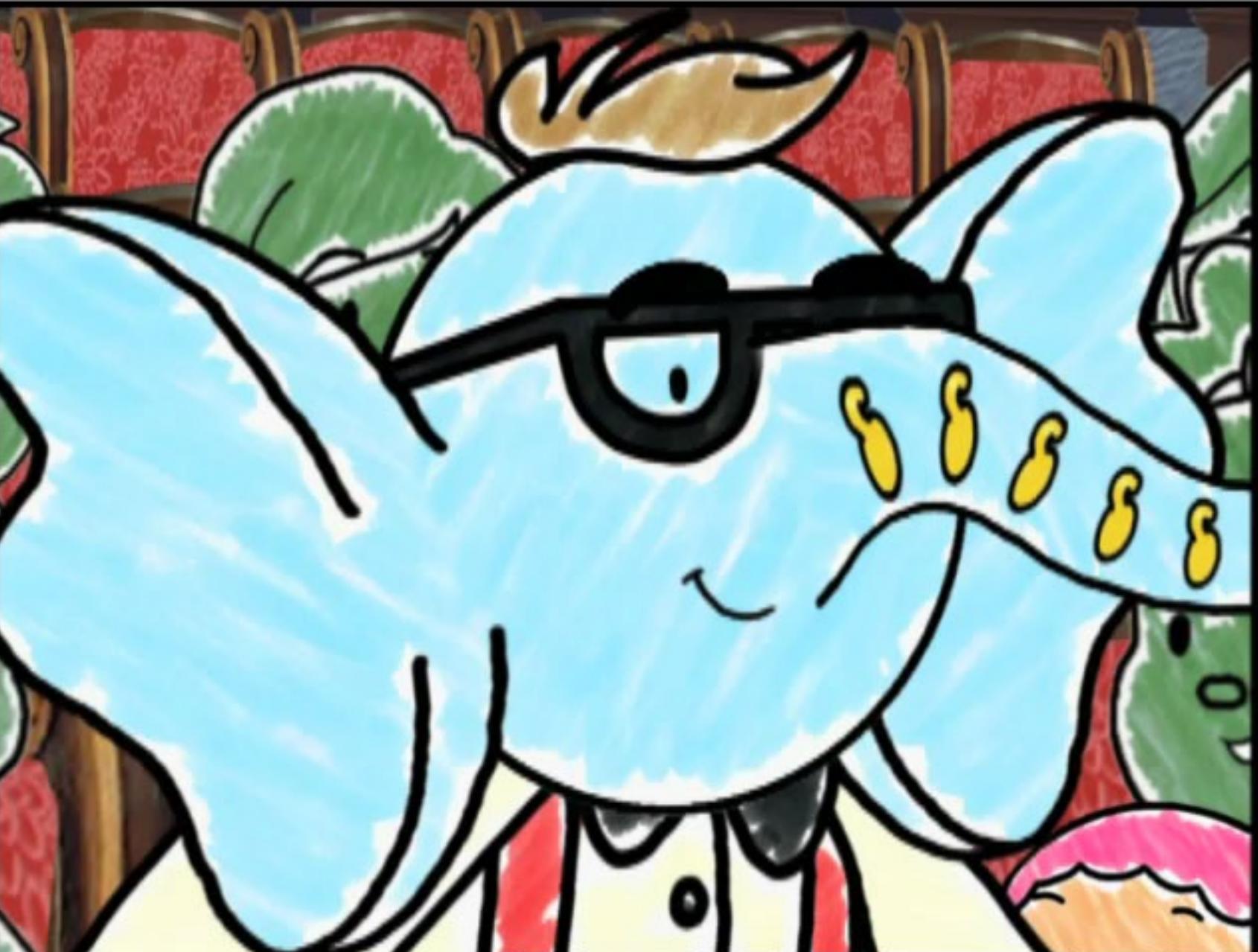


Blockiness / Blocking

Blockiness / Blocking  
Banding (at the horizon with the sea)



Moire

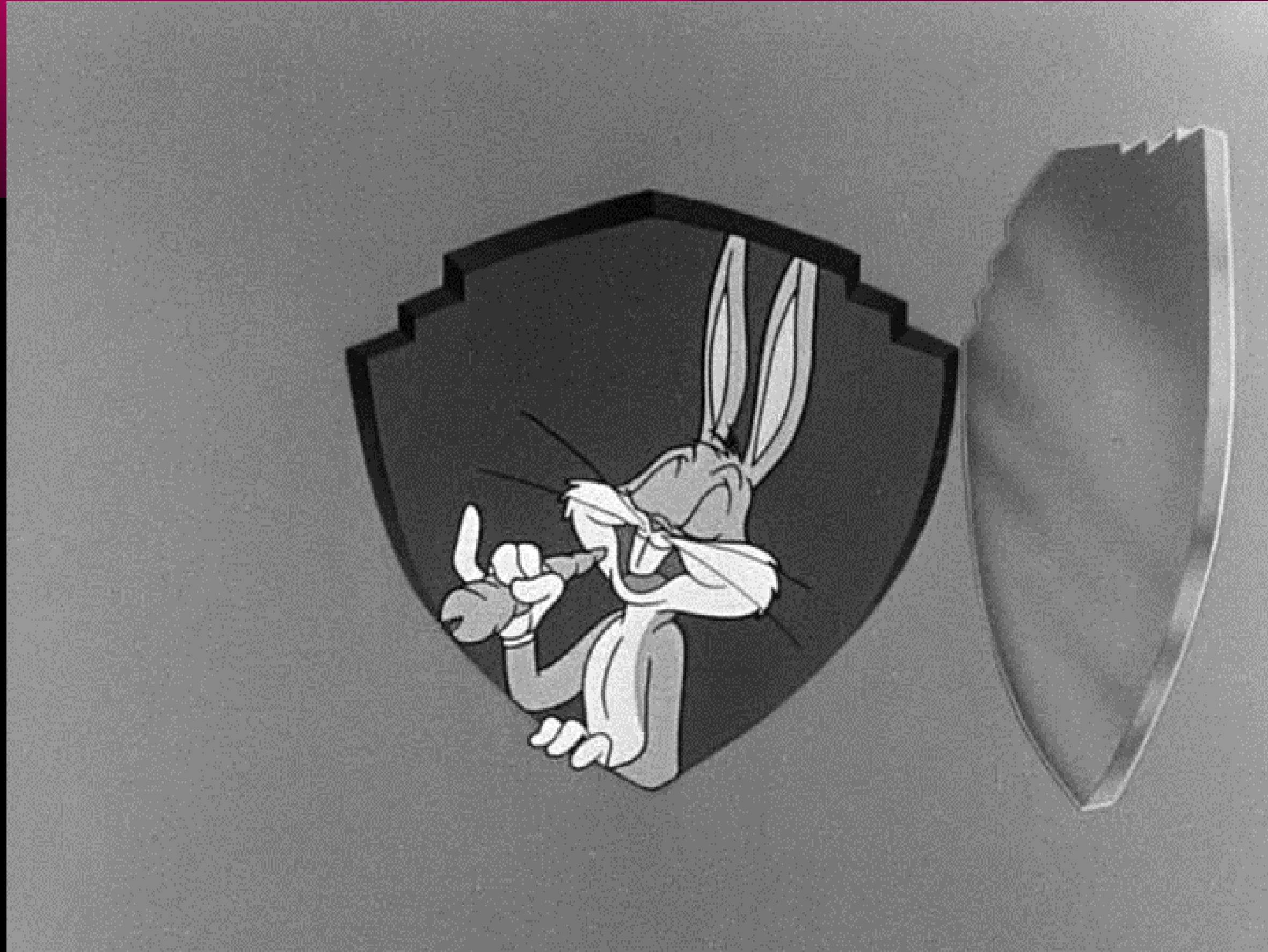


Pixelation



Tearing

## Noise/Grain Noise



# Metrics (reference and no reference)



A reference metric answers the question “Do the output pictures look like a given set of golden reference” images?

A no-reference metric answers the question “How good does this picture look?”

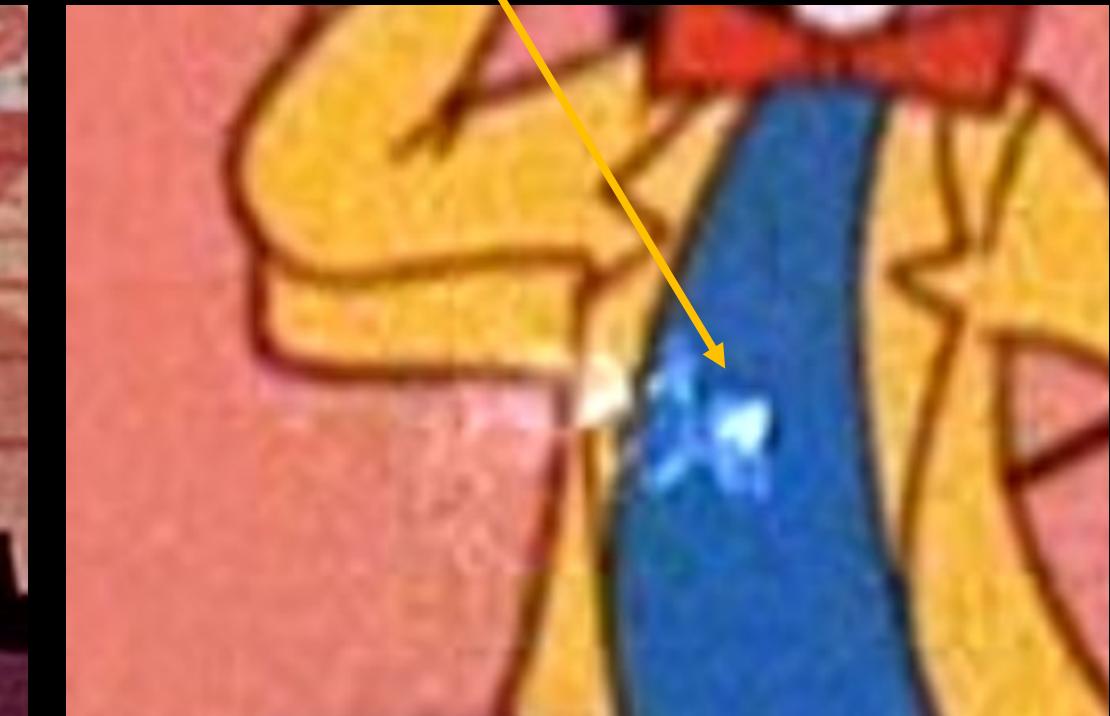
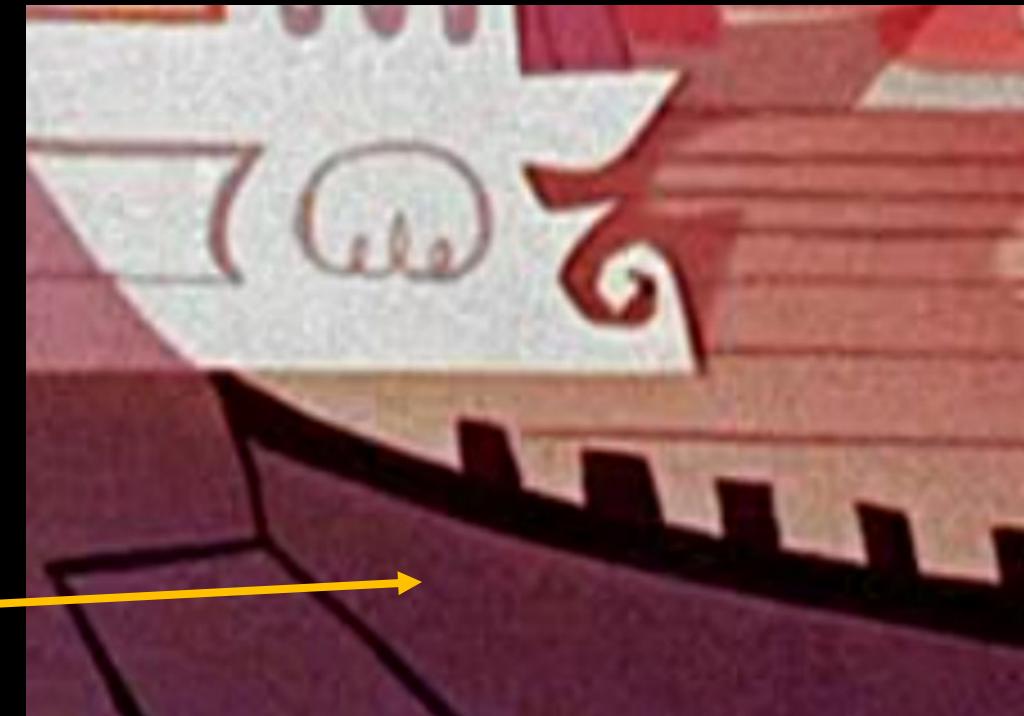
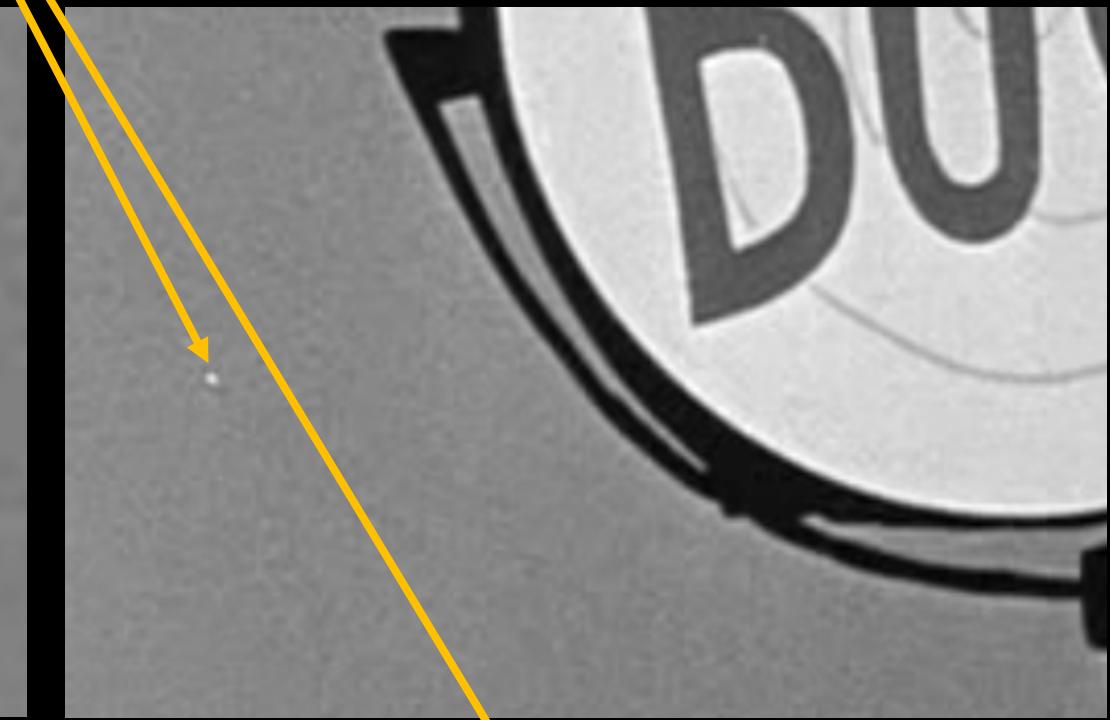
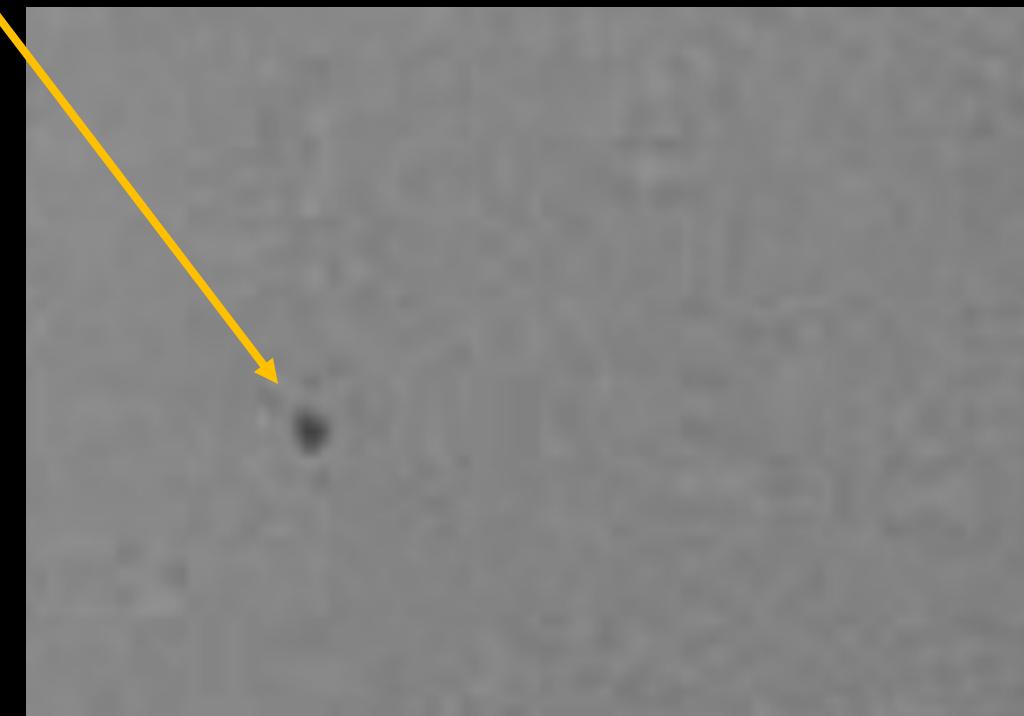
Blotches  
(Missing Data)

Dirt

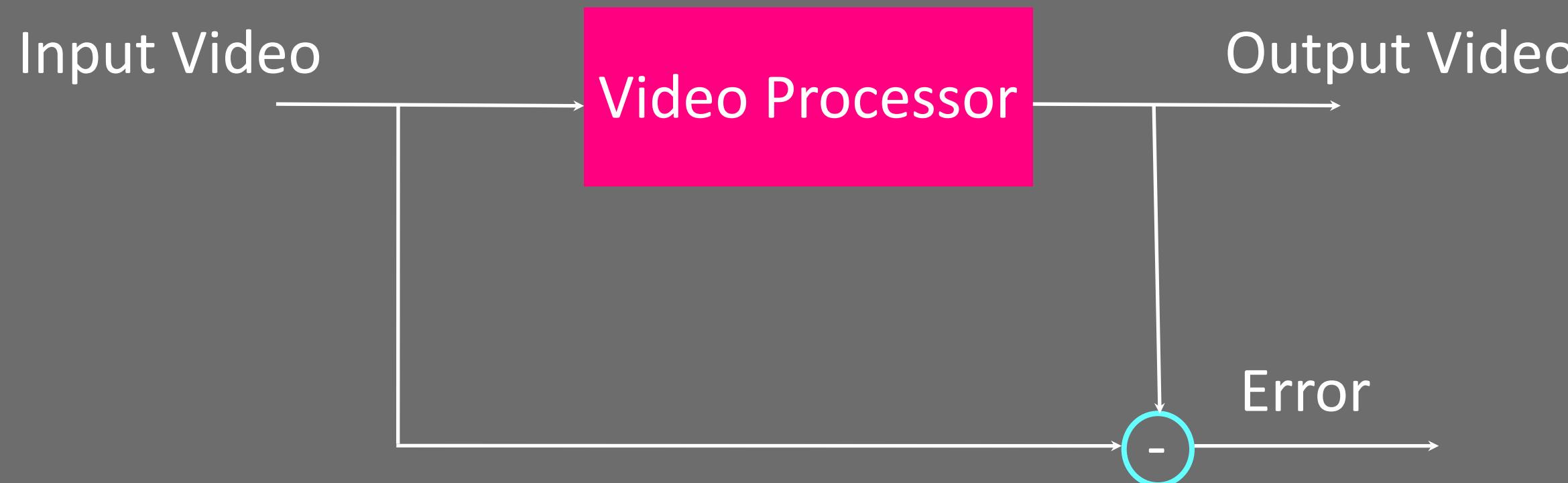
Sparkle

Old Film  
Artefacts

Grain



# Metrics (reference and no reference)



In video compression we want the output of the compression/decompression process to look like the input so the reference image frames are the input themselves

For high quality content provided by Netflix, AppleTV etc, we have a known reference so “reference metrics” are important there.

