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.

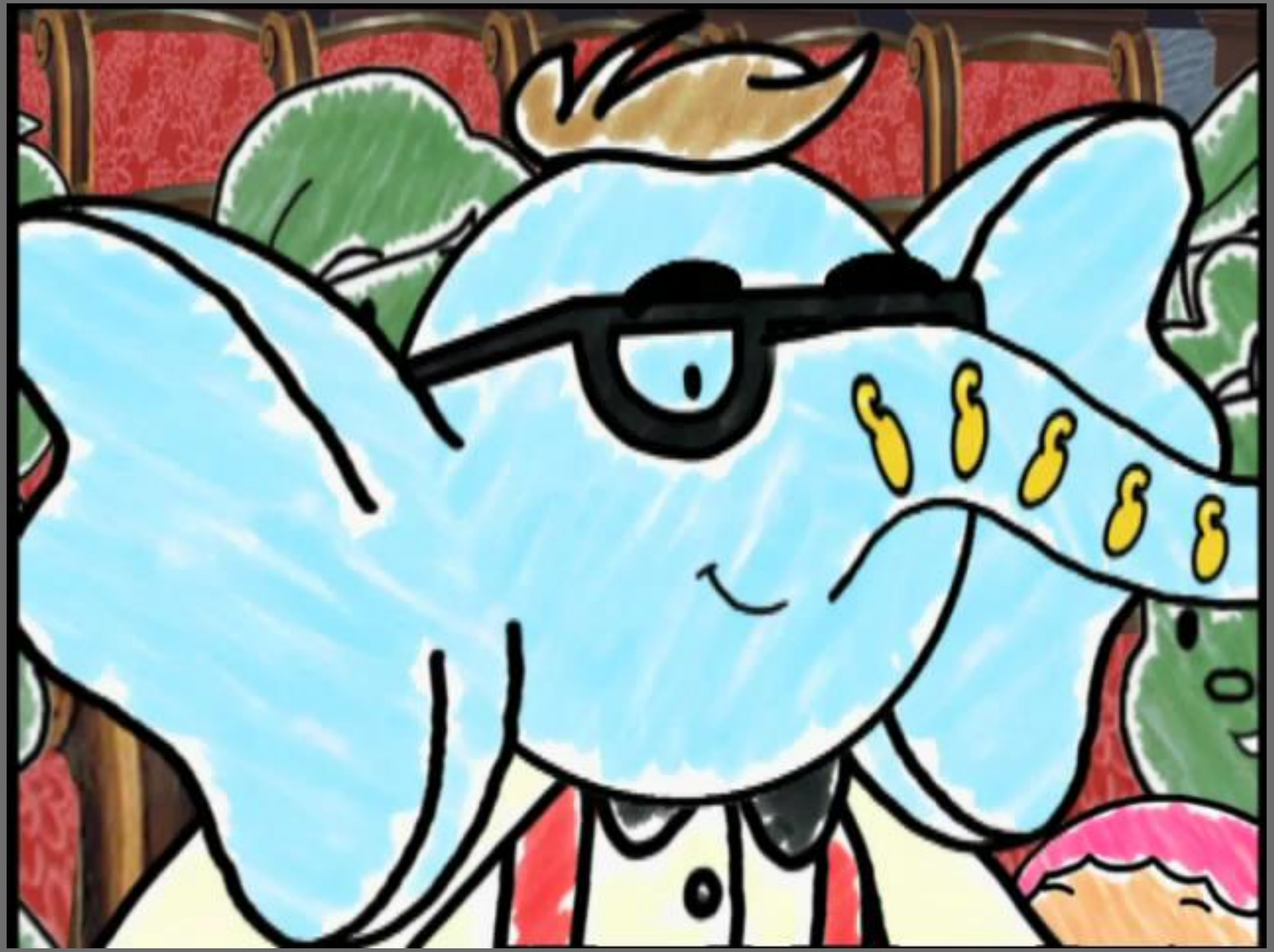
What are our questions?

Do the output pictures look better than the input?

Do the output pictures look like the input?