For User Generated Content (like mobile camera uploads) we don’t have a reference so “non-reference metrics” are important there.

# Consider these ...

Do the output pictures look like the reference?

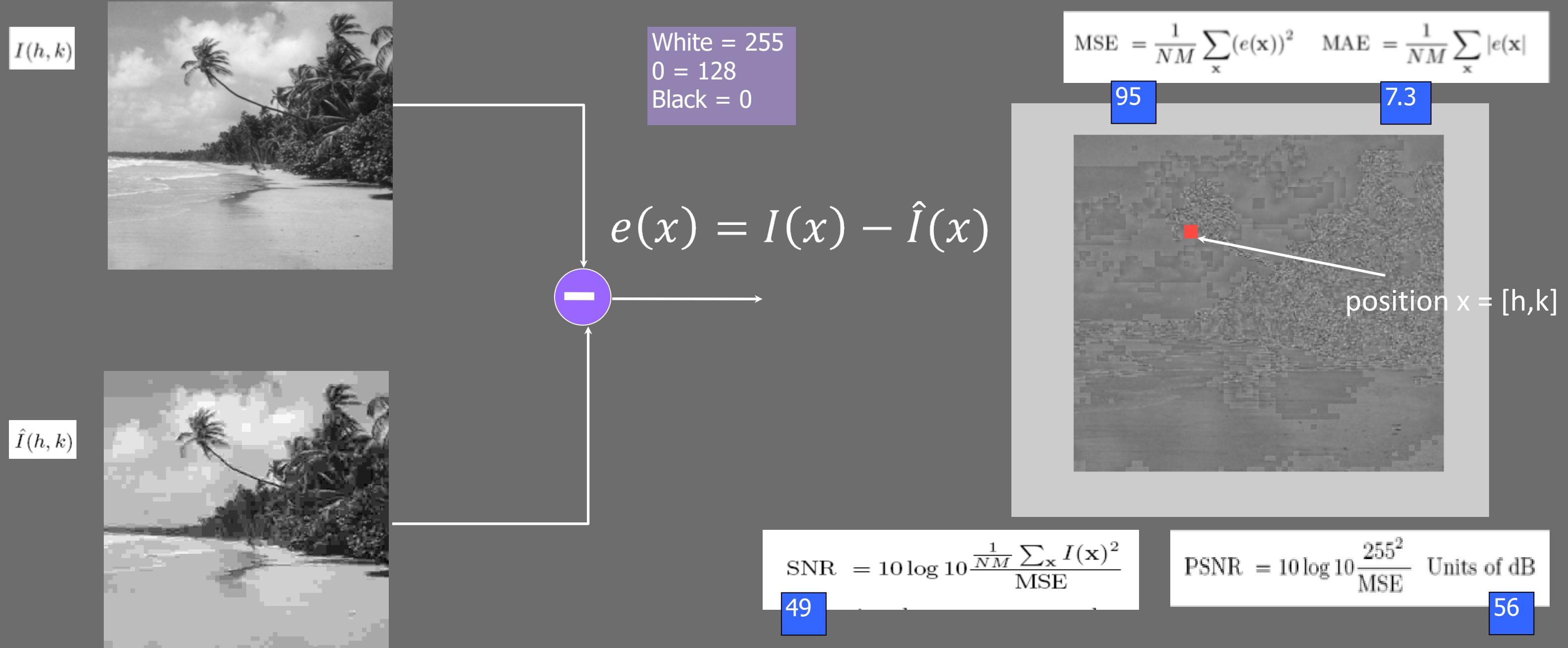


Reference



Output

# Squared Error



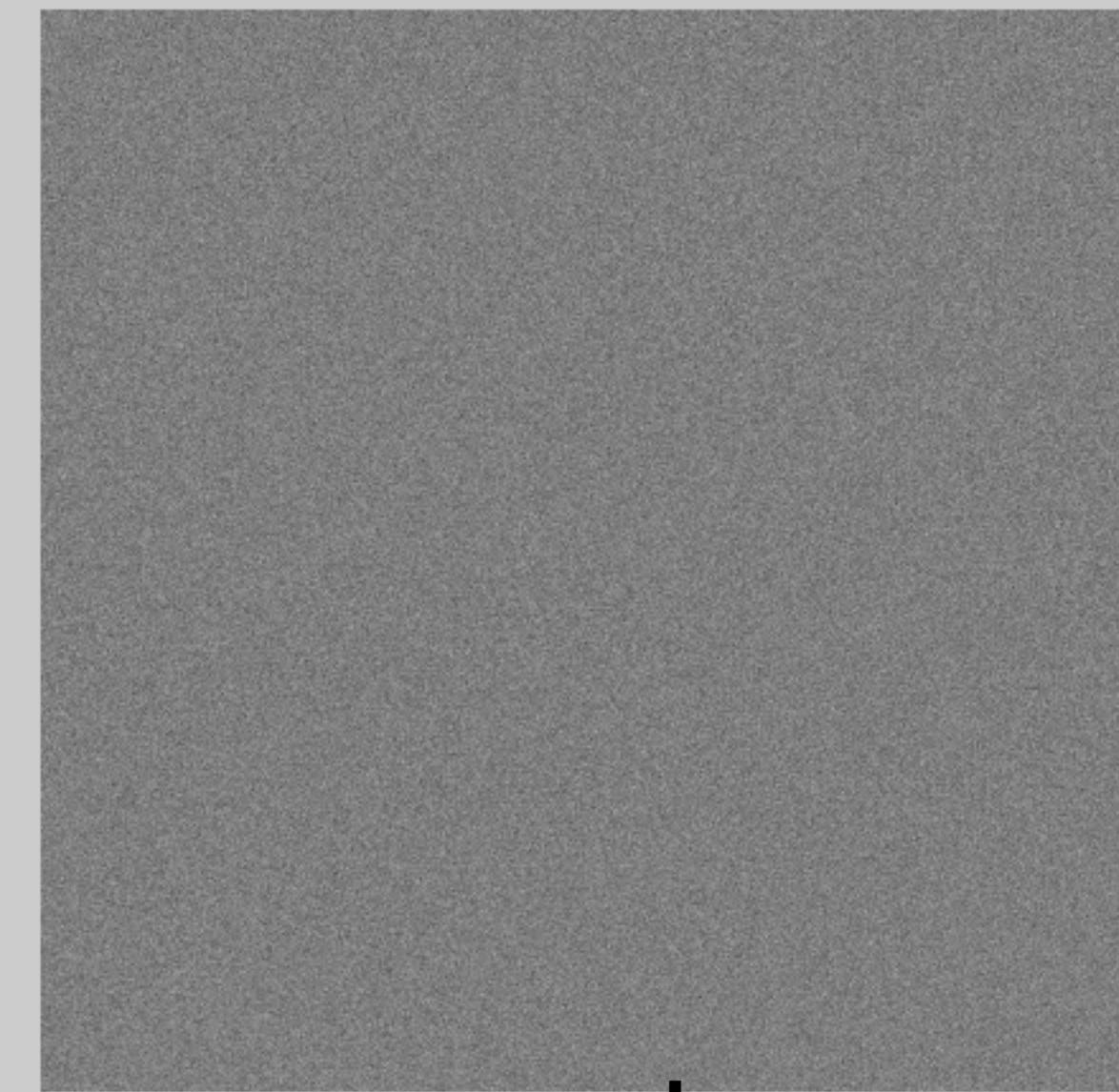
# Output contains added noise

SNR ~ 27.5dB, MSE ~ 116



PSNR = 27.5 dB, MSE = 116

Error between original and degraded



$$e(\vec{x})$$

# Output contains missing blocks

SNR ~ 27.5dB, MSE ~ 116



PSNR = 27.5 dB, MSE = 116

Error between original and degraded



$$e(x)$$

# Are these the same?



28 dB



27dB

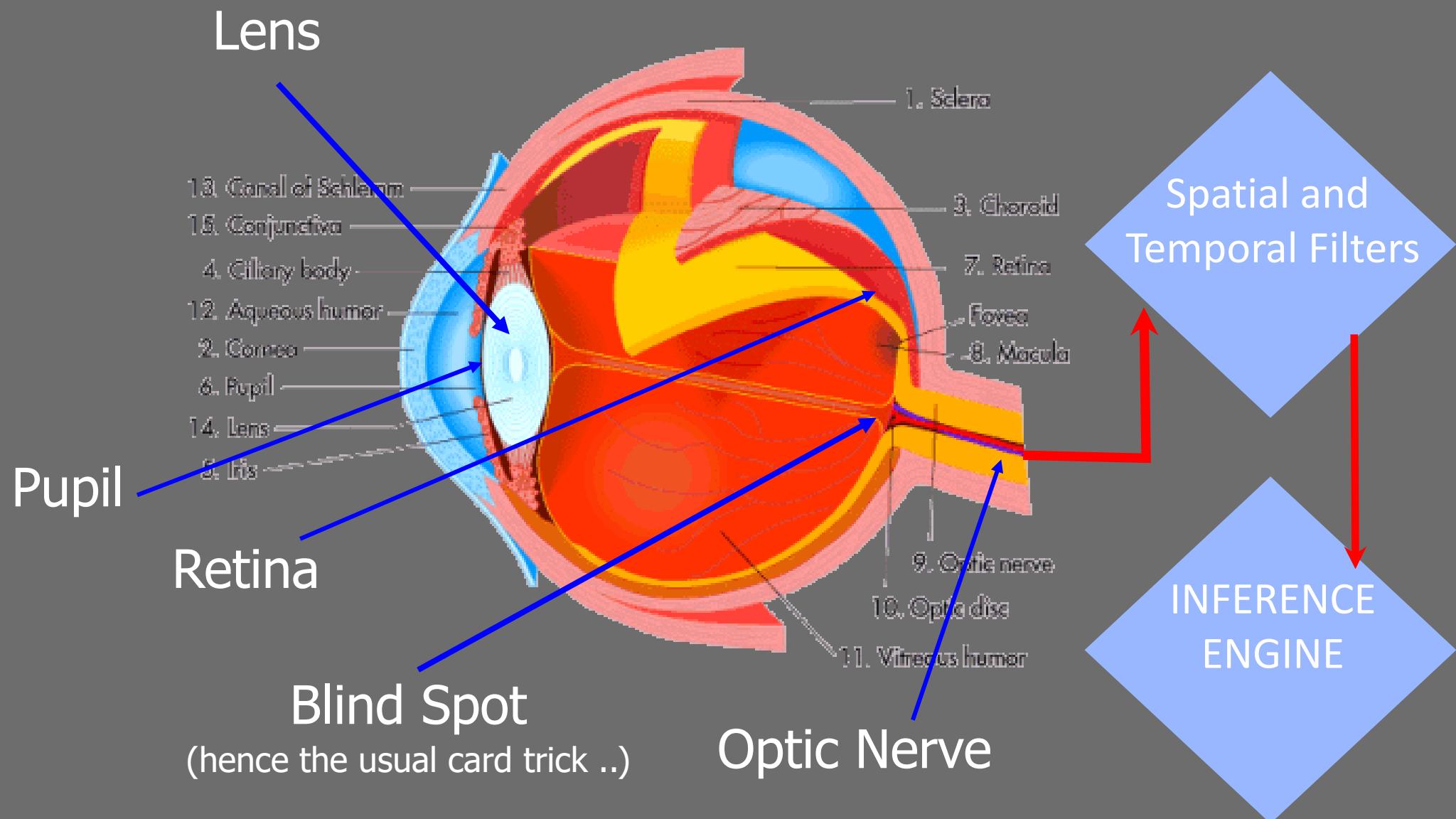


27dB

# PSNR/MSE (used as reference metrics)

- PSNR not that bad in fact : if the error is distributed evenly in the picture, and is the same type of error from picture to picture, then it can compare one picture to another
- These are big IFs
- PSNR used a lot because it is amenable to mathematical analysis: you can optimize systems for PSNR (mean squared error leads to differentiable loss functions)
- But pictures are for People. So to understand what “looks like” means .. you have to study how people see what they see.

# The eye is a pinhole camera attached to impressive DSP post-processor and Learning units



- Light is focussed onto the retina
- Electrical Impulses from the retina are channelled by the optic nerve to the Visual Cortex
- The Visual Cortex does a whole bunch of smart things including filtering, object recognition, edge detection.
- In 'primitive animals' A LOT of processing happens just behind the retina. Frogs and Rabbits have TEMPLATES for spotting birds of prey.
- Our motion sensitivity is better at the periphery of vision than at the centre. [Helps to avoid people sneaking up on you.]

# Intensity Perception and Weber's Law $\Delta I / I = k$

$I = 130$

$I = 60$

$I + \Delta I$

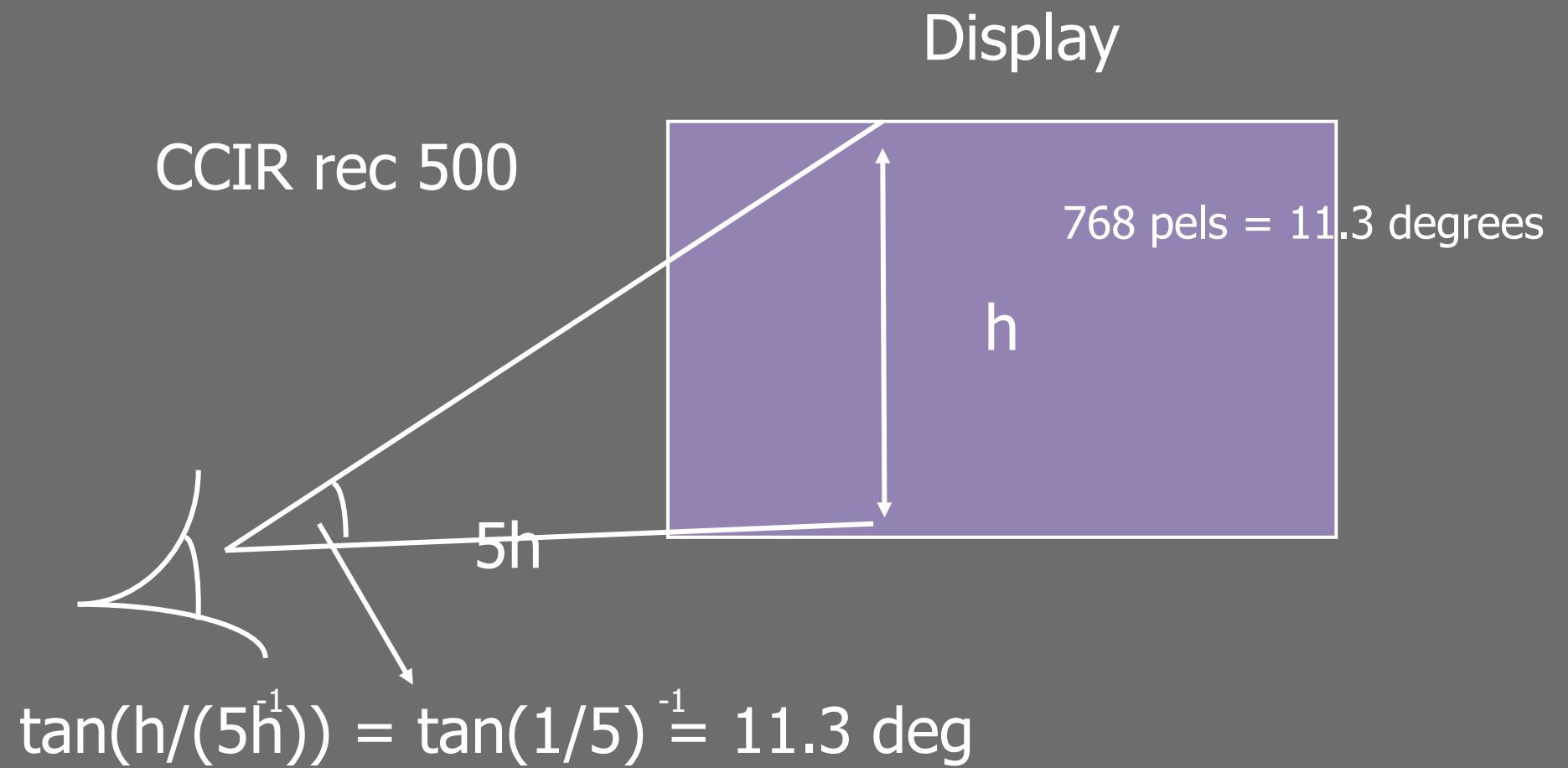


This is crucial for understanding High Dynamic Range image and video data formats

Contrast on left worse than on right but  $\Delta I = 10$  in both cases

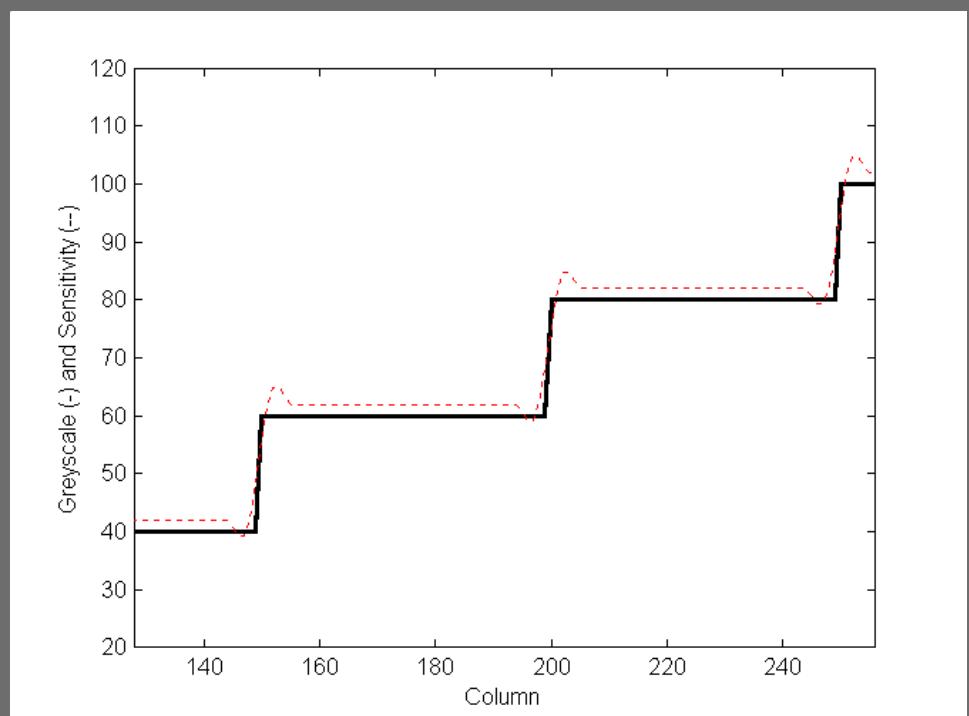
# Perception and angles

- What matters is the degrees of arc subtended by an object on the retina.
- Hence distance to the image plays a part.
- A standard exists for viewing distance.