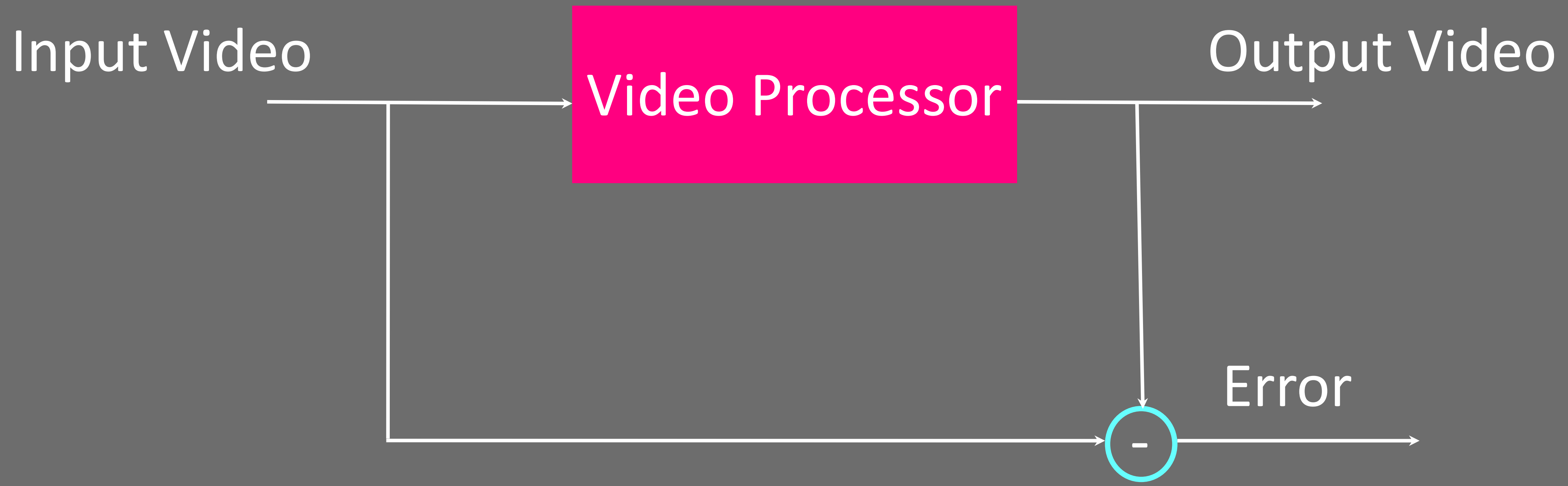
Do the output pictures look good?

Things go wrong




This isn't right

Metrics (reference and no reference)



A reference metric answers the question “Do the output pictures look like the input?”
A no-reference metric answers the question “How good does this picrure look?”

Consider these ...

Do the output pictures look like the input?

Input

Output

Squared Error

$I(h,k)$

$\hat{I}(h,k)$

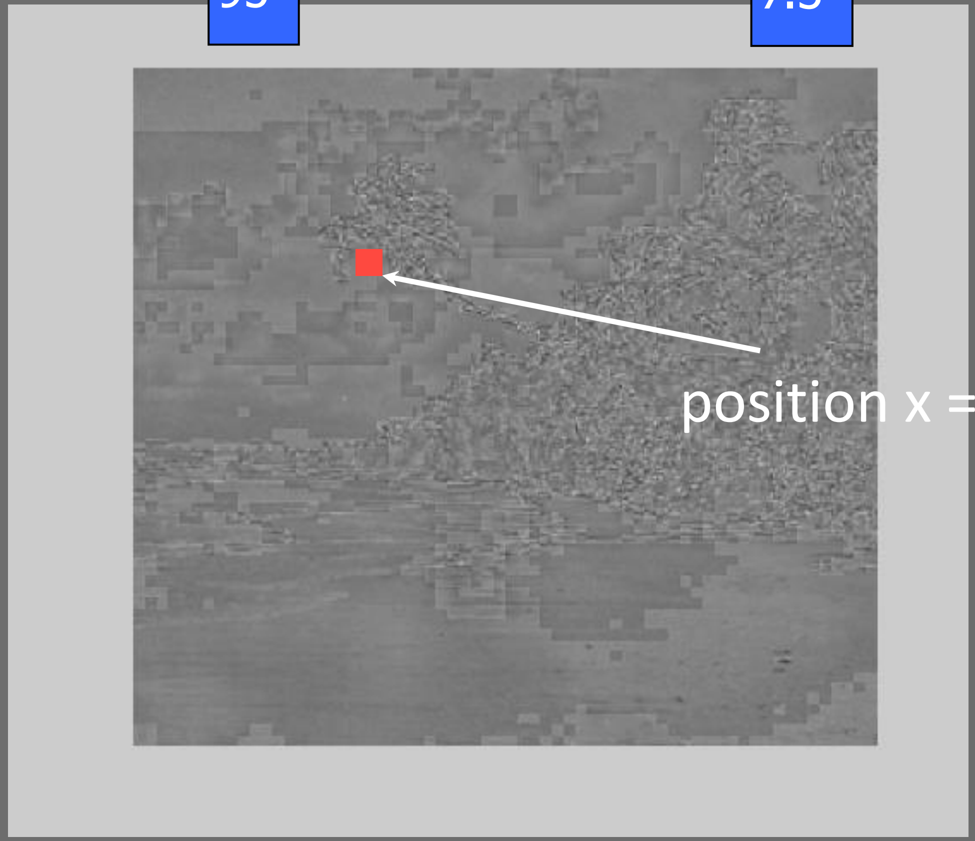
White = 255
0 = 128
Black = 0

$$e(\vec{x}) = \hat{I}(\vec{x}) - I(\vec{x})$$

$$\mathrm{MSE} = \frac{1}{NM}\sum_{\mathbf{x}}(e(\mathbf{x}))^2 \qquad \mathrm{MAE} = \frac{1}{NM}\sum_{\mathbf{x}}|e(\mathbf{x})|$$

95

7.3



position x = [h,k]

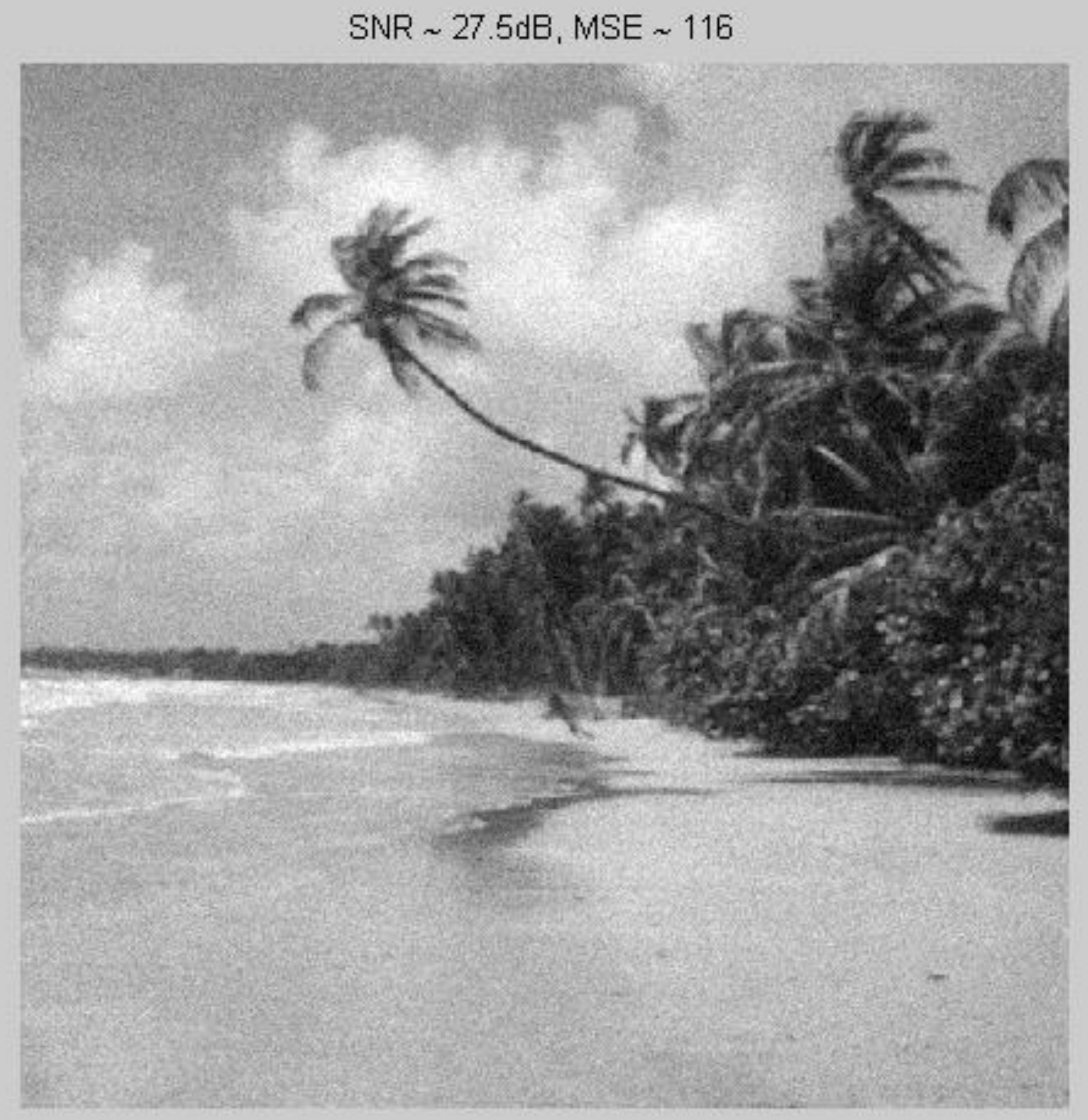
$$\mathrm{SNR} = 10\log 10\frac{\frac{1}{NM}\sum_{\mathbf{x}} I(\mathbf{x})^2}{\mathrm{MSE}}$$