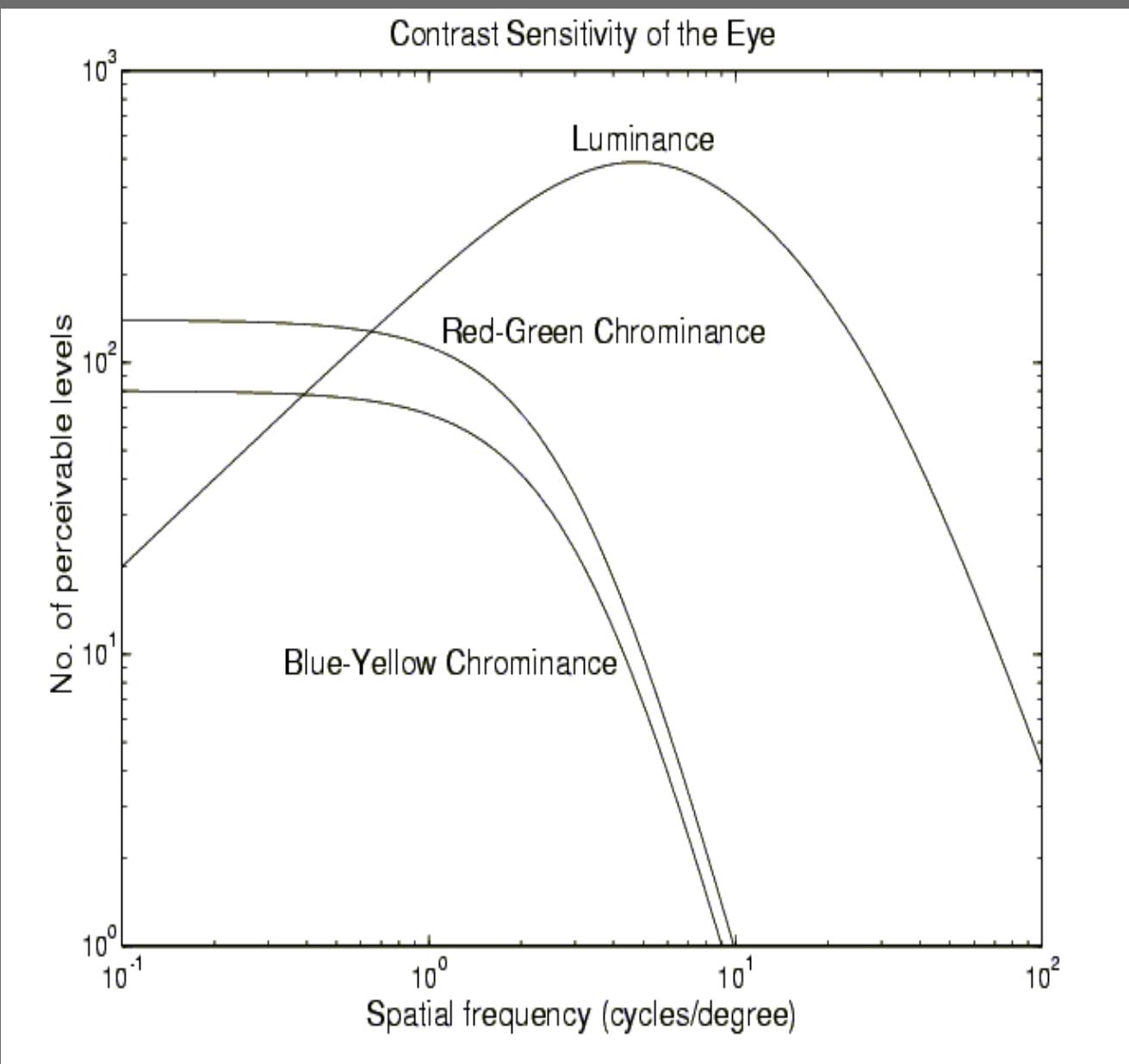


Visual acuity is 1/60 cpd, i.e. you can just resolve 2 lines separated 1/60 degrees.  
Ipad is 2048x1536 .. which is at ths limit when held 41cm away

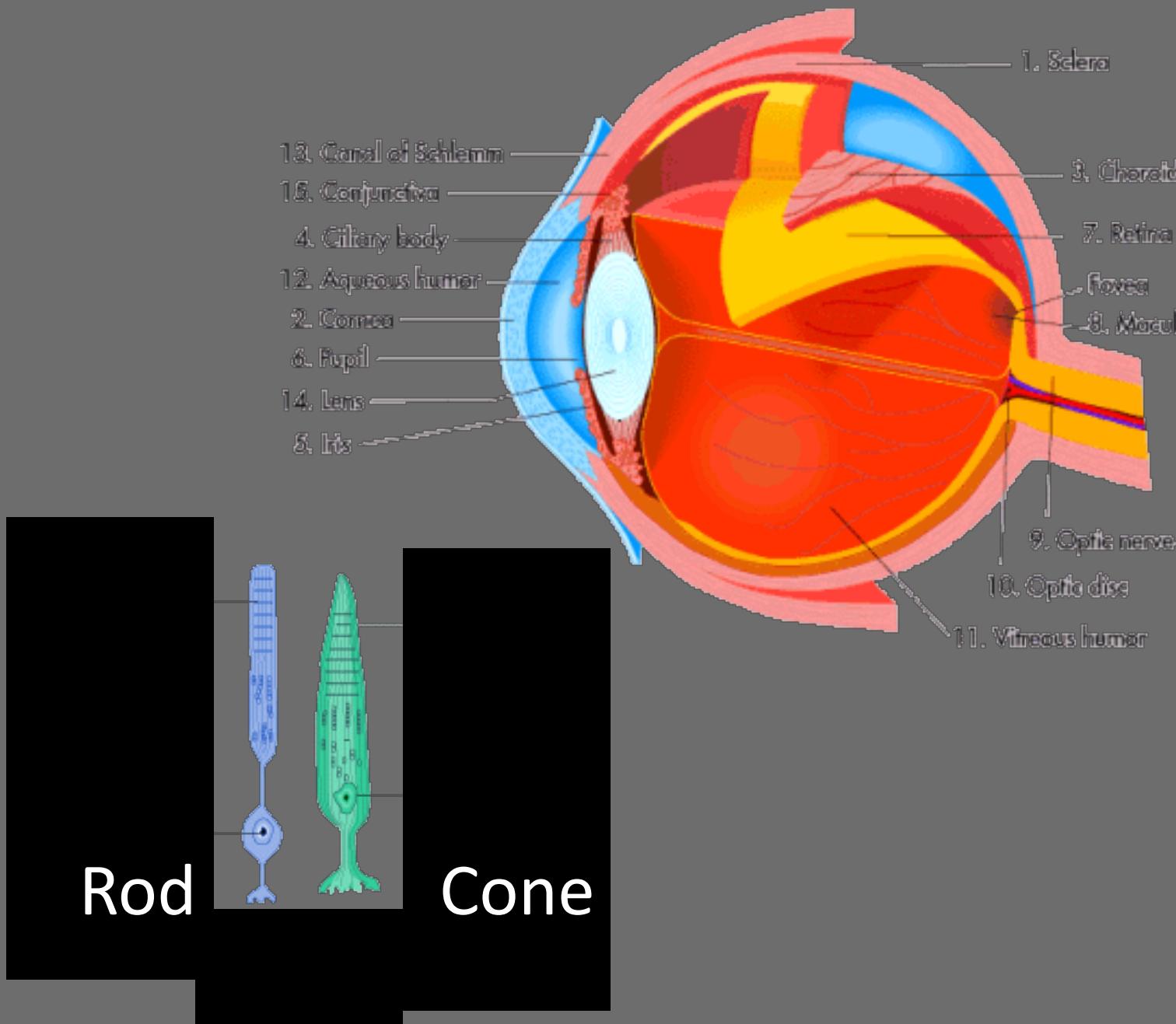
# Mach Bands and Visual Filtering



# Sensitivity to Spatial Frequency



# Eye and colour sensitivity



- Cones allow colour vision, fovea has only cones. 3 types of cones.
- Rods remain active at low light levels
- 120 Million Rods and 7 Million Cones. Hence luminance sampled ALOT more finely than colour
- Hexagonal arrangement of cells.

# How DTV took advantage of Colour Frequency Perception



In 1952 Colour Broadcasts had to be compatible with B/W TV sets.  
How to allow RGB colour to make sense on a B/W set?

# YUV Colour Space

Encode colour as two colour difference signals U/V, as quadrature modulated components in the line signal

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = C \begin{bmatrix} R \\ G \\ B \end{bmatrix}$$

Where  $C = \begin{bmatrix} 0.3 & 0.6 & 0.1 \\ -0.15 & -0.3 & 0.45 \\ 0.4375 & -0.3750 & -0.0625 \end{bmatrix}$



# 1984 CCIR 601 Digital Rec.

- Frame rates chosen to be 30/25 because of interference between scan line clocks in analog TVs and power lines. So 60/50 Hz power lines are to blame.
- Turns out that 13.5 (14.3?) MHz sampling frequency gives the same integer number of pixels in a sec for US and Europe
- Choose YUV-RGB conversion matrix as previously shown.
- “full swing (0-255)” and “studio swing (16-235)” definitions : Studio Swing preserves the black/white guard bands that used to allow analog ccts to have signals below black and white. This is still useful today. Now called Superwhite or whiter-than-white (and for black).
- 1984 Committee made a mistake and used 16-235 .. it should be 16-240 to match the chroma guards

# Exploiting Colour Perception in BT601



Downsample U/V 2:1, Leave Y alone



Downsample U/V 4:1, Leave Y alone

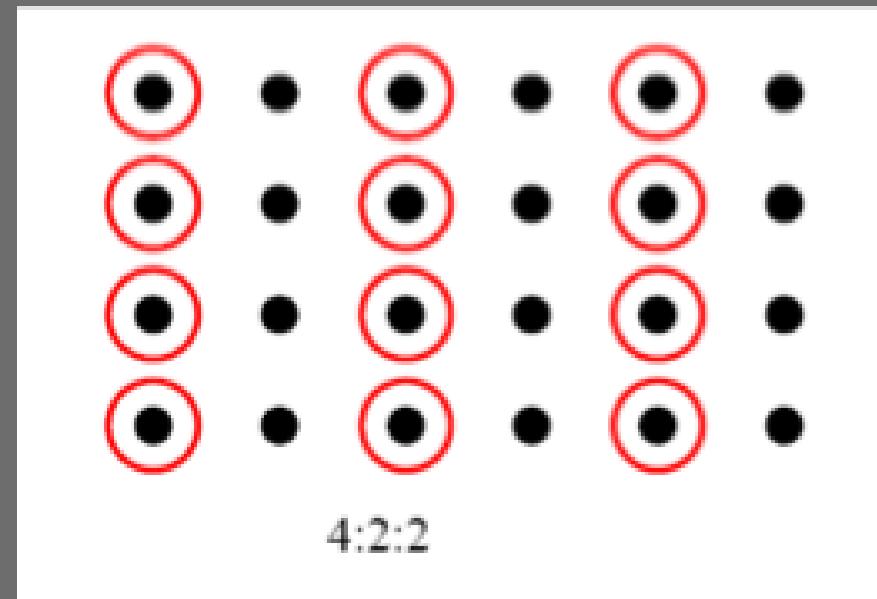


# Downsample Y/U/V 4:1



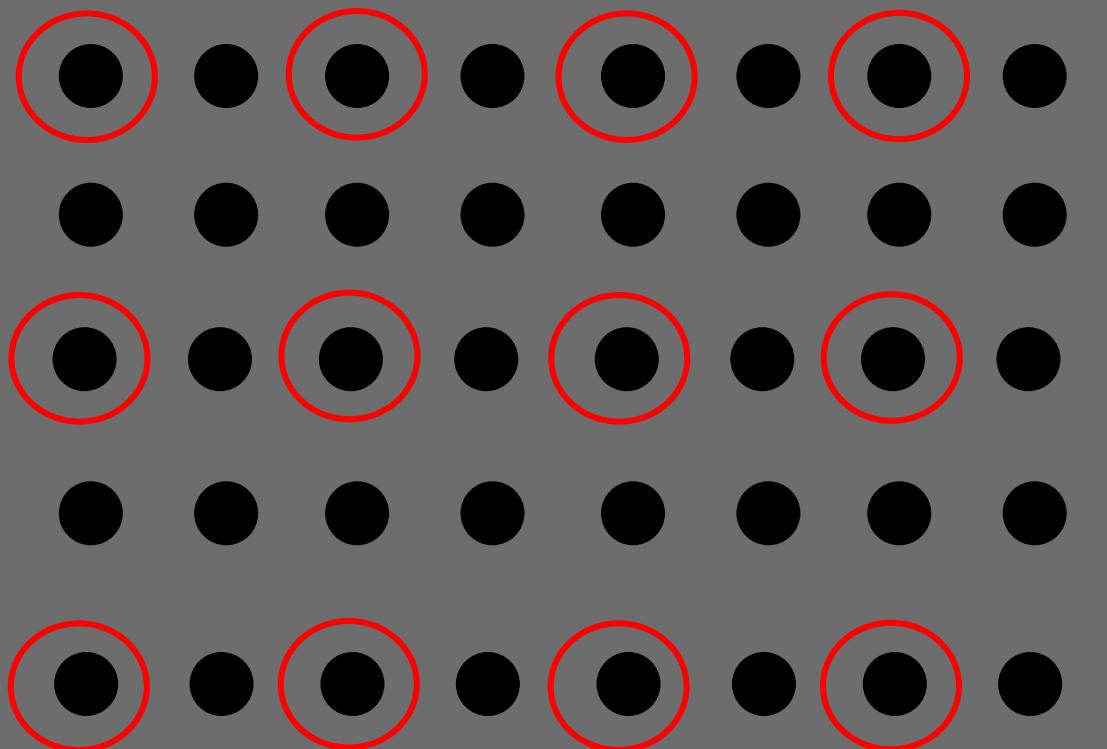
# Hence YUV 420, 422 etc

- We can throw away quite a lot of colour data without perceived effect!
- Digital data gets compressed even before “compression”

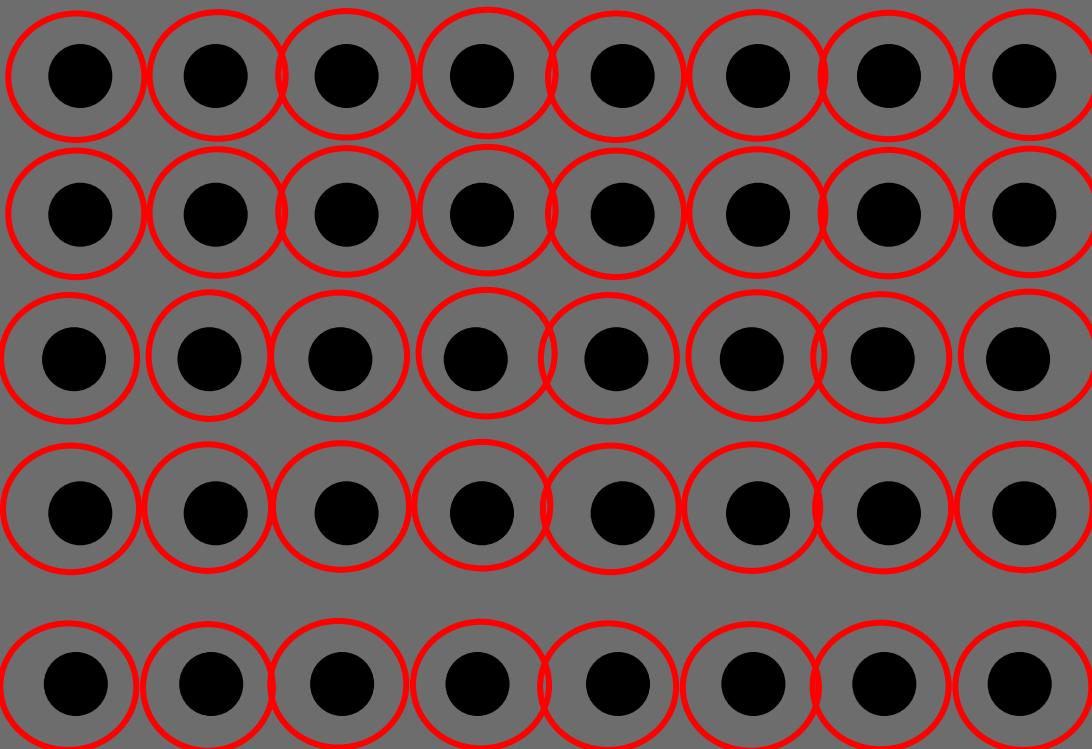


Actually this was also done in Analog TV because the bandwidth for the colour signals was 1/2 that of the Luma.

# YUV 420, 444

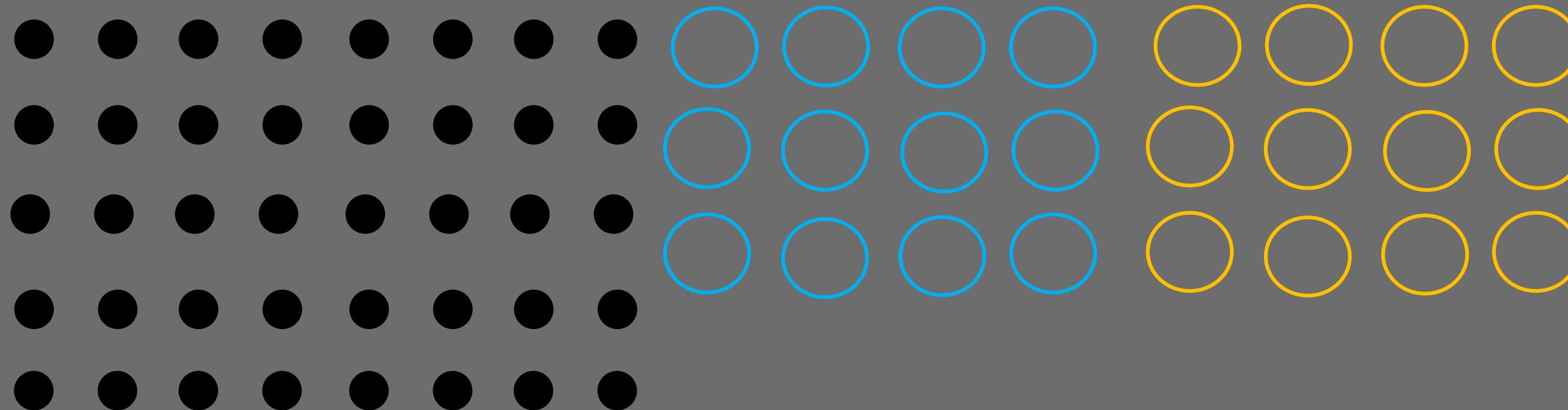


YUV 420 (Streaming media format)  
Y sampled at every site  
UV downsampled by factor of 2 in  
each direction