49

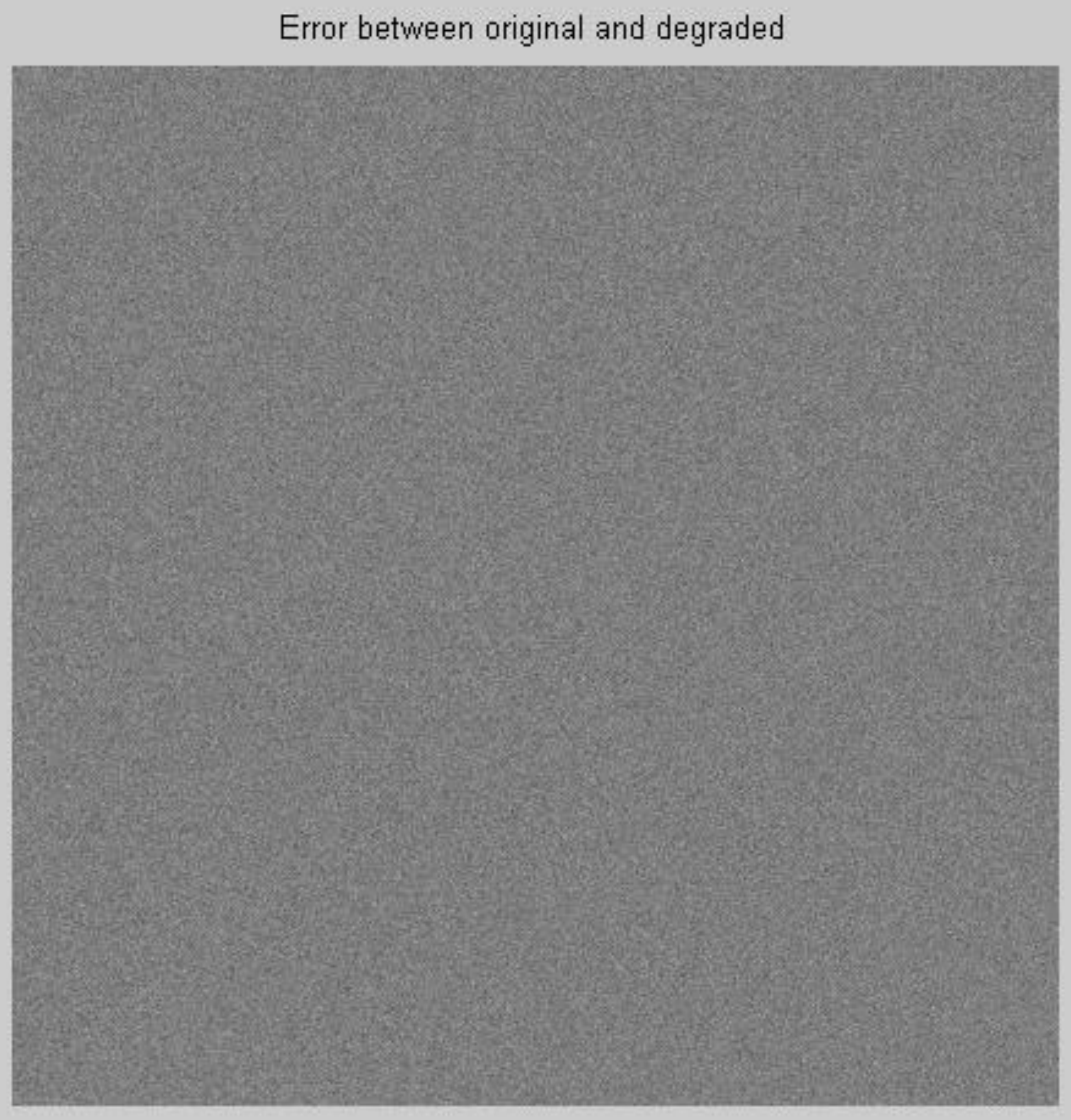
$$\mathrm{PSNR} = 10\log 10\frac{255^2}{\mathrm{MSE}} \quad \text{Units of dB}$$

56

Output contains added noise



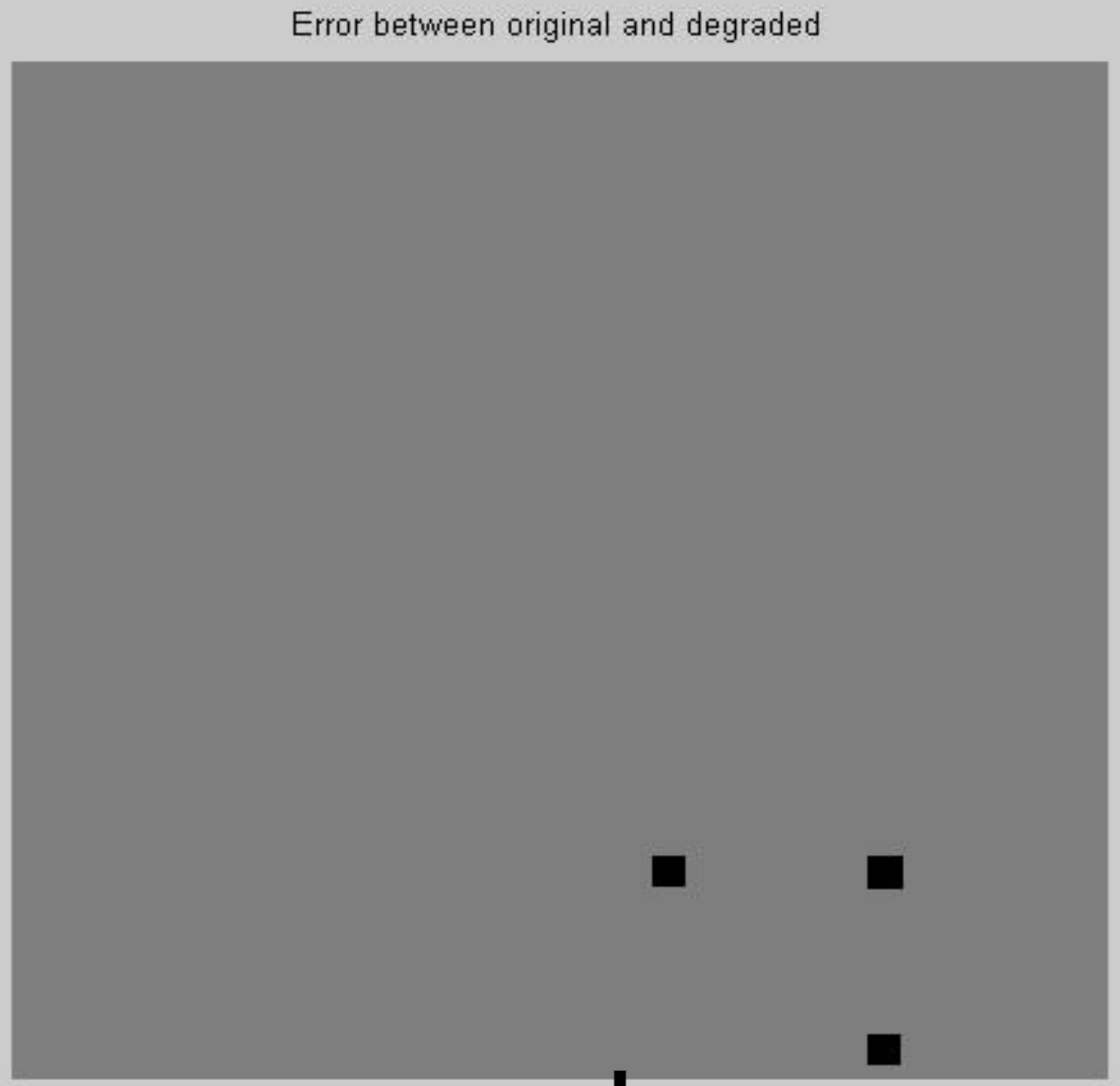
PSNR = 27.5 dB, MSE = 116



$e(\vec{x})$

Output contains missing blocks

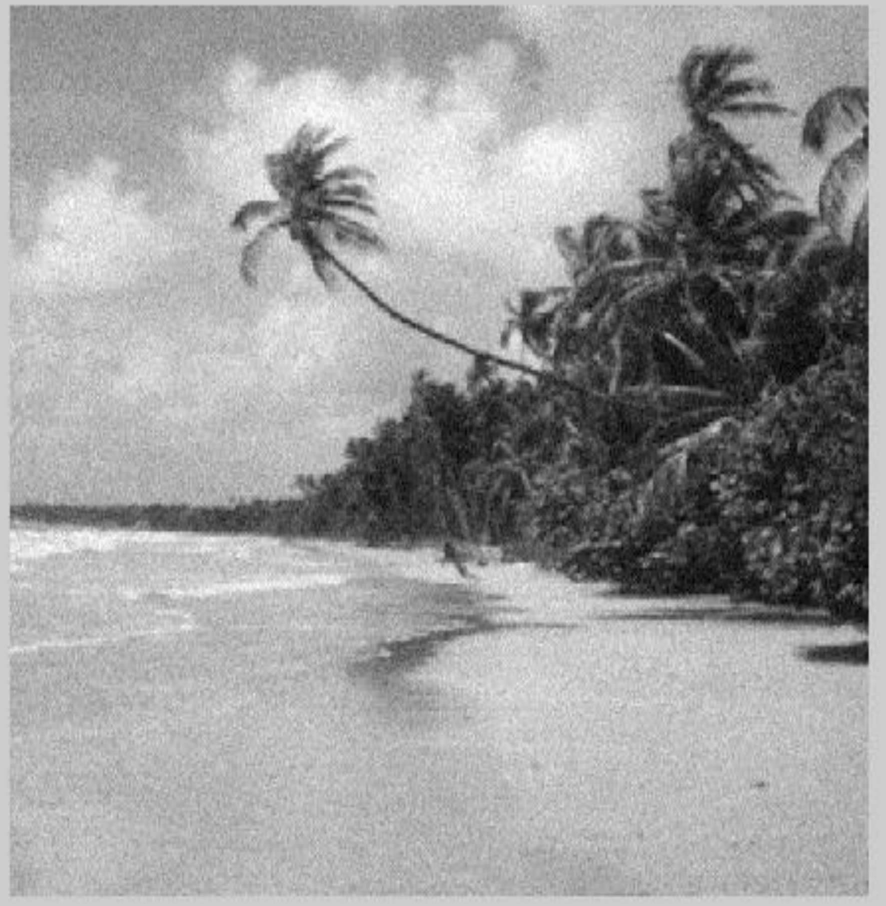
PSNR = 27.5 dB, MSE = 116



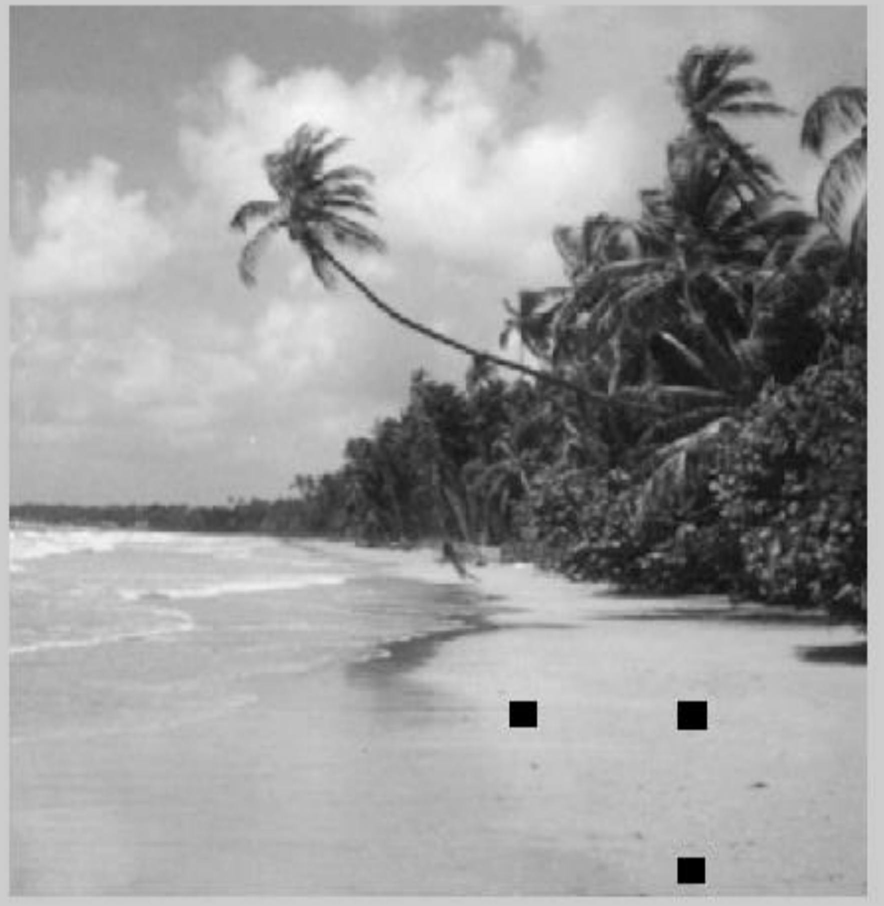
$e(x)$