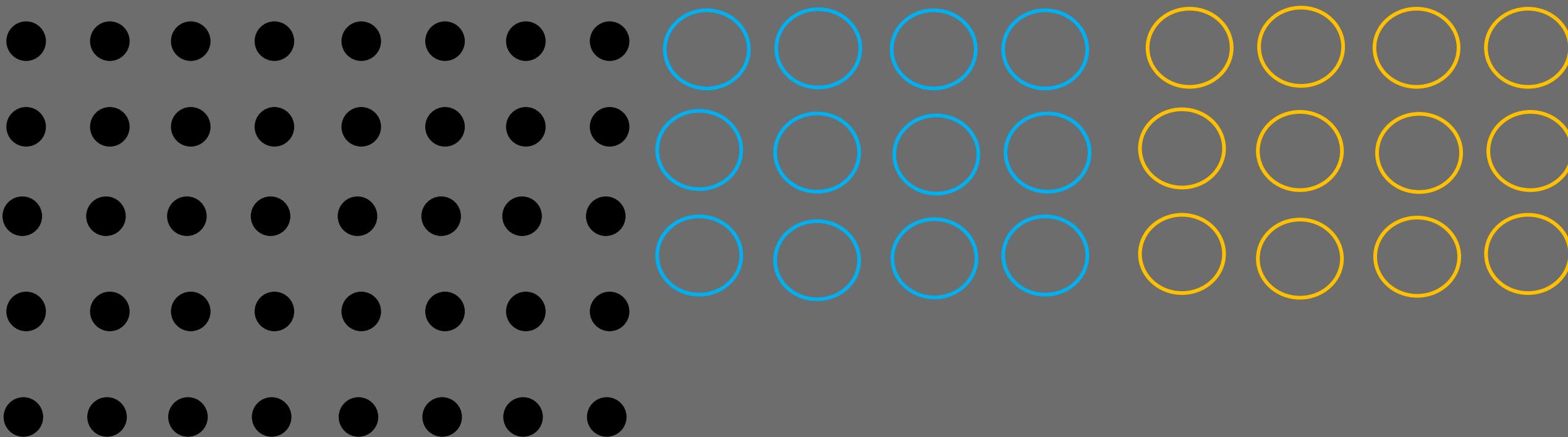


YUV 444  
YUV sampled at every site

# YUV 420 video format



Frame 1



Frame 2

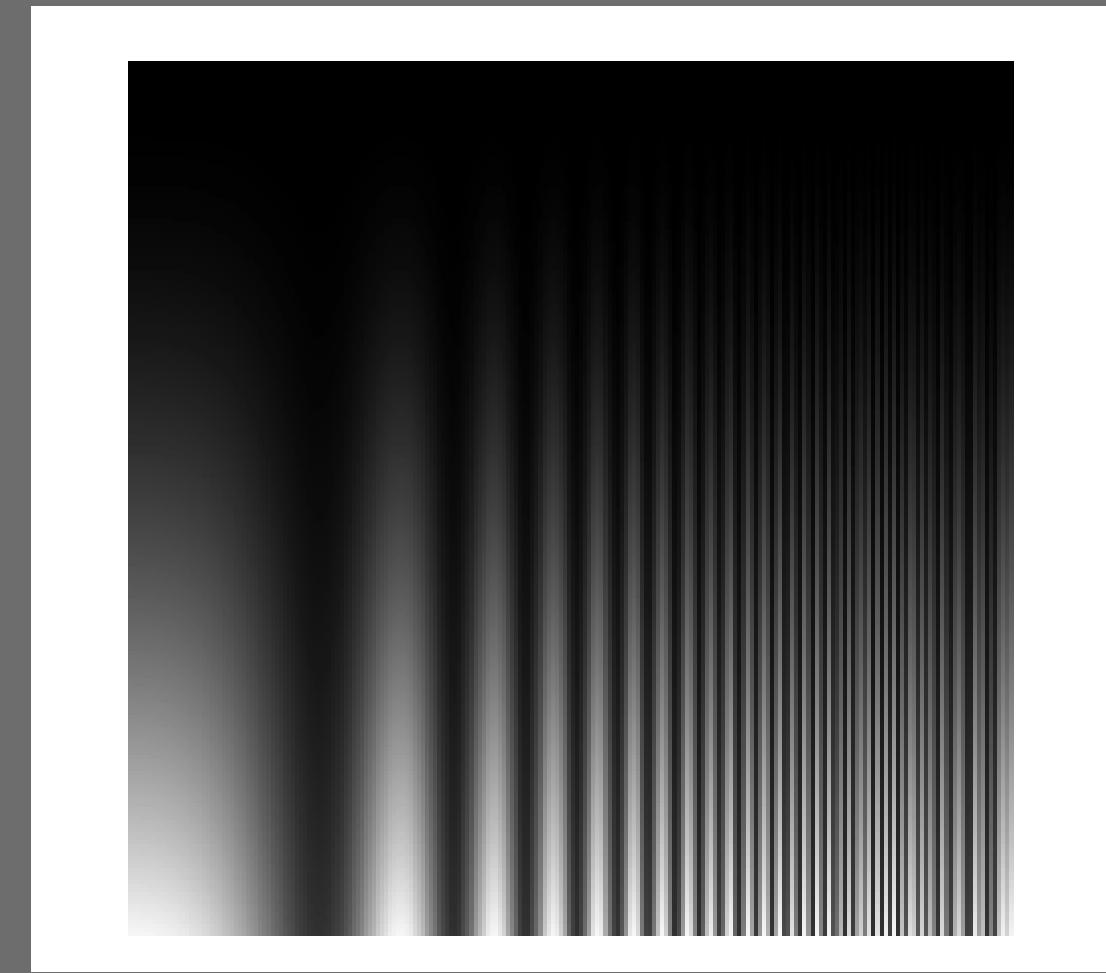
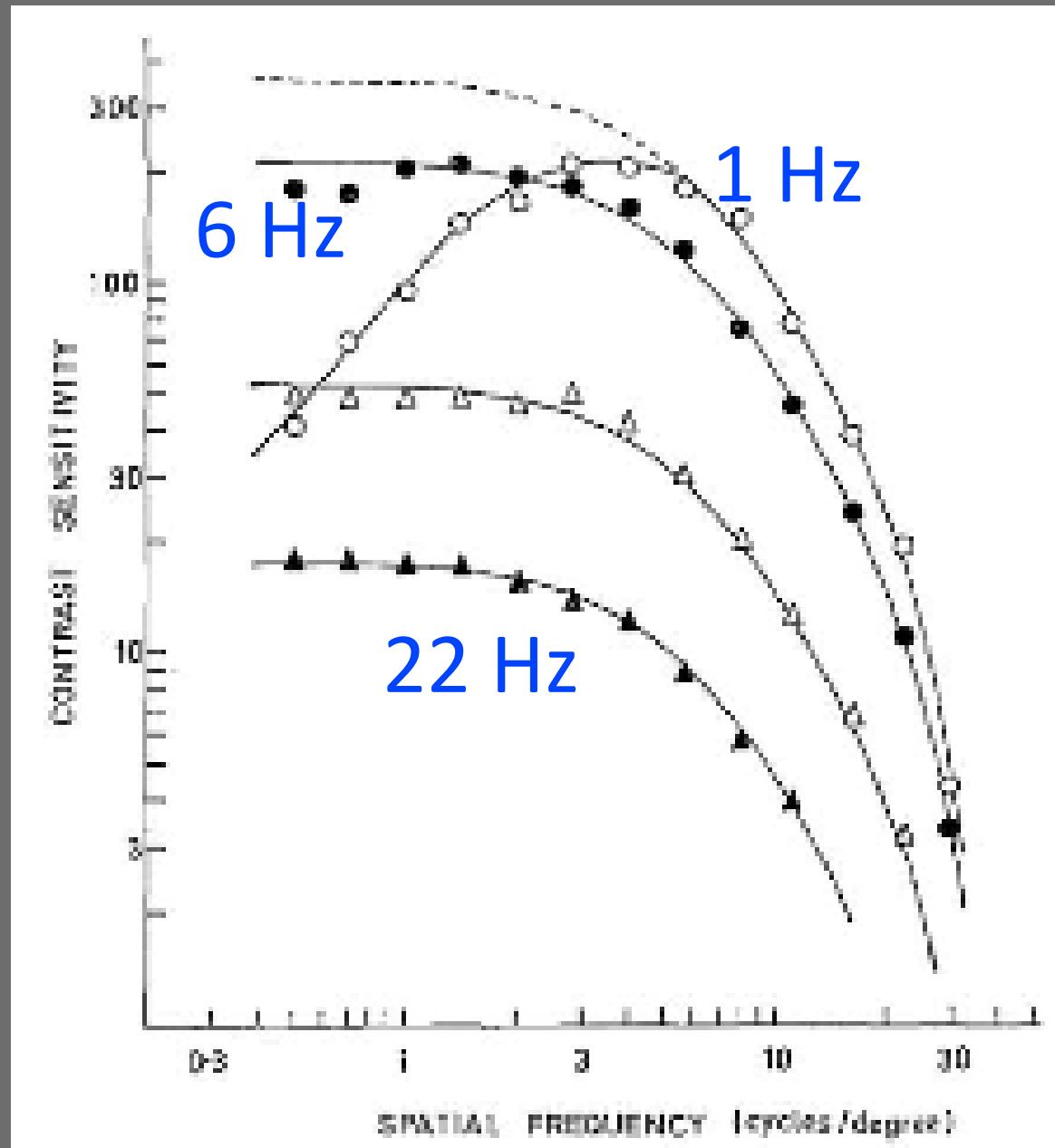
# A note on 601 Vs 709

- In 1984 when 601 was established, the tristimulus colour references were chosen according to SMPTE RP 145, and without a defined gamma to convert from electrical to optical stimulus.
- In 709 it was decided to “fix” all of that. And a new colour reference was defined BT 709 with a defined elec/opt mapping BT1886. This meant almost nothing for U/V but it changed Y alot!
- Accordig to C. Poynton (SMPTE award for square pixels in HD 2012):  
Bummer!

# More perception: Activity Masking



# Temporal Perception/Masking

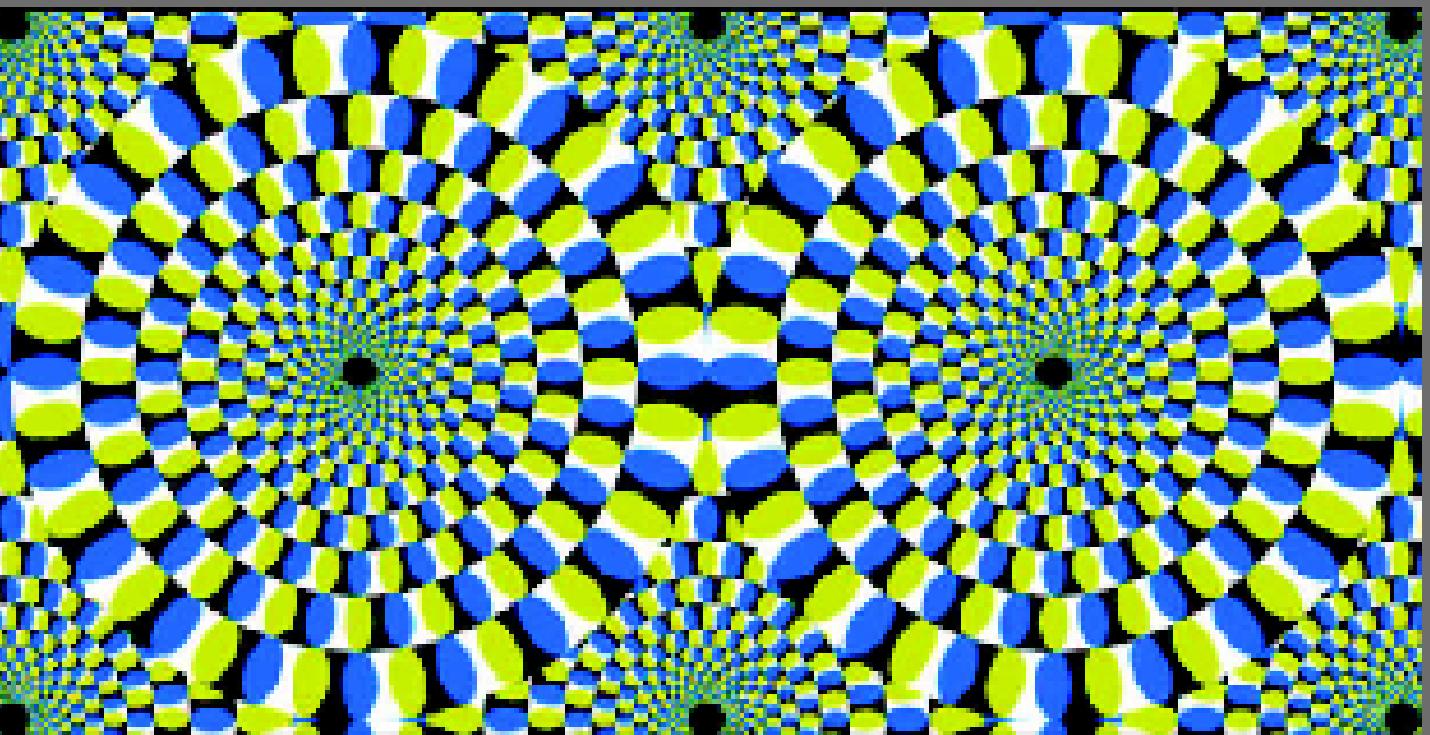


Spatial and Temporal Contrast-Sensitivity Functions of the Visual System

J. G. Rouson  
Physiological Laboratory, Cambridge, England  
(Received 3 March 1966)

# Recent Motion Studies

- Journal of Vision 2005 Backus and Oruc,
- Stirling at Irvine
- Motion perception is controlled by 3 processes
- Global Motion estimator, tiny motion estimator (probably gradient motion estimation) AND a point/feature tracker.



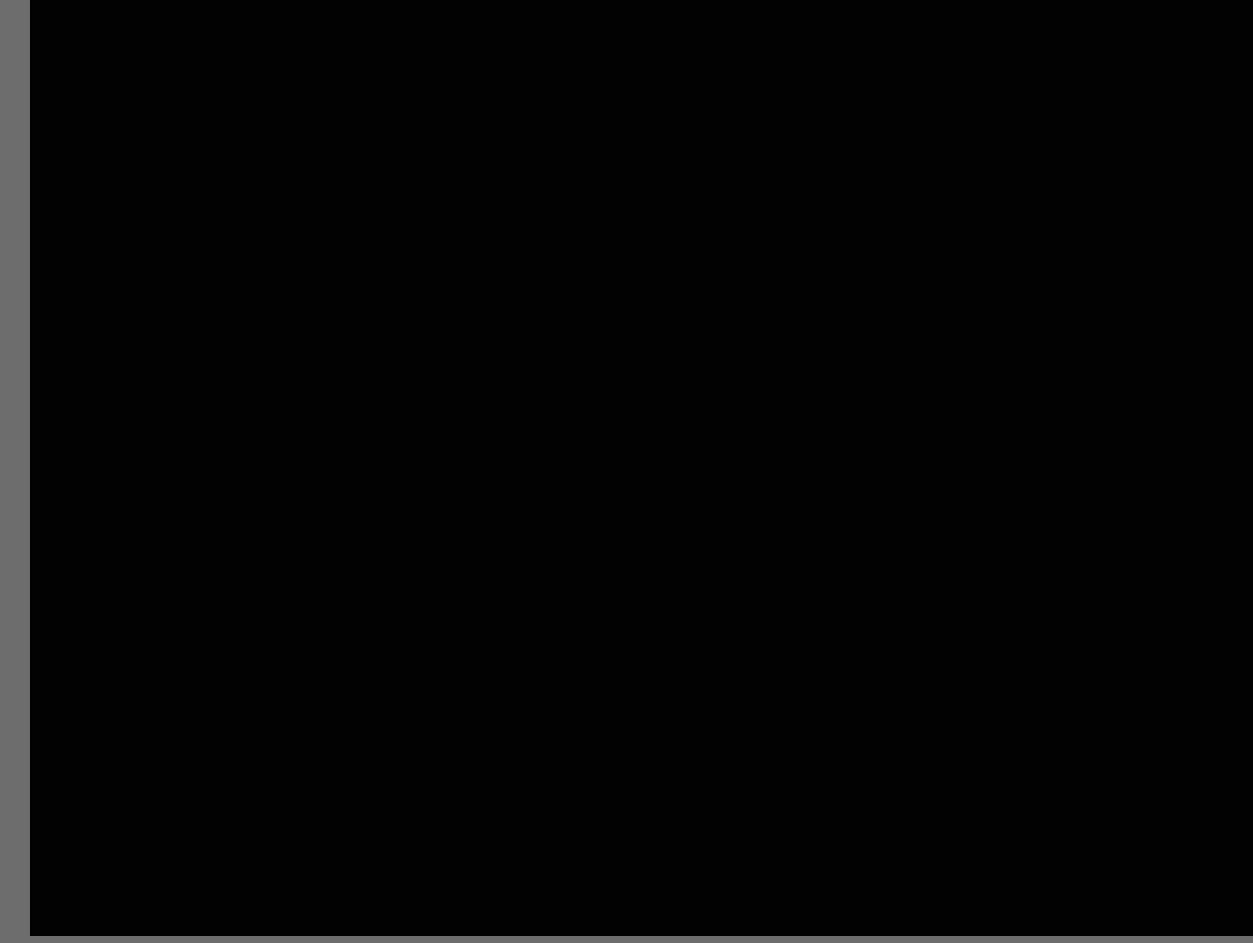
# Subjective Studies

- Designed to measure “perceptual” criteria. Ask people to rate images according to a 5 level “likert scale”.
- Defined first in BT500 standard
- Mean Opinion Score (MOS) measures the average level of perception of some feature
- Difference MOS (DMOS) measures the average level of difference perceivable given some property change between images.
- JND : Just Noticeable Difference defined as the minimum level in some property of an image at which that property is just noticeable. JND usually used in comparing differences between images (stimuli)

# Subjective Video Quality Assessment



A



B

Now how would you rate example B relative to example A?