Are these the same?

28 dB



27dB

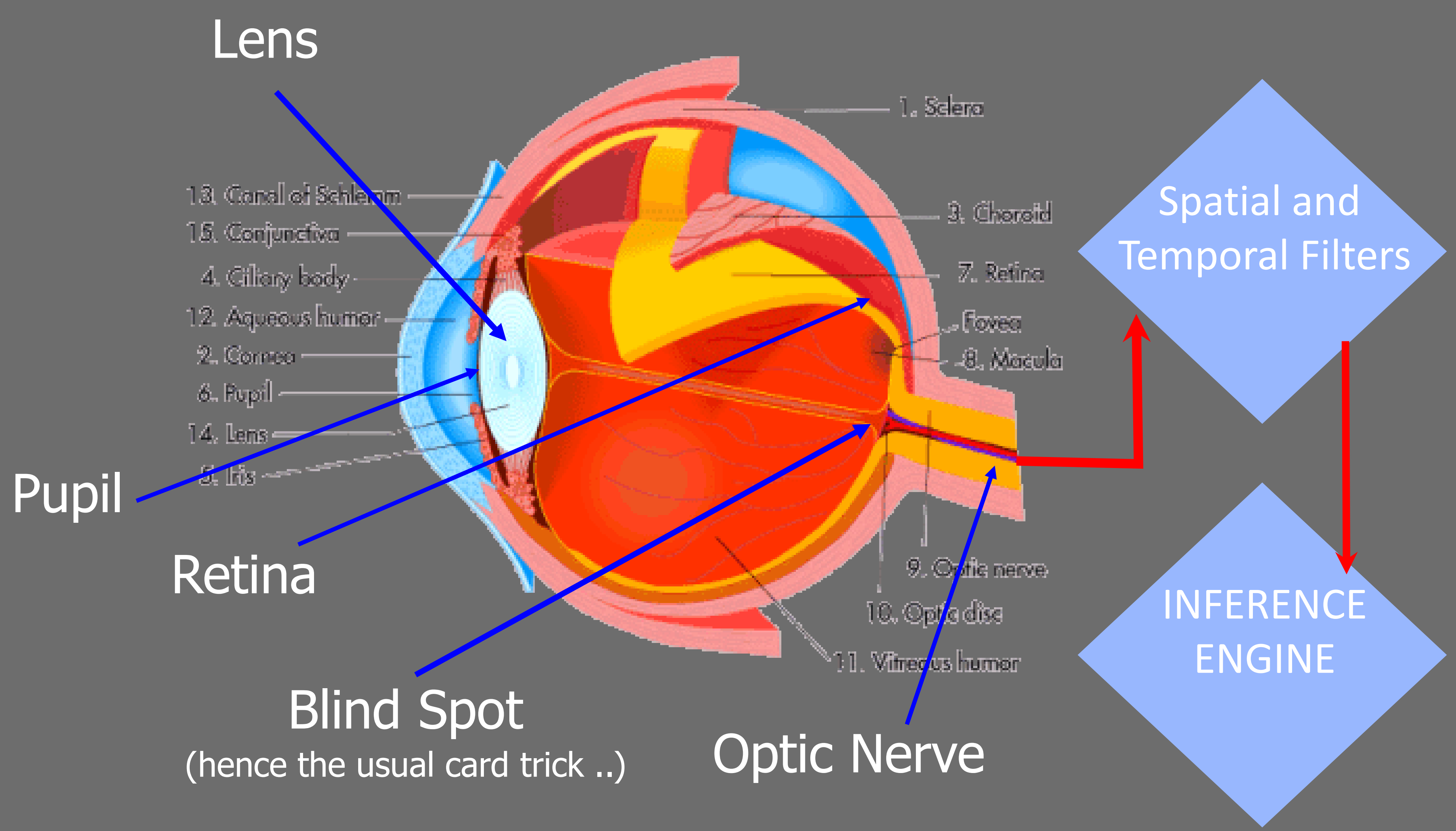


27dB

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 alot because it is amenable to mathematical analysis: you can optimize systems for PSNR
- 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