Difference Mean Opinion Score

Mean Opinion Score

No noticeable Difference

Excellent

Perceptible but  
not Annoying

Good

Slightly Annoying

Fair

Annoying

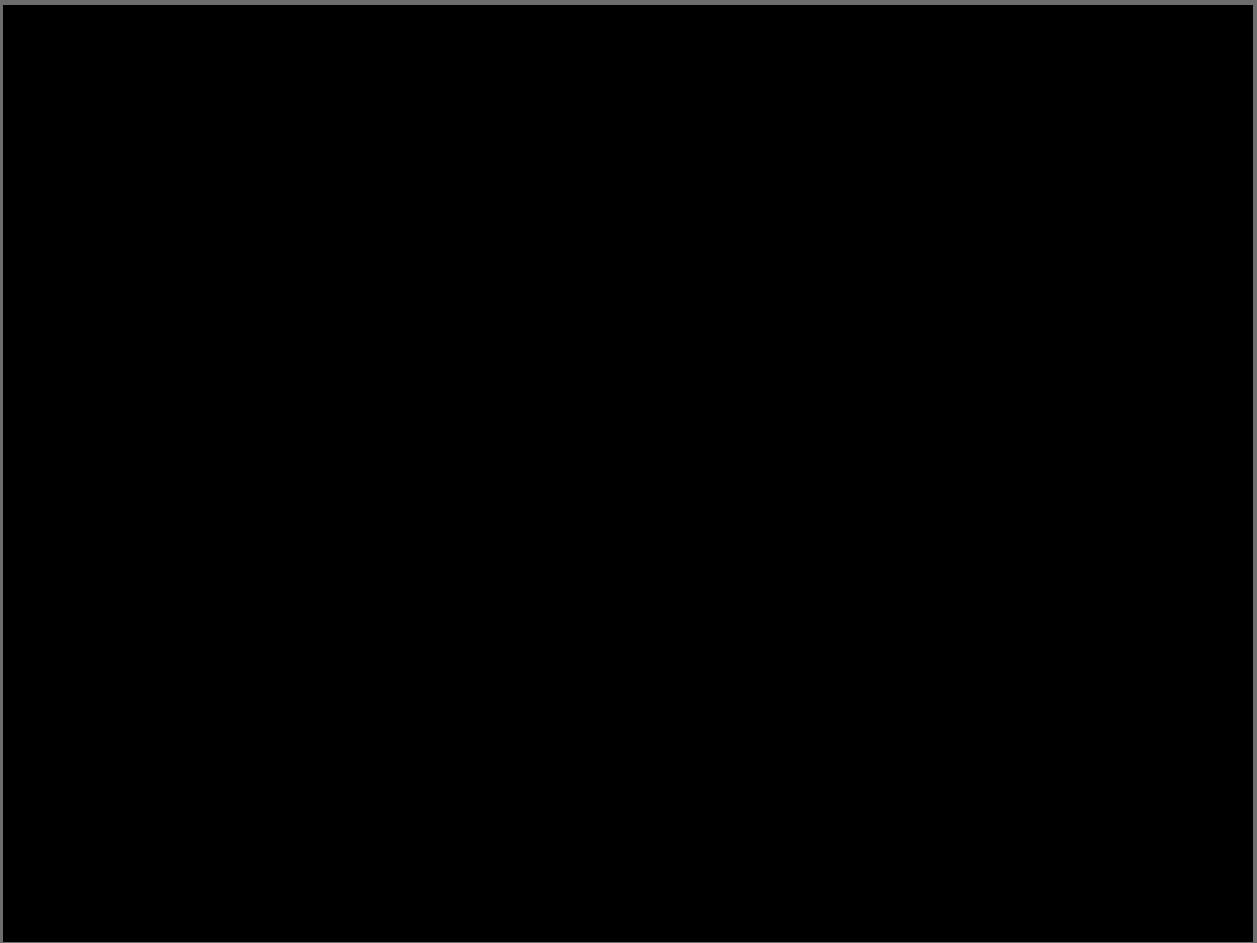
Poor

Extremely Annoying

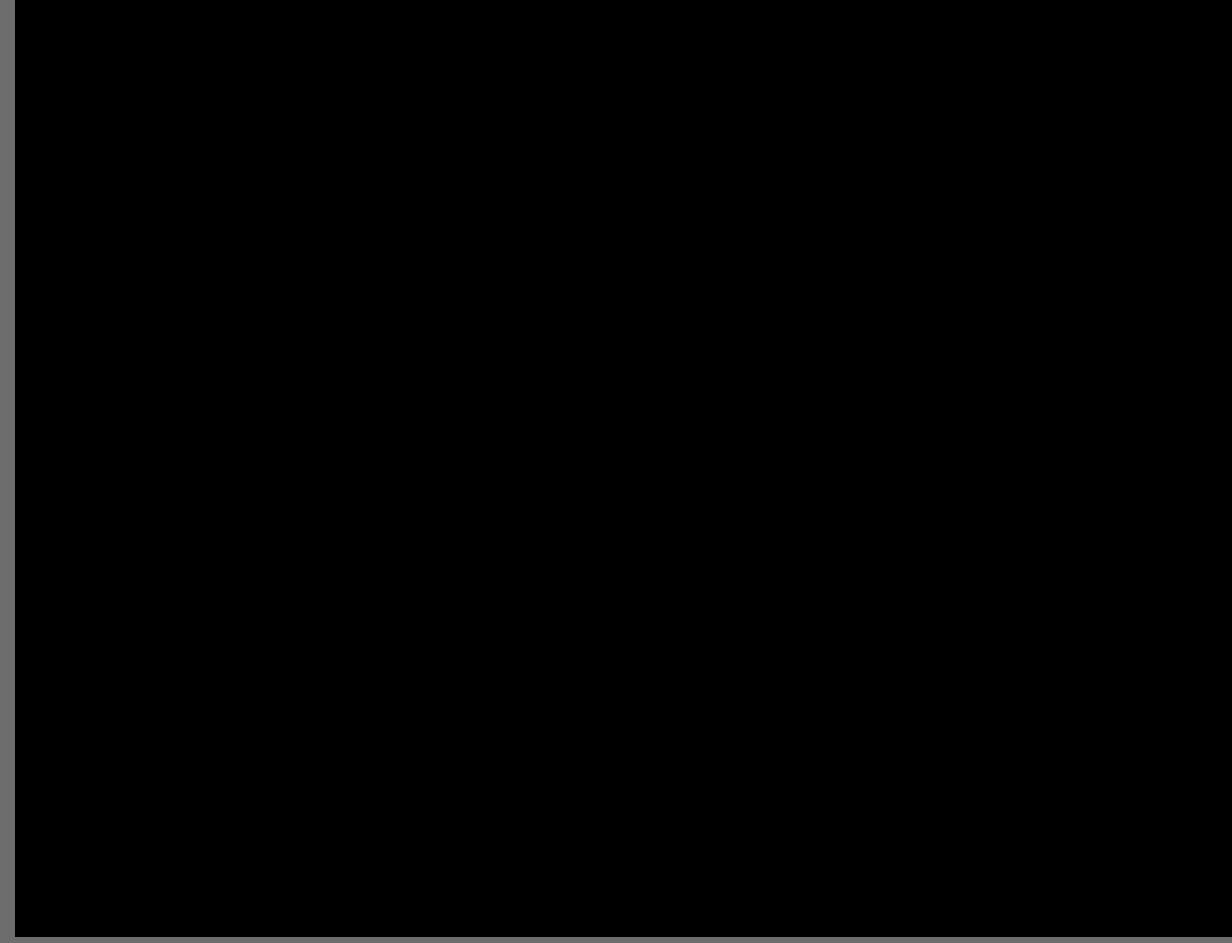
Bad



# Subjective Video Quality Assessment

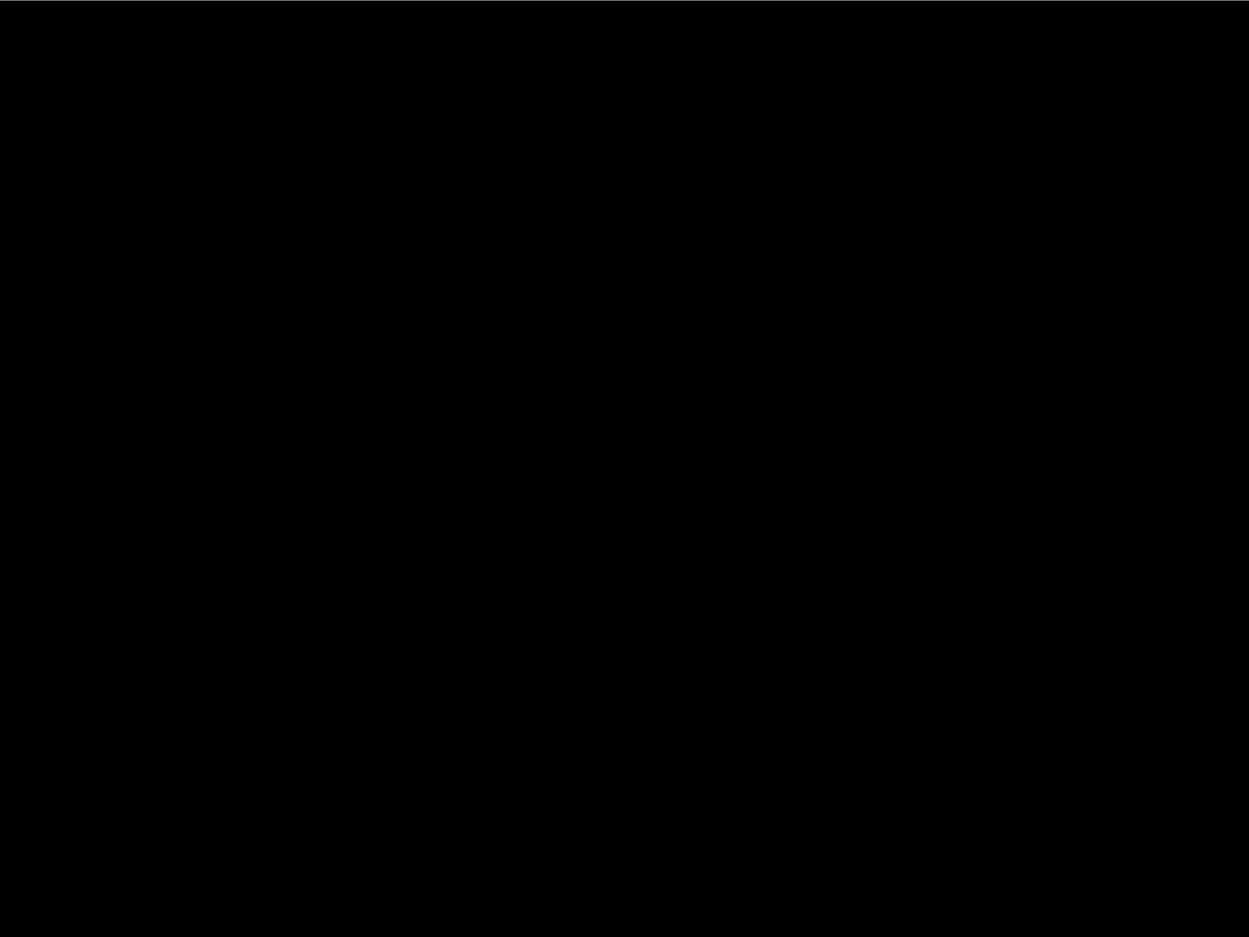


A

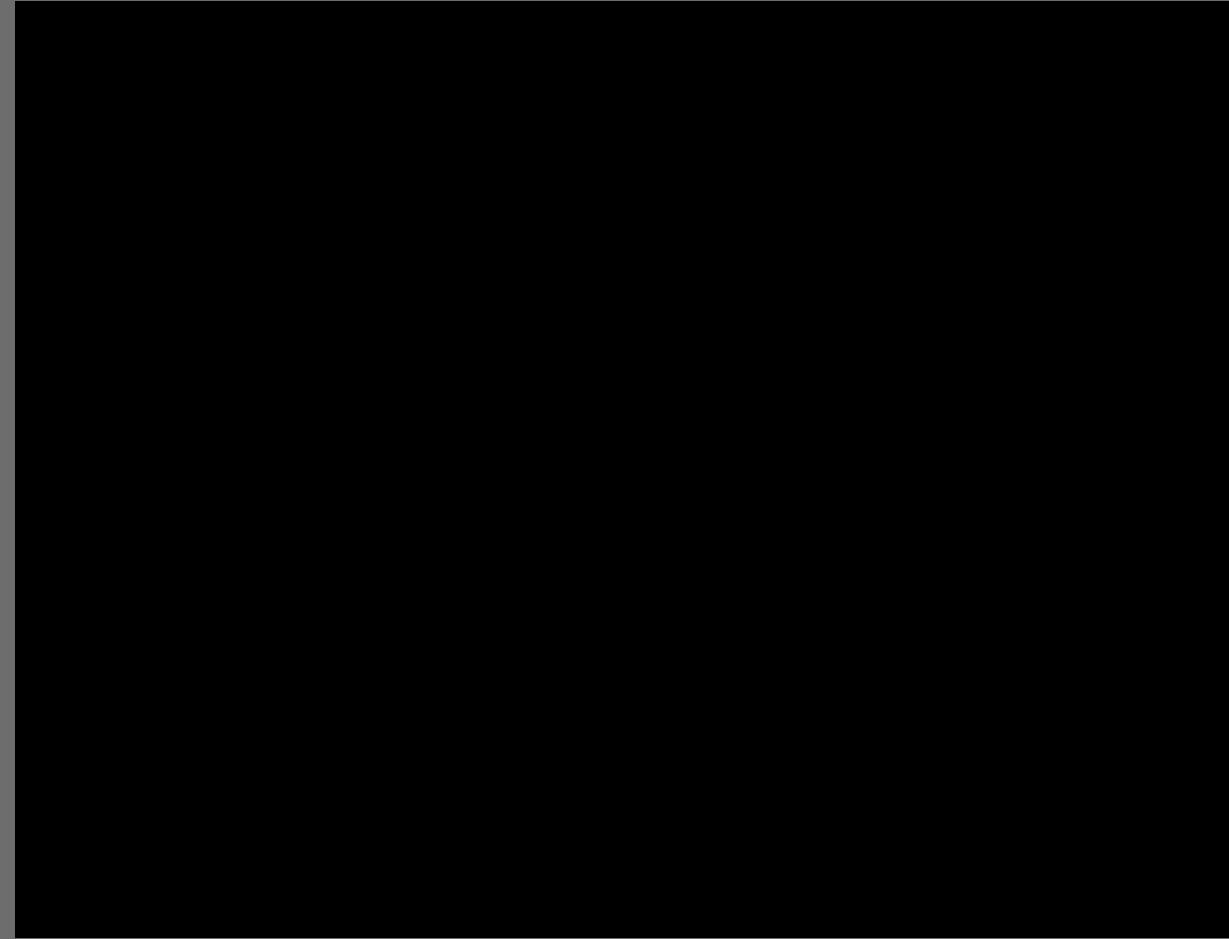


B

# Subjective Video Quality Assessment



A



B

# Collect the data and analyse

	<b>Example 1 (CRF = 50)</b>	<b>Example 3 (CRF = 35)</b>	<b>Example 2 (CRF = 25)</b>
PSNR	26.4 dB	34.13 dB	38.74 dB
SSIM	0.8	0.876	0.94
DMOS	1.34	4.17	4.47

Usually data collected from > 20 participants in a study conducted in controlled lighting conditions

Confidence limits of means are important

Result is a curve of DMOS versus algorithm parameters which allows tradeoffs to be made quantitatively in terms of visual quality

# Subjective Tests and Predicting MOS

- VQEG Video Quality Experts Group (since 1997)
- ITU Recommendation 2004 (BT500)
- Double stimulus continuous quality scale DSCQS
- Viewing distance etc defined.



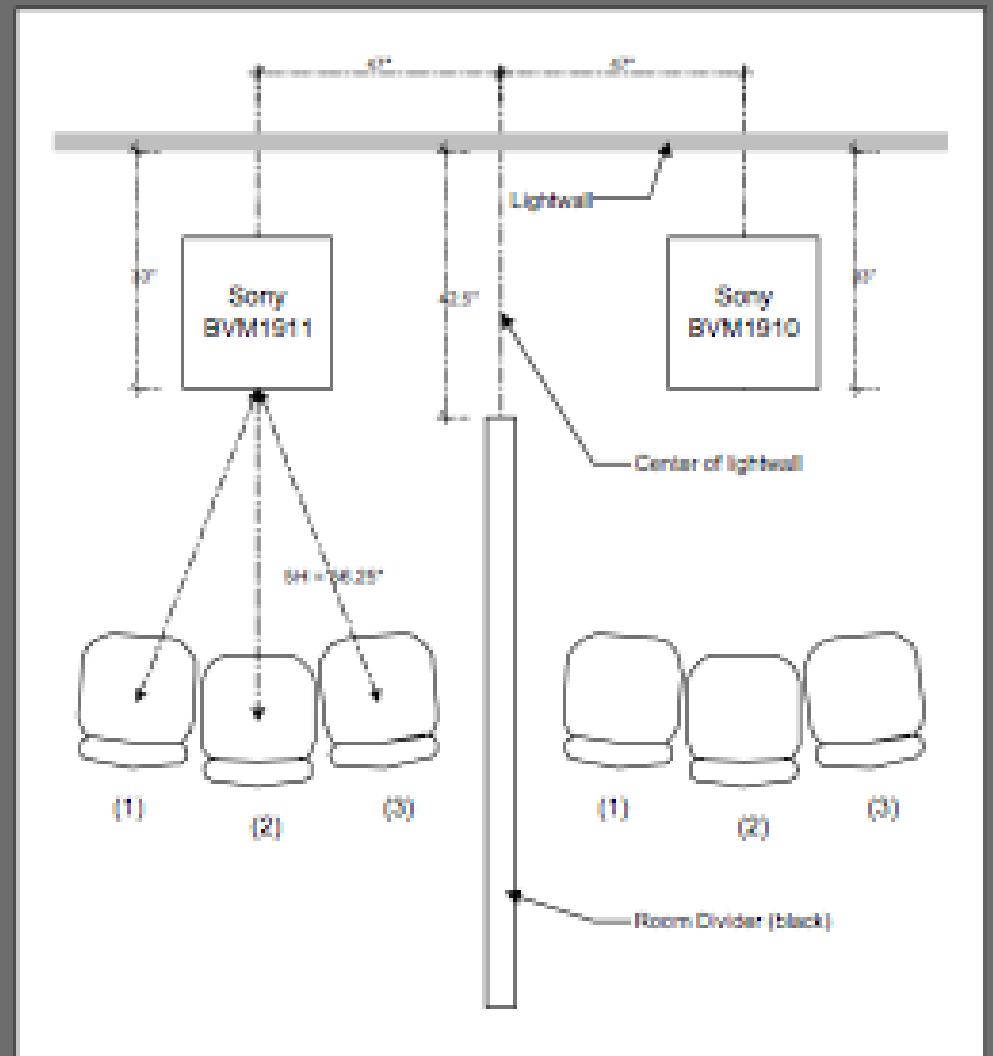
Excellent

Good

Fair

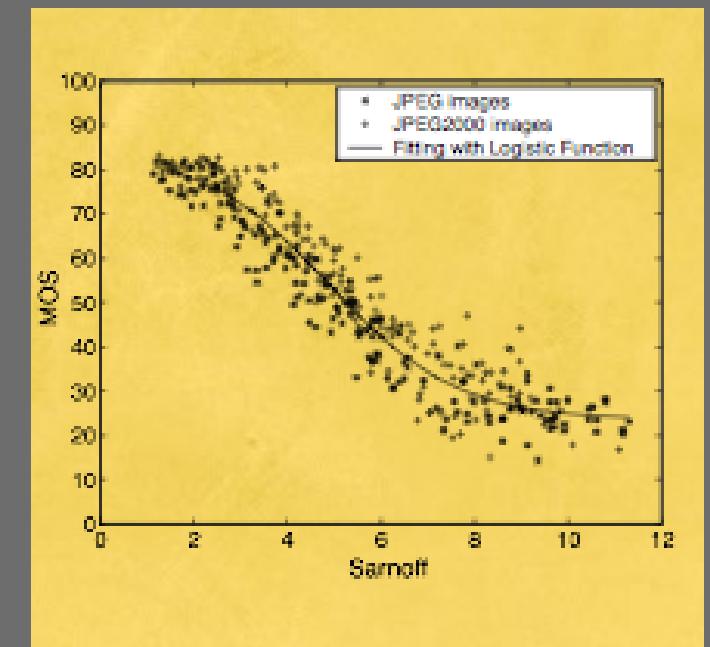
Poor

Bad



# VQEG: Video Quality Experts Group

- BT500 : Defines environment/method for testing
- Defines picture quality tests i.e. picture sets
- Defines tests of “goodness of the metric”
  1. Regression line fit to the data vs MOS
  2. Correlation coefficients between objective/subjective scores, spearman rank and outlier ratios



# Structural Similarity Metric

(Reference Metric from Bovik et al at Austin Texas 2004)

- Motivated by all various psychophysical measurements shown here (except motion)
- Weber's law and Texture/Edge masking in particular
- Defined a REFERENCE metric that aligns better with MOS than PSNR

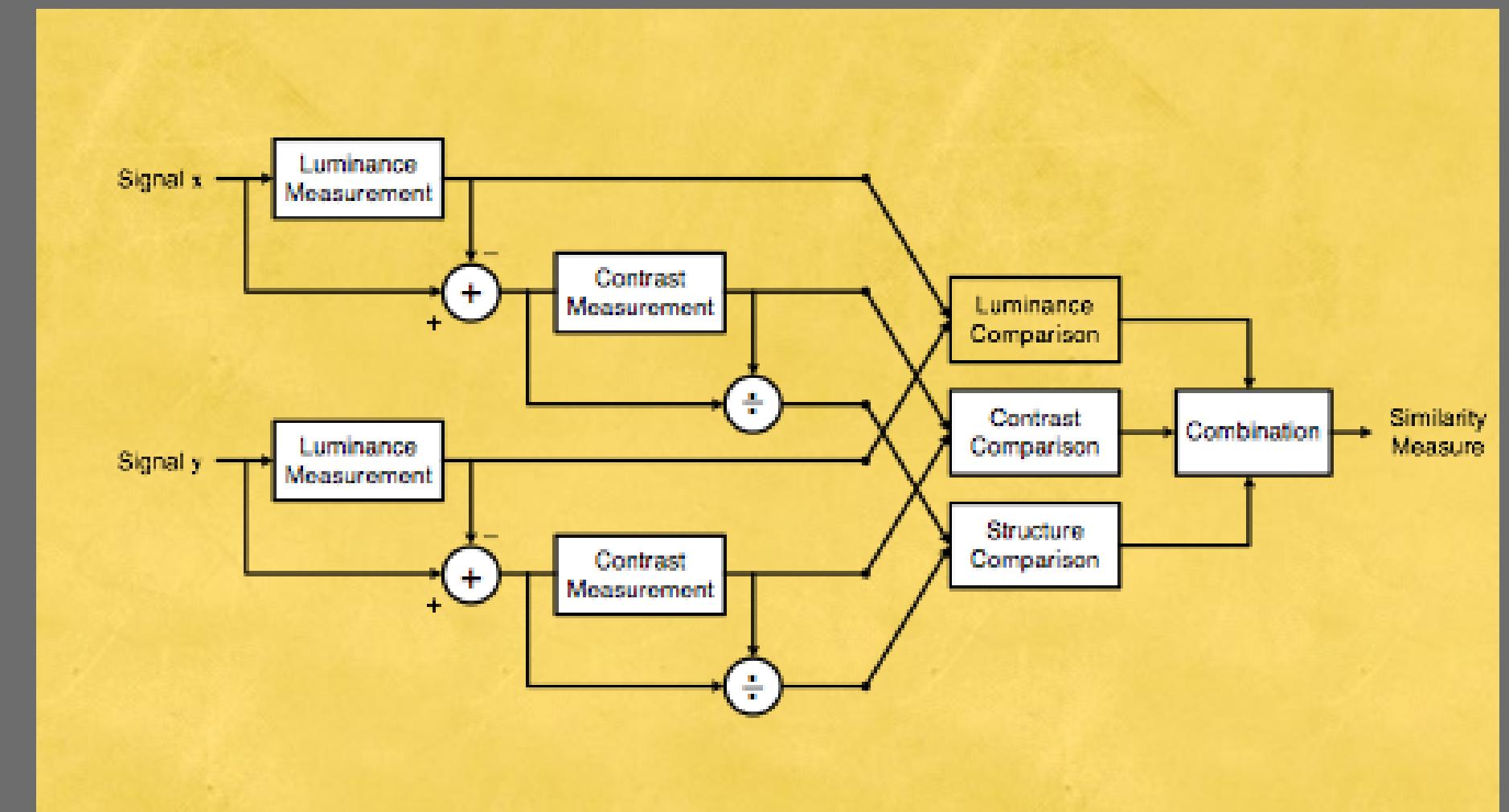


Image Quality Assessment: From Error Visibility to, Structural Similarity, Zhou Wang, Alan Conrad Bovik, Hamid, Rahim Sheikh, and, Eero P. Simoncelli.