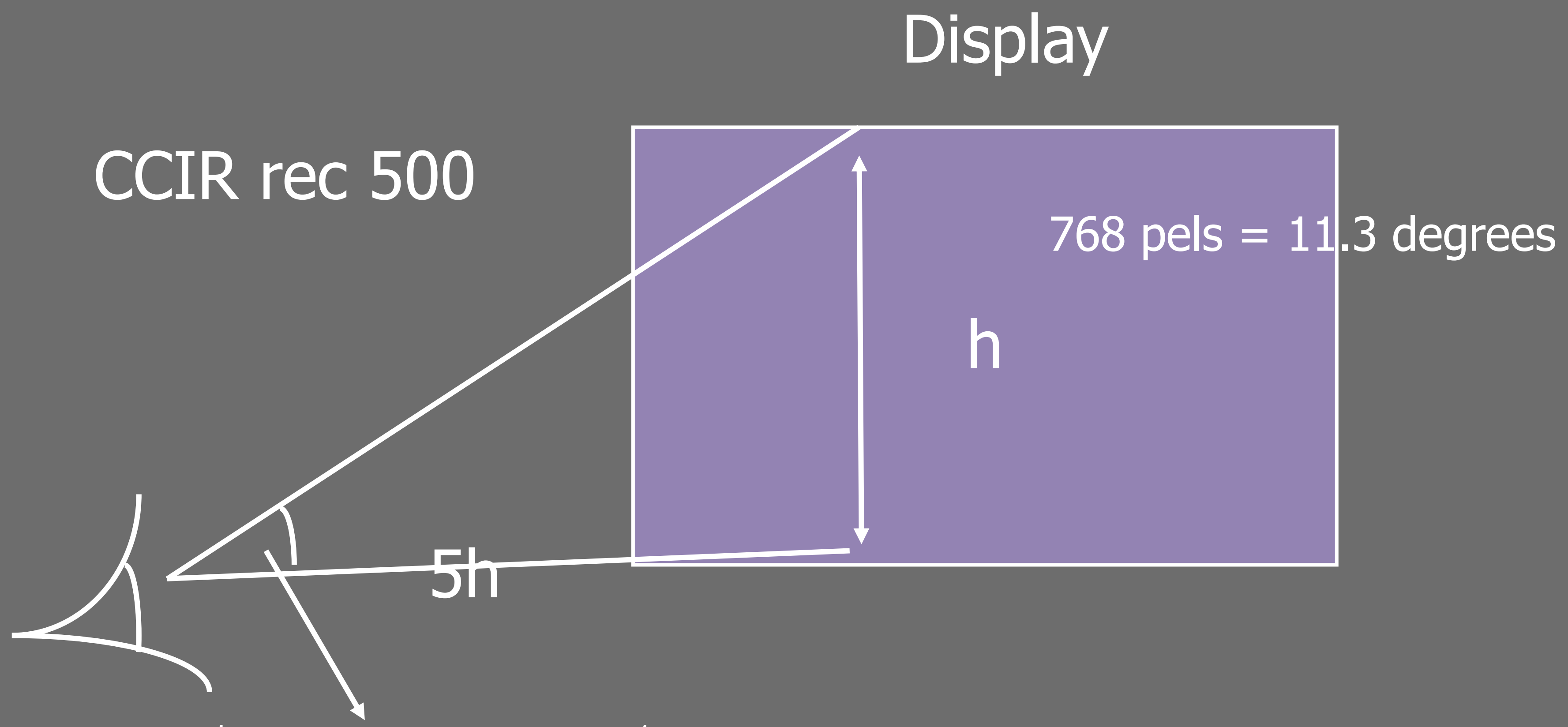
- Light is focussed onto the retina
- Electrical Impulses from the retina are chanelled by the optic nerve to the Visual Cortex
- The Visual Cortex does a whole bunch of smart things including filtering, object recognition, edge detection.
- In `primitive animals' A LOT of processing happens just behind 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