# Some details

$$l(a,b) = \frac{2\mu_a\mu_b + K_1}{\mu_a^2 + \mu_b^2 + K_1}$$

Luminance visibility depends  
on ratio between objects

$$\varsigma(a,b) = \frac{2\sigma_a\sigma_b + K_2}{\sigma_a^2 + \sigma_b^2 + K_2}$$

Contrast operates on variance of patches and is  
less sensitive when there is a lot of local variance  
.. which is in keeping with CSF masking

$$s(a,b) = \frac{2\sigma_{ab} + K_3}{\sigma_a\sigma_b + K_3}$$

Structure is a variance measure after luma and  
contrast differences are compensated

$$SSIM(a,b) = l(a,b)^\alpha c(a,b)^\beta s(a,b)^\gamma$$

# SSIM Maps



(a)



(b)



(c)



(d)



(e)



(f)



(g)



(h)



(i)

# SSIM Comparison



PSNR 28.3  
SSIM 0.78



PSNR 27.5  
SSIM 0.99

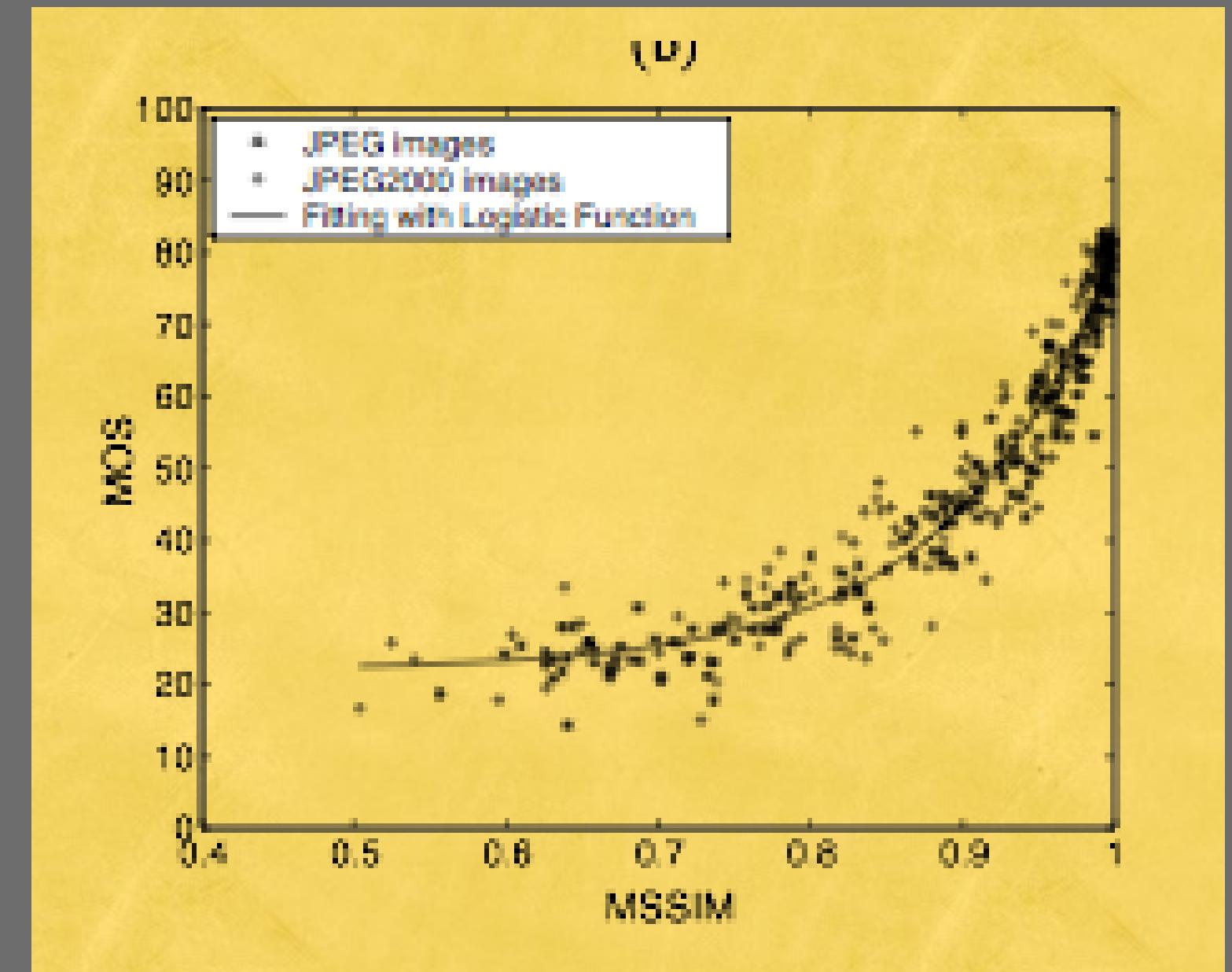


PSNR 27.5  
SSIM 0.53

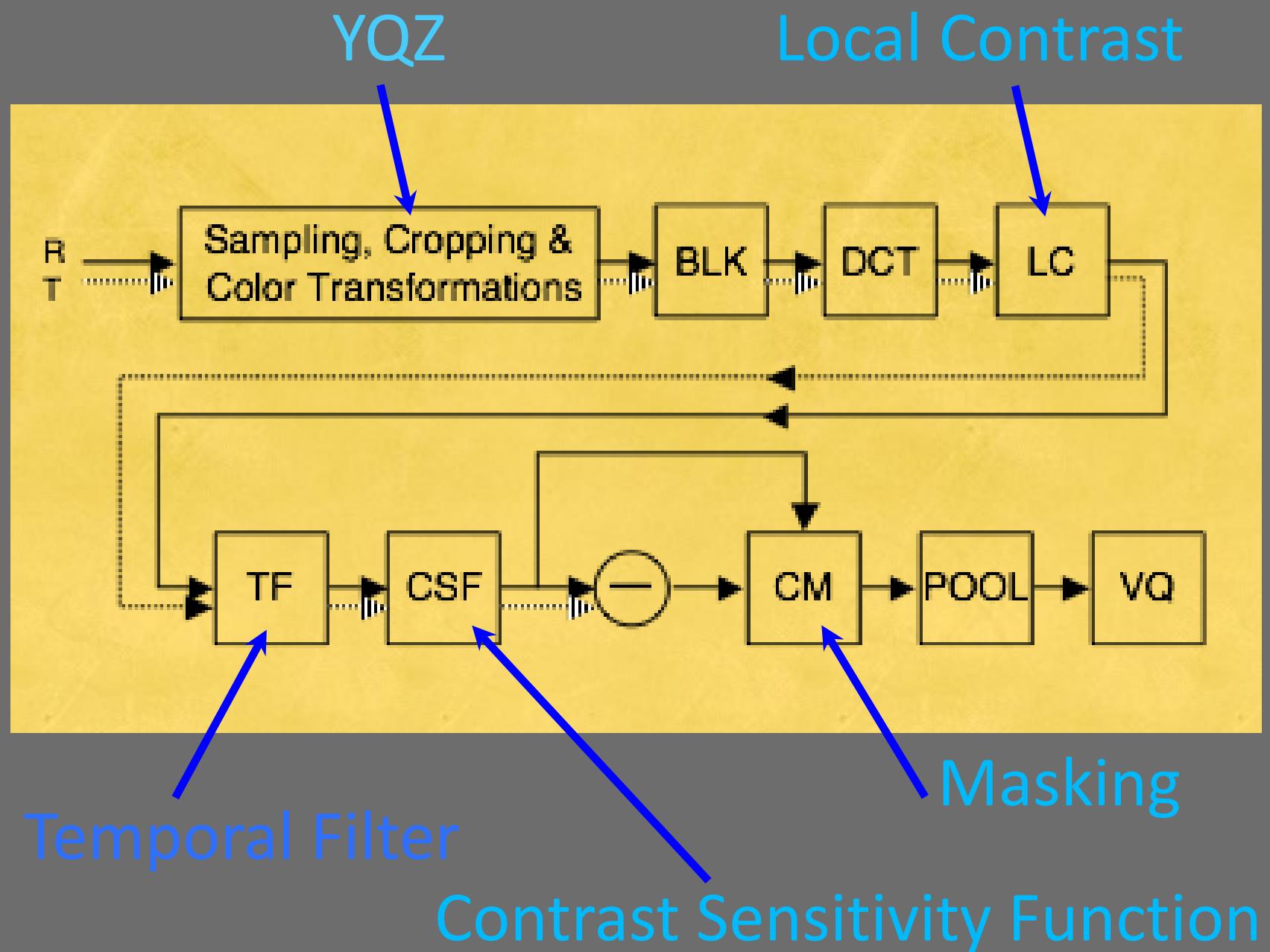


# Predicting MOS

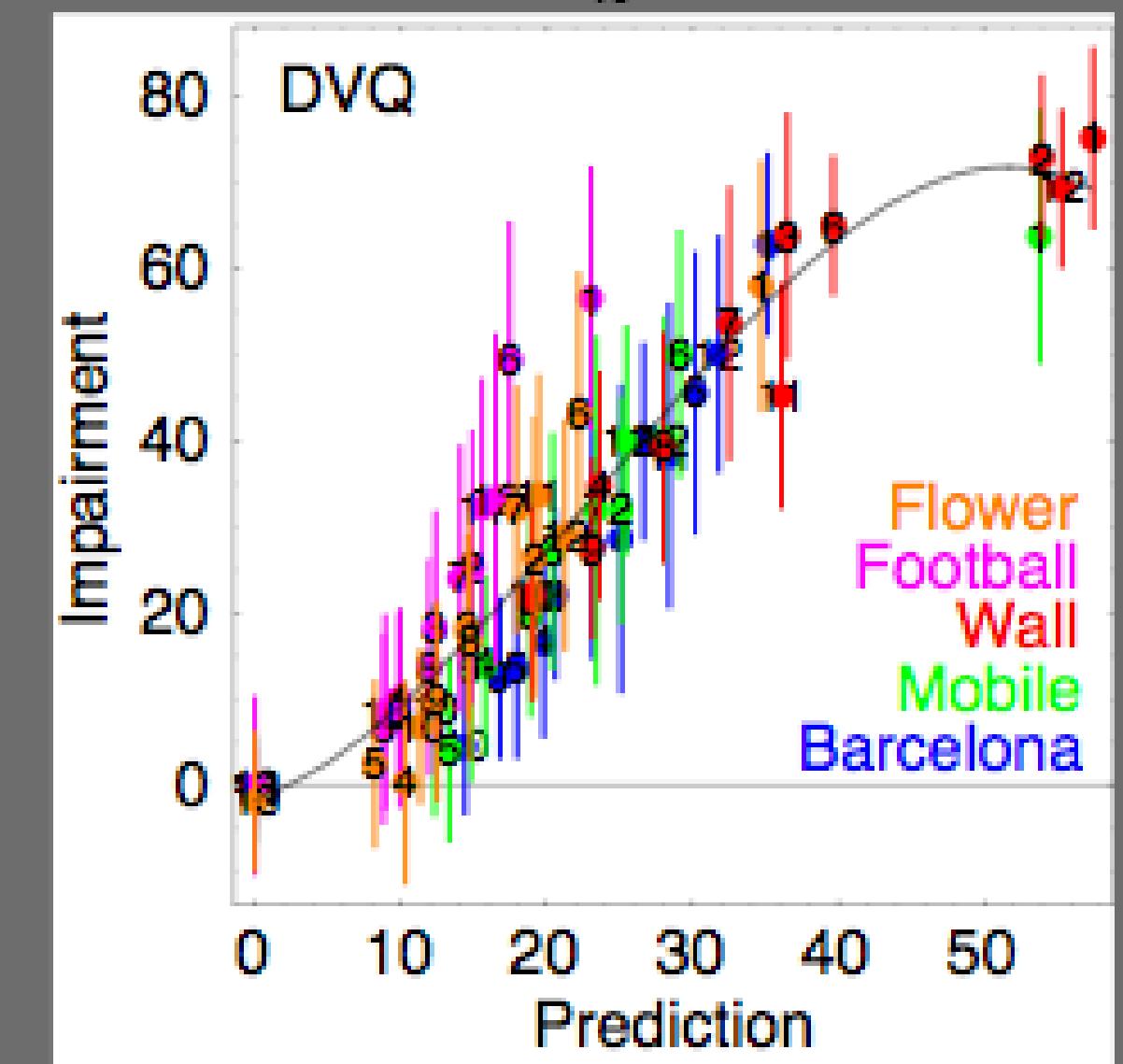
- A complaint about the SSIM is that it does not vary much numerically, and we do not know what a good number is
- We know 40dB PSNR in coding is good .. but not what 0.998 means in SSIM
- In their original paper they actually provide a mapping to MOS.



# Temporal Video Quality Metric



4th order polynomial fit



# VMAF (Video Multimethod assessment Fusion)

- VMAF is a reference metric developed at USC sponsored by Netflix
- It varies between 0 – 100
- Now becoming an industry standard : implemented in ffmpeg too
- No explicit motion information is used for giving temporal feature
- Deep Learning methods for modelling MOS from pictures hampered by lack of large amounts of reliable data

# But wait .... what about...?

Do the output pictures look better than the input?



Do the output pictures look good?

# No reference assessment

- Even more affected by viewing conditions
- Need to investigate fundamental limits on visibility of artefacts
- Experimental design seems to point to earlier work on RT (reaction time) experiments

# Specifically

- Noise/Blocking/Banding visibility seems something we can achieve without a reference

# An example : Measuring detail

$$Q = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2}$$

Milanfar's Q-Metric  
(implementation modified a little)



$\lambda_1, \lambda_2$  are the eigenvalues of the  $G'G$  matrix that was introduced in the motion estimation lectures. They align with max and min values of contrast in the block. The red vectors in this picture show the eigenvectors scaled by the eigenvalues. These show both the amount and direction of max/min contrast in a block.  $16 \times 16$  blocks used here.

# No-reference detail



$Q = 24$



$Q = 7.4$

# The future

Combining metrics to predict MOS scores through some non-linearity seems sensible

Colour perception not well incorporated

There is a place for traditional modelling still

Visual perception knowledge more advanced than we might think.

Recent work on DNNs for perception is promising

Colour at Weta

Peter Hillman (Oscar Winner 2014)

Thursday 25<sup>th</sup> March 2021

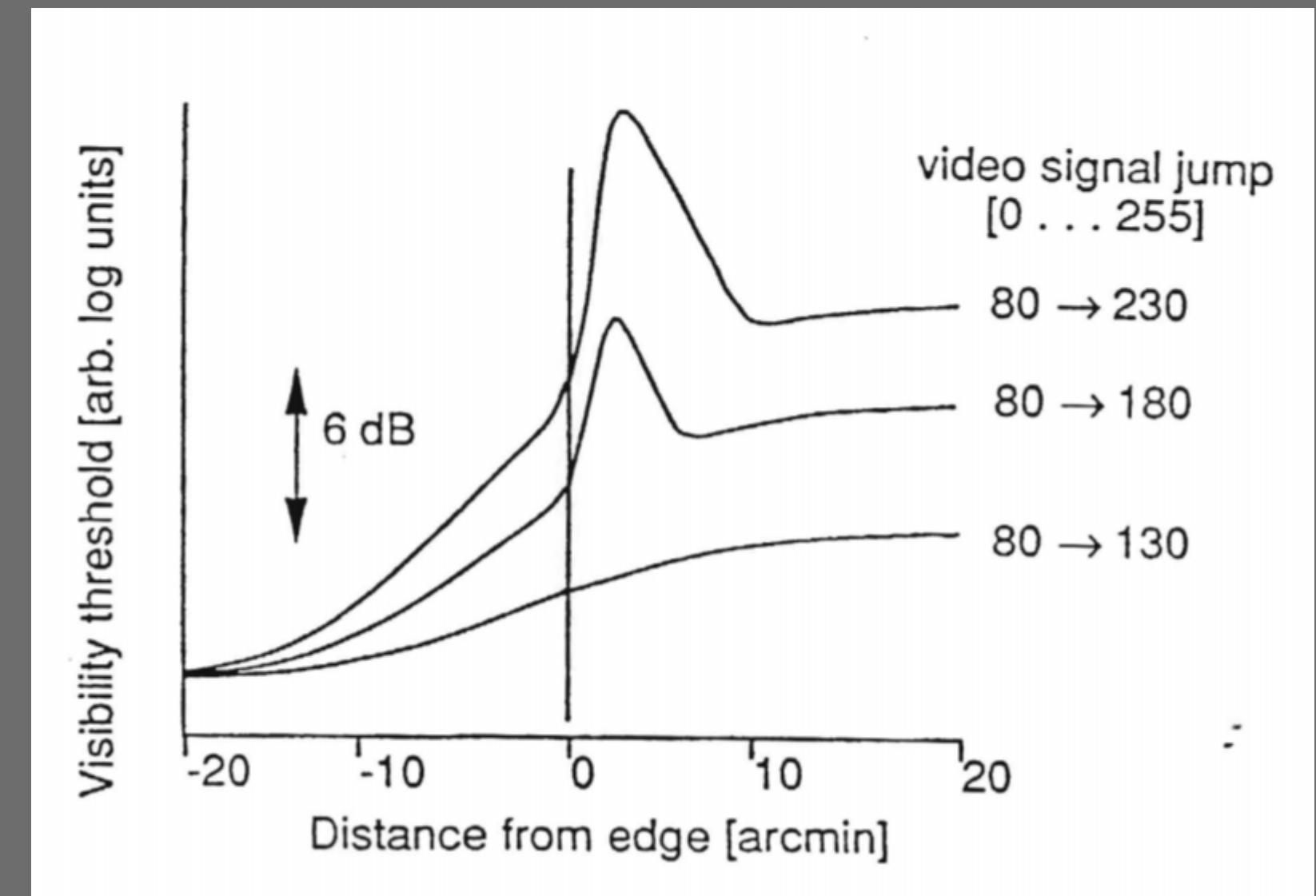
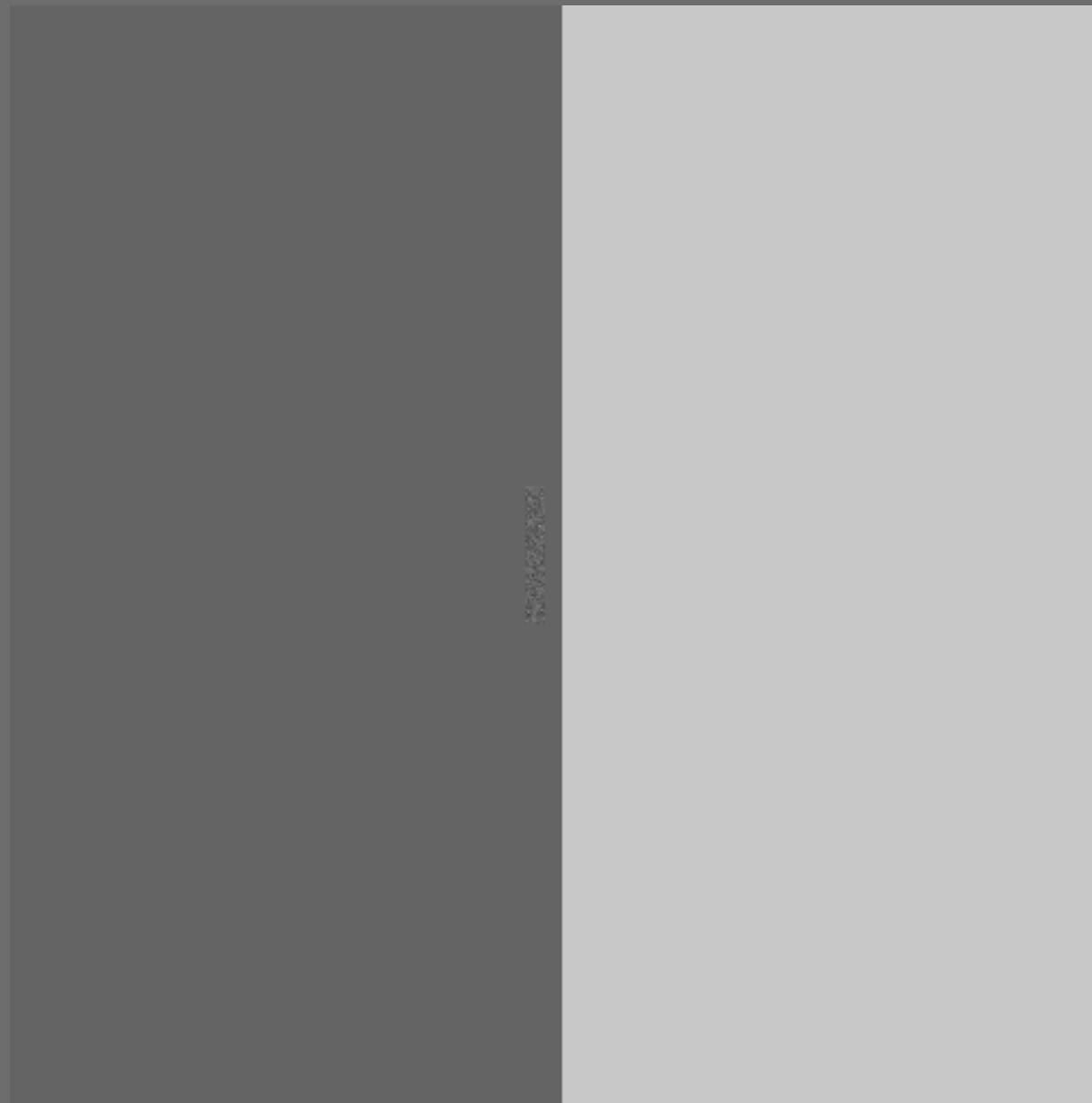
8pm

## FILMS+

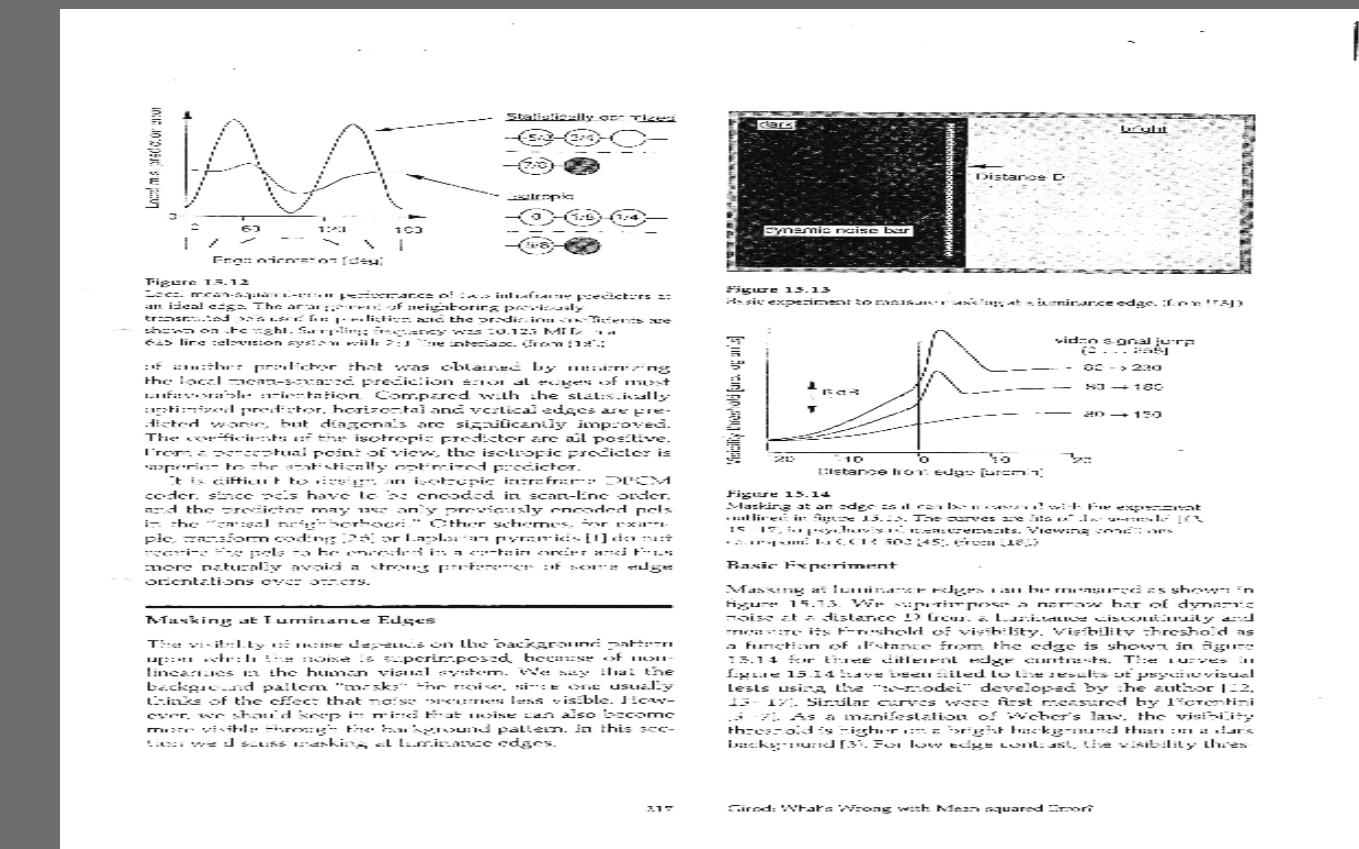
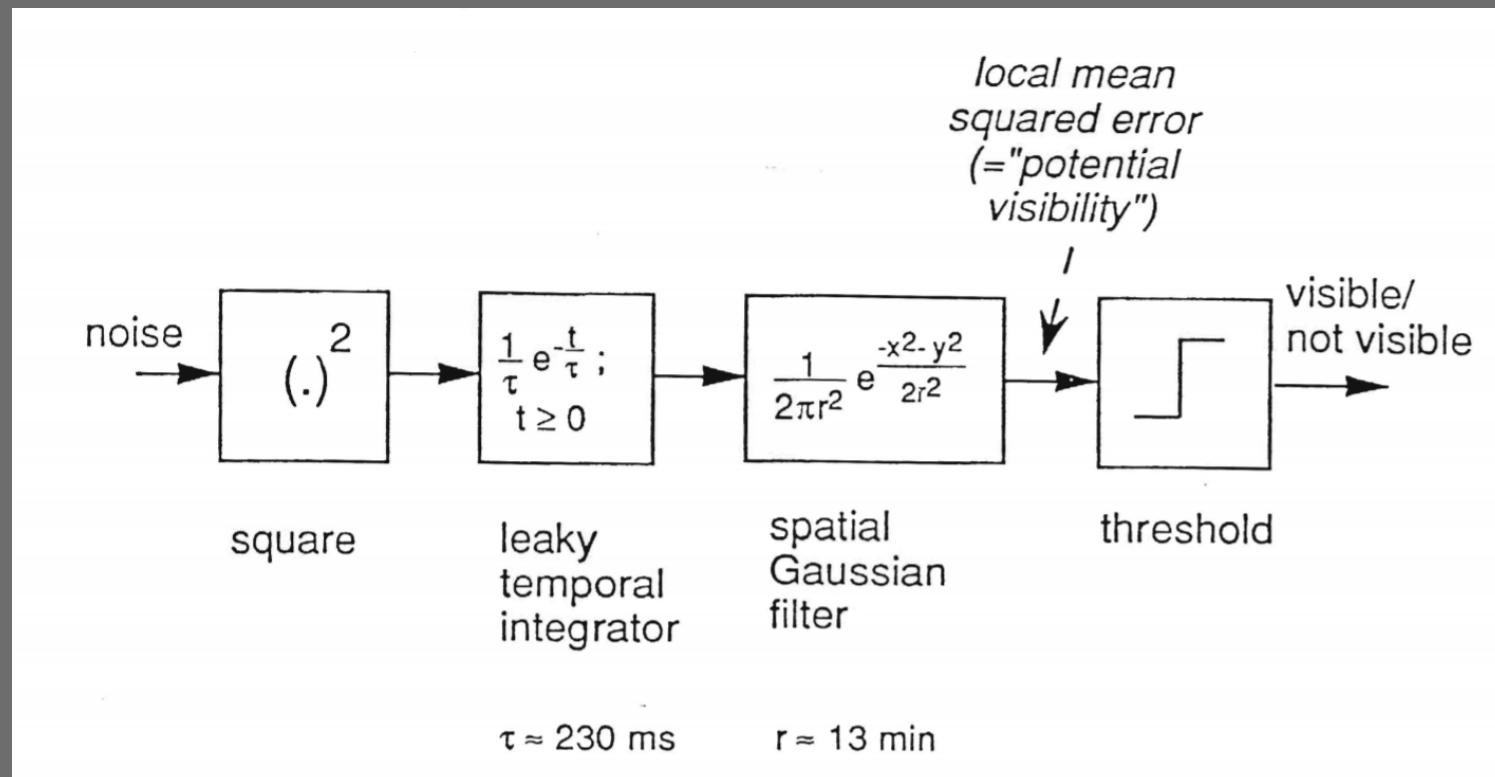
Creating some of the most  
memorable visual effects for  
over twenty years.

*Interlude on earlier work*

# Early work in perceptual quality prediction



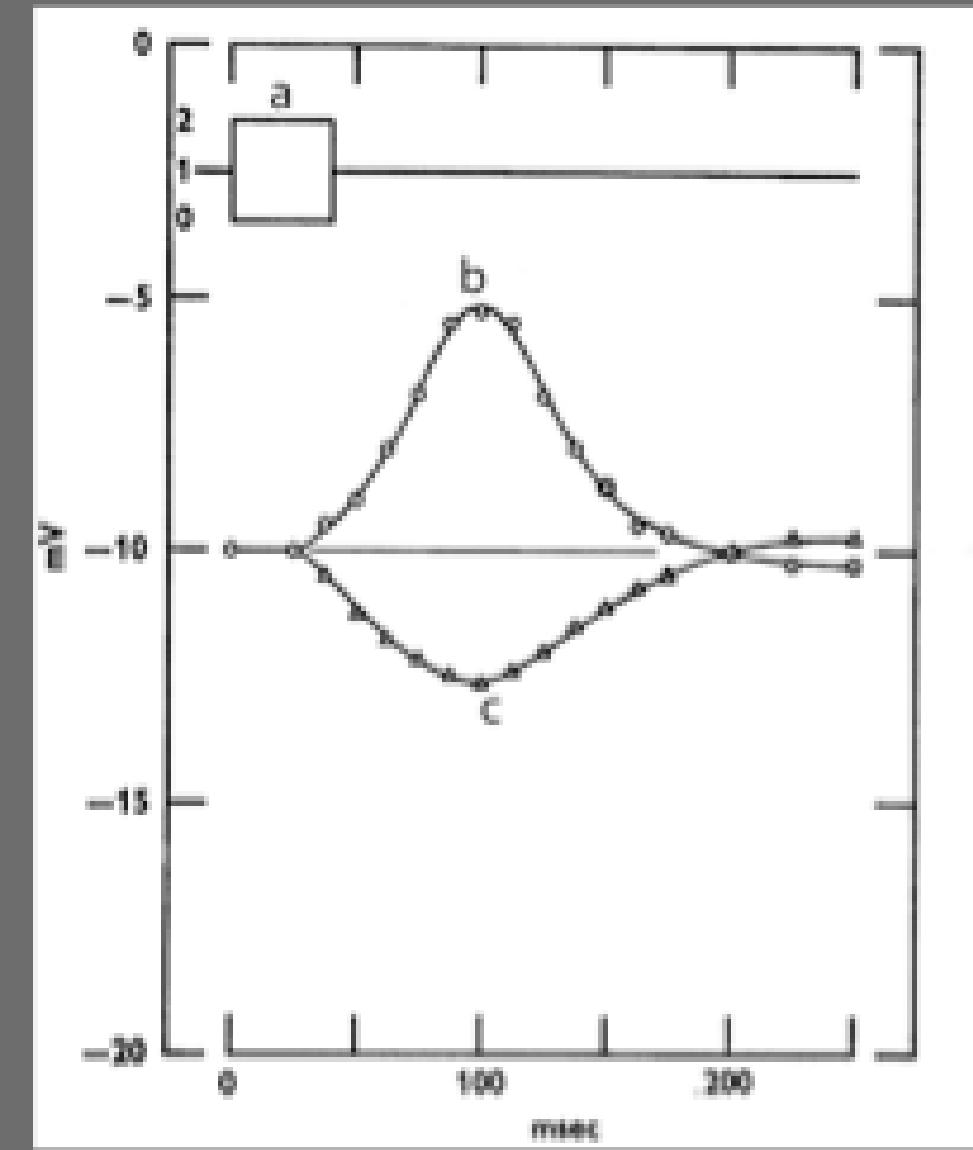
# Model



Use the visibility to control quantisation in a predictive coder

# Incidentally in 2005 ...

- People who design codecs know that you can hide more errors on the dark side of an edge than the bright side.
- Spirling et al, at Irvine CA and Ohio State, did a series of experiments to model the observation.
- Turns out that cones respond differently to turning the lights down than turning the lights up.

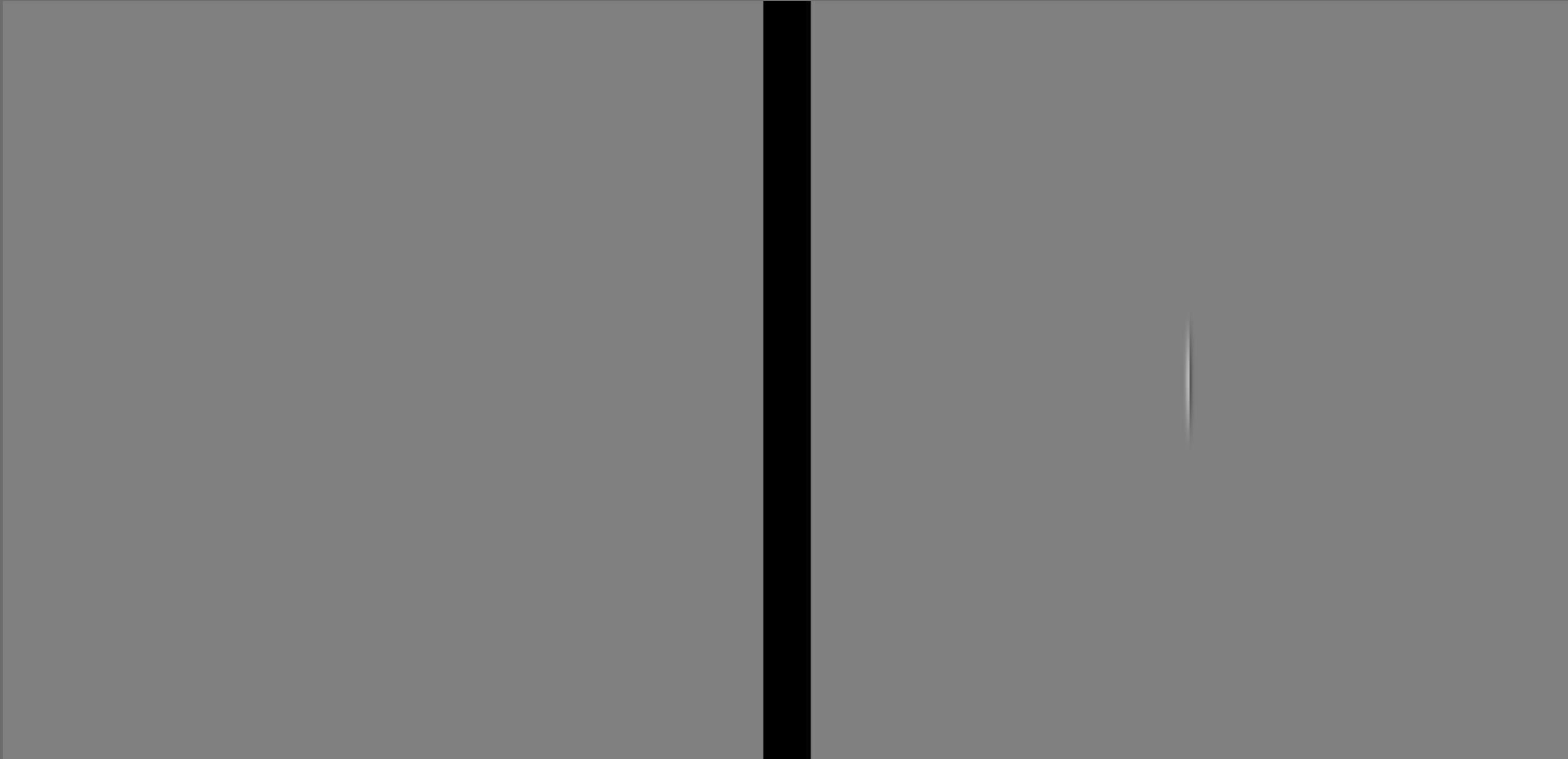


Experiment on turtle Cones. Light on/off for 40 ms and then measured mV.

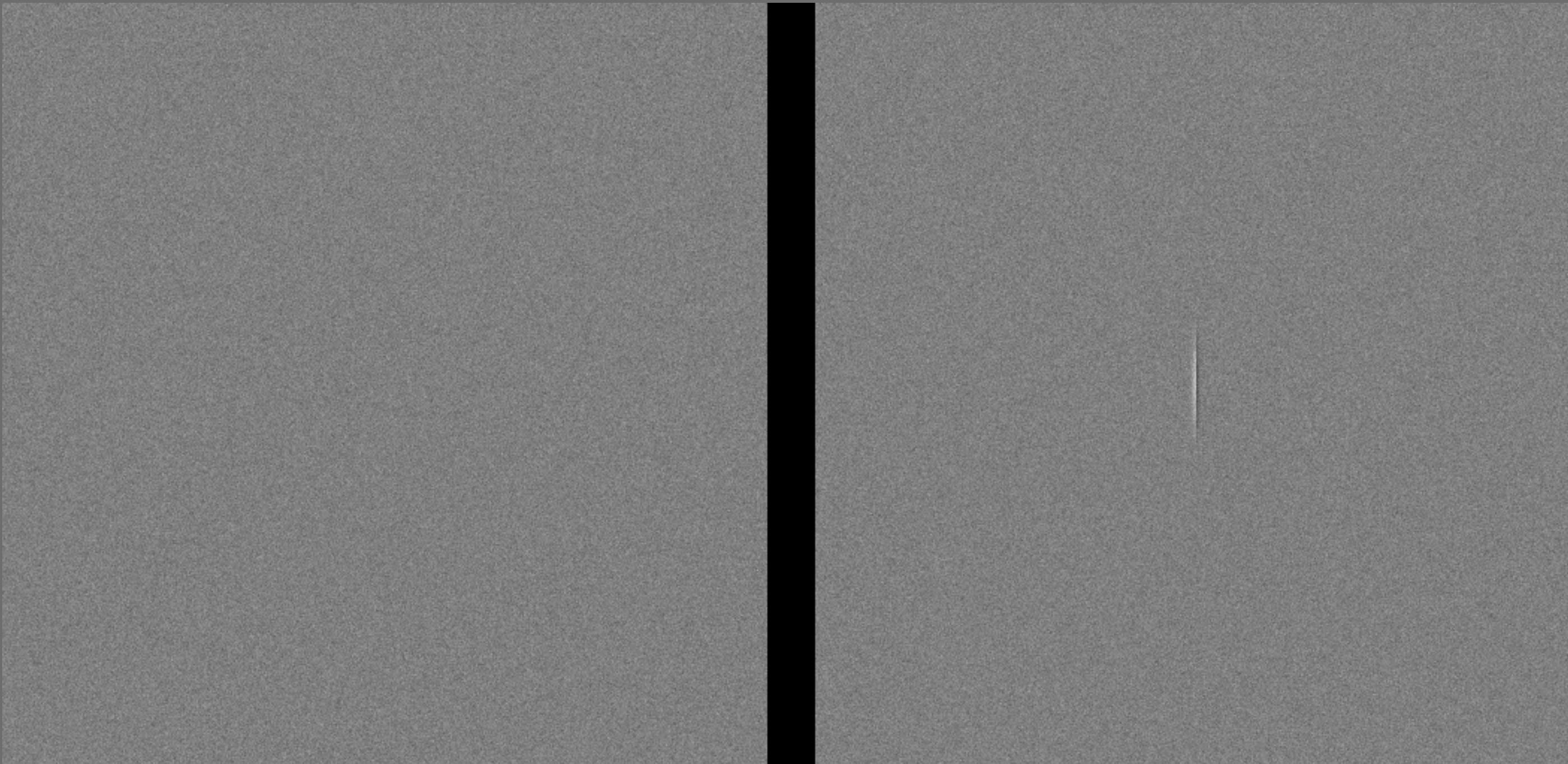
# In 90's some experimented with direct assessment of artefacts

- Model directly the type of artefact you are interested in and then use it to predict the picture quality.
- Use stylistic edge stimulus.
- Get away from “threshold measurements” and use “reaction time” measurements for supra-threshold behaviour.

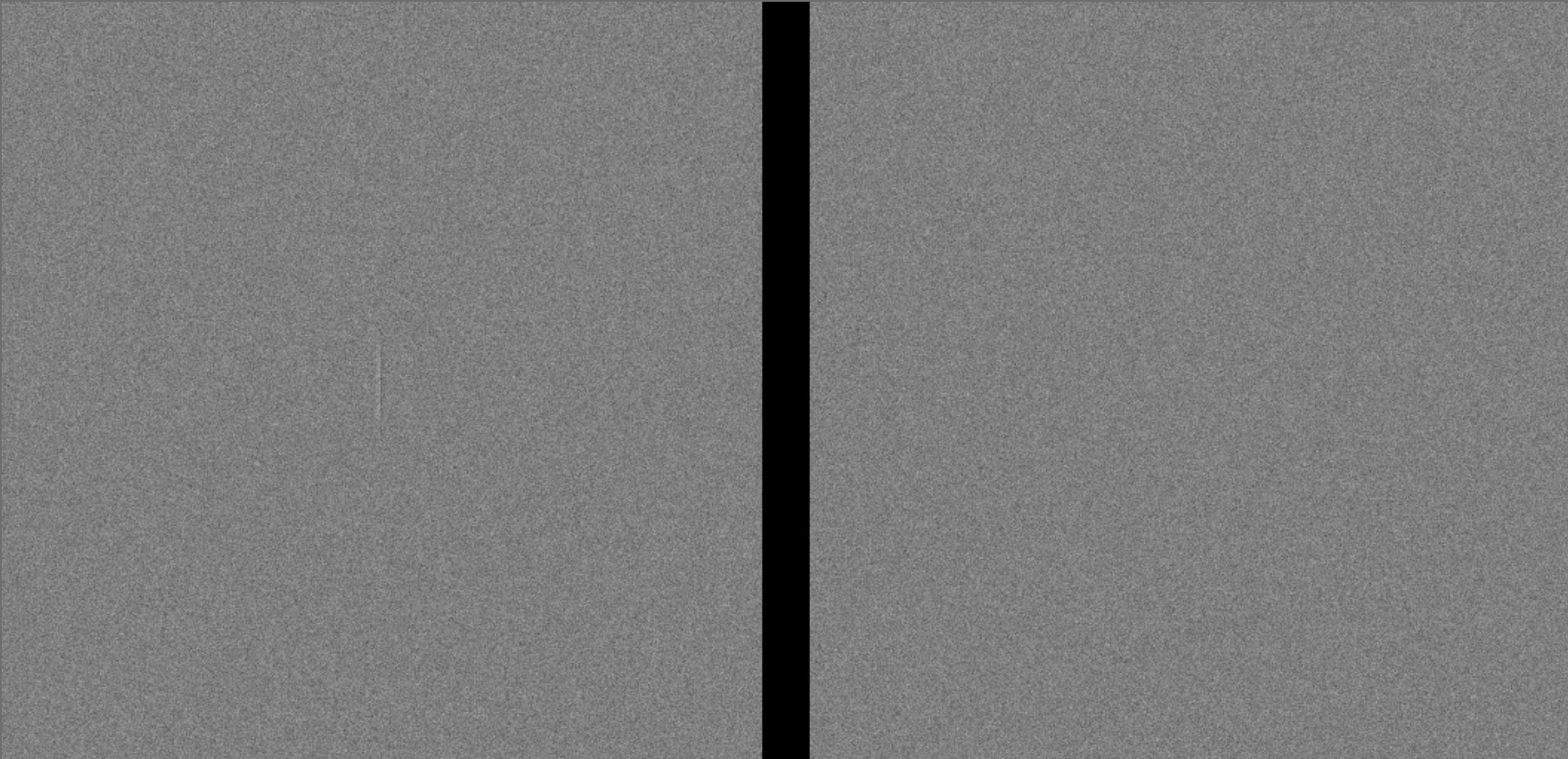
# Reaction time experiments



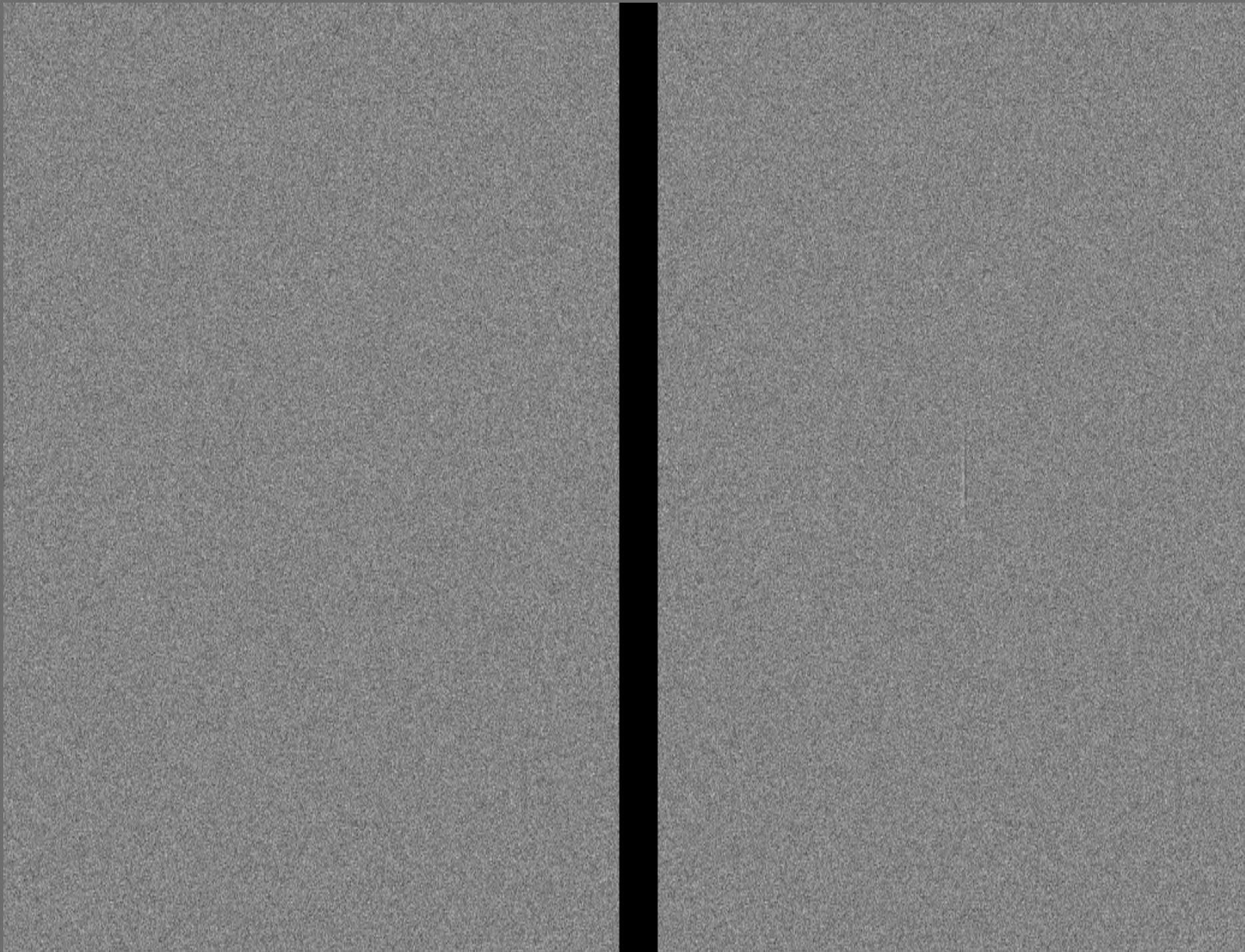
# Edge Masking with noise



Which side? (10 sec timeout)



# Motion perception/masking



# Reaction time results

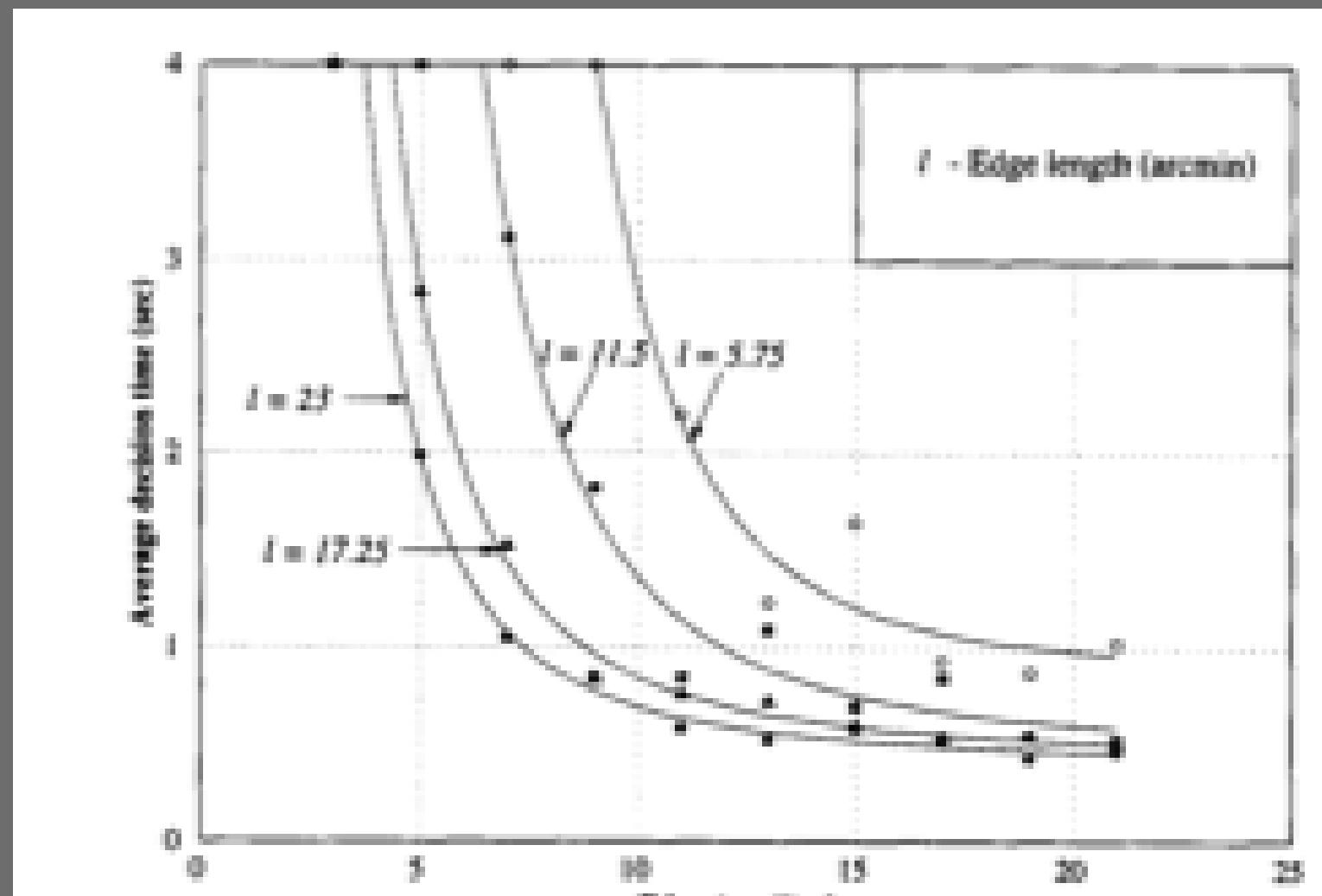
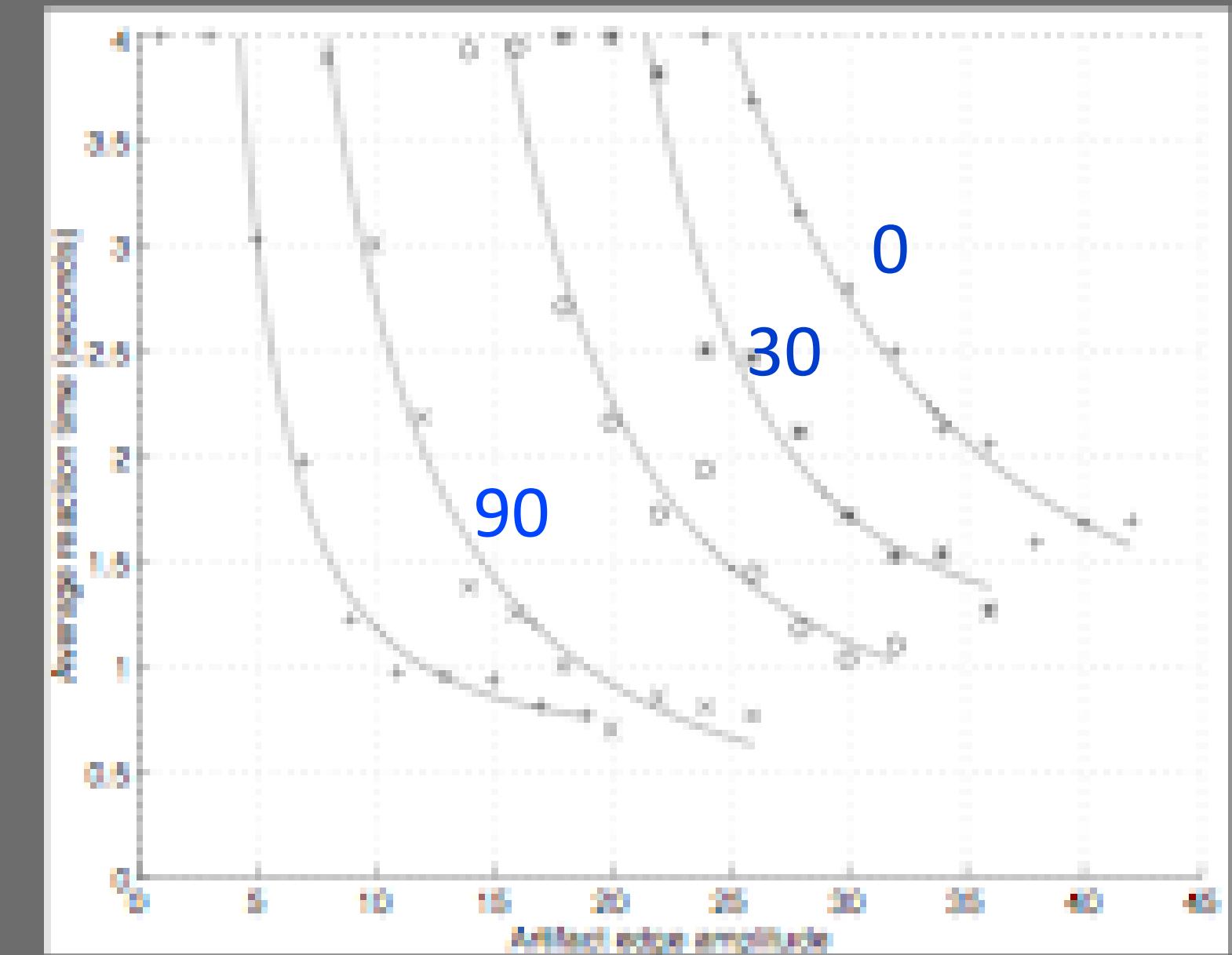
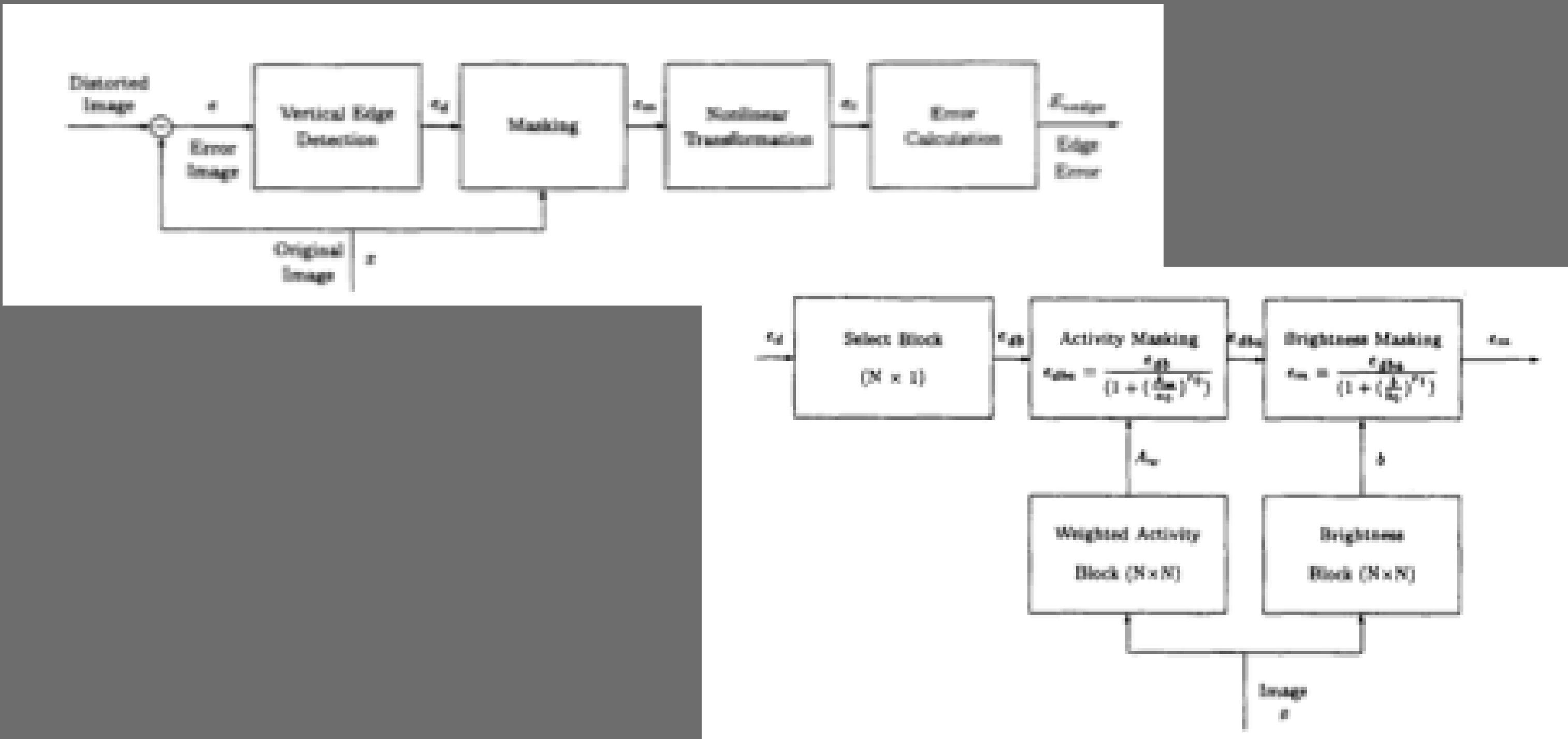


Fig. 6. Decision time versus edge amplitude for constant edge length.



RT versus edge amplitude in noise

# Models: Reference Metrics



# Edge Error Quality Prediction

- Classic 1990's image processing approach
- Lena image corrupted by DCT artefacts to create 12 pictures
- Ranked by a few people, and then compare the ranking to the prediction.
- Errors chosen to have similar MSE but different visibility

Picture	MSE	E <sub>visibility</sub>	Subjective Ranking								
			1	2	3	4	5	6	7	8	9
1	1.50	0.19	1	1	1	1	1	1	1	1	1
2	1.50	0.33	2	3	3	3	3	3	3	2	4
3	1.50	0.35	3	2	4	2	4	4	4	3	3
4	1.51	0.68	5	4	2	4	3	2	7	2	2
5	1.50	0.79	4	5	7	7	8	7	4	5	
6	1.50	0.80	6	6	5	5	5	6	5	5	7
7	1.51	0.82	7	7	6	6	6	5	6	6	6
8	1.50	0.90	8	8	8	8	7	8	8	8	8
9	1.50	0.93	9	9	9	9	9	9	9	9	9
10	1.50	0.99	10	10	10	10	10	10	10	10	10
11	1.50	1.08	11	11	11	11	11	11	12	11	11
12	1.51	1.24	12	12	12	12	12	12	11	12	12

*Done Interlude on earlier work*