$$\Delta I \,/\, I = k$$

Perception and angles

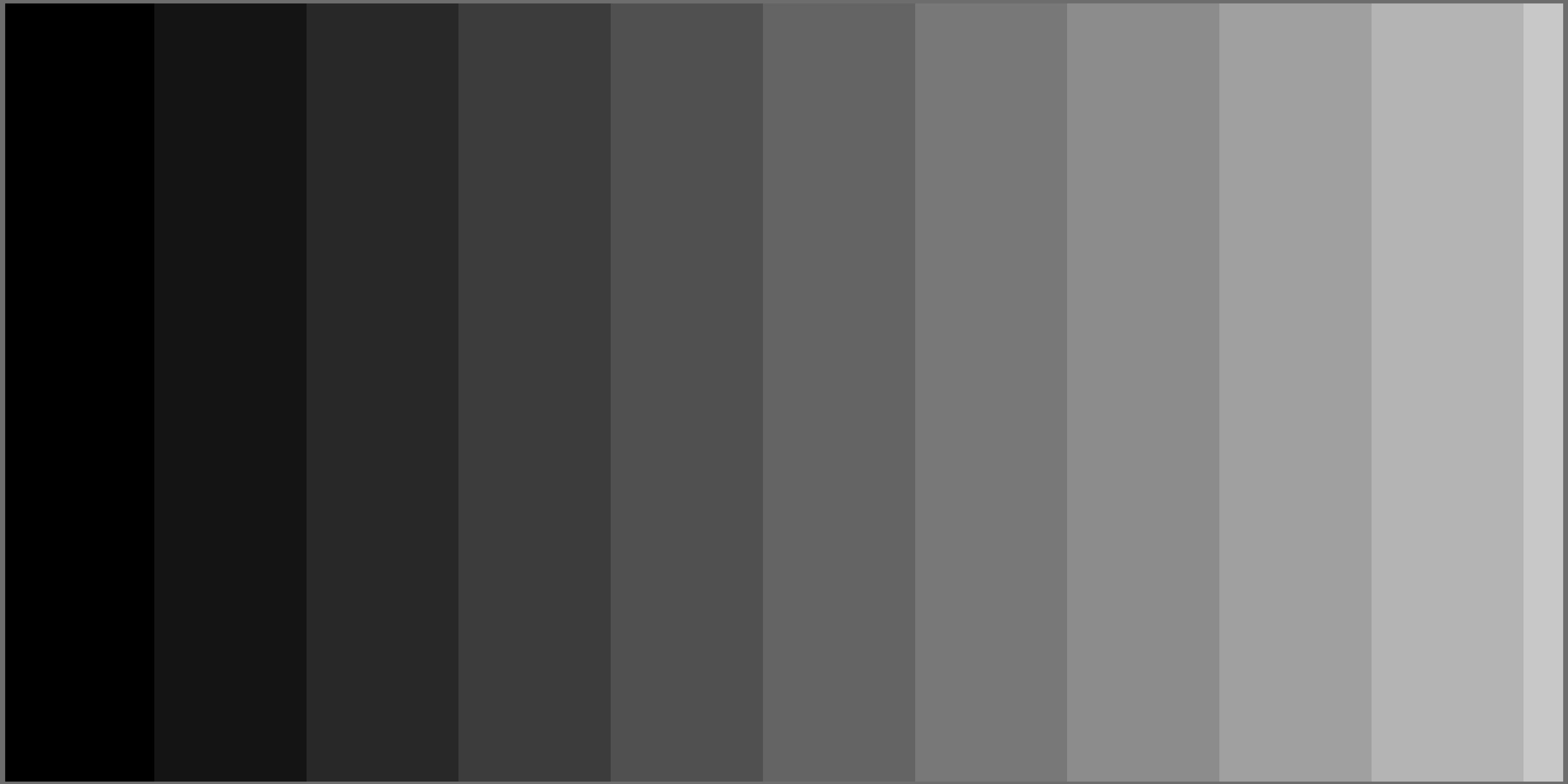
- What matters is the degrees of arc subtended by an object on the retina.
- Hence distance to the image plays a part.
- A standard exists for viewing distance.

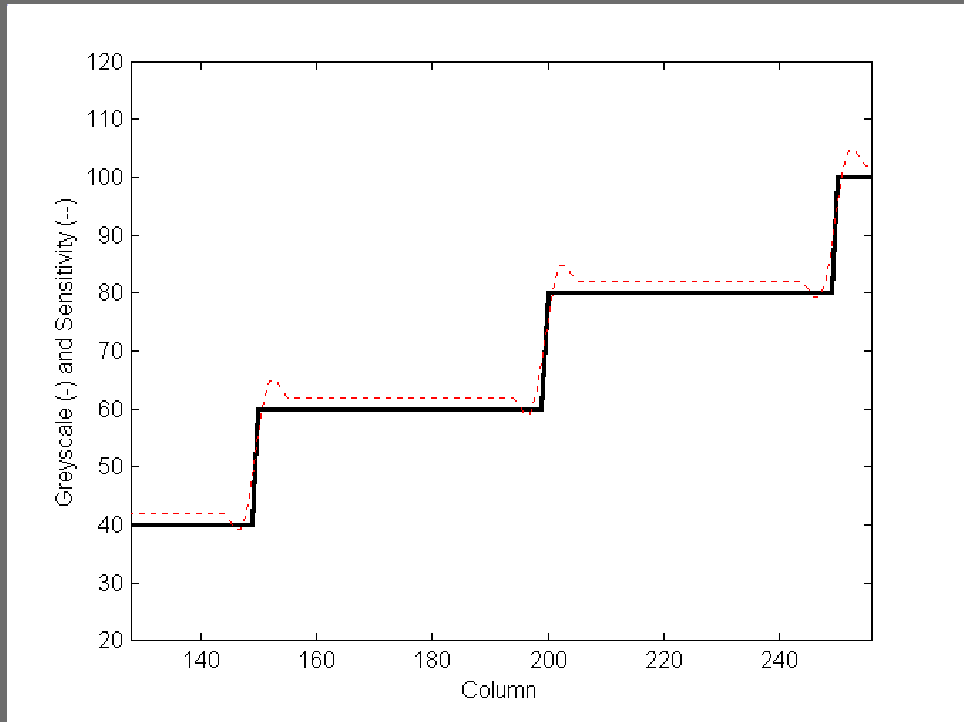


$$\tan^{-1}(h/(5h)) = \tan^{-1}(1/5) = 11.3 \text{ deg}$$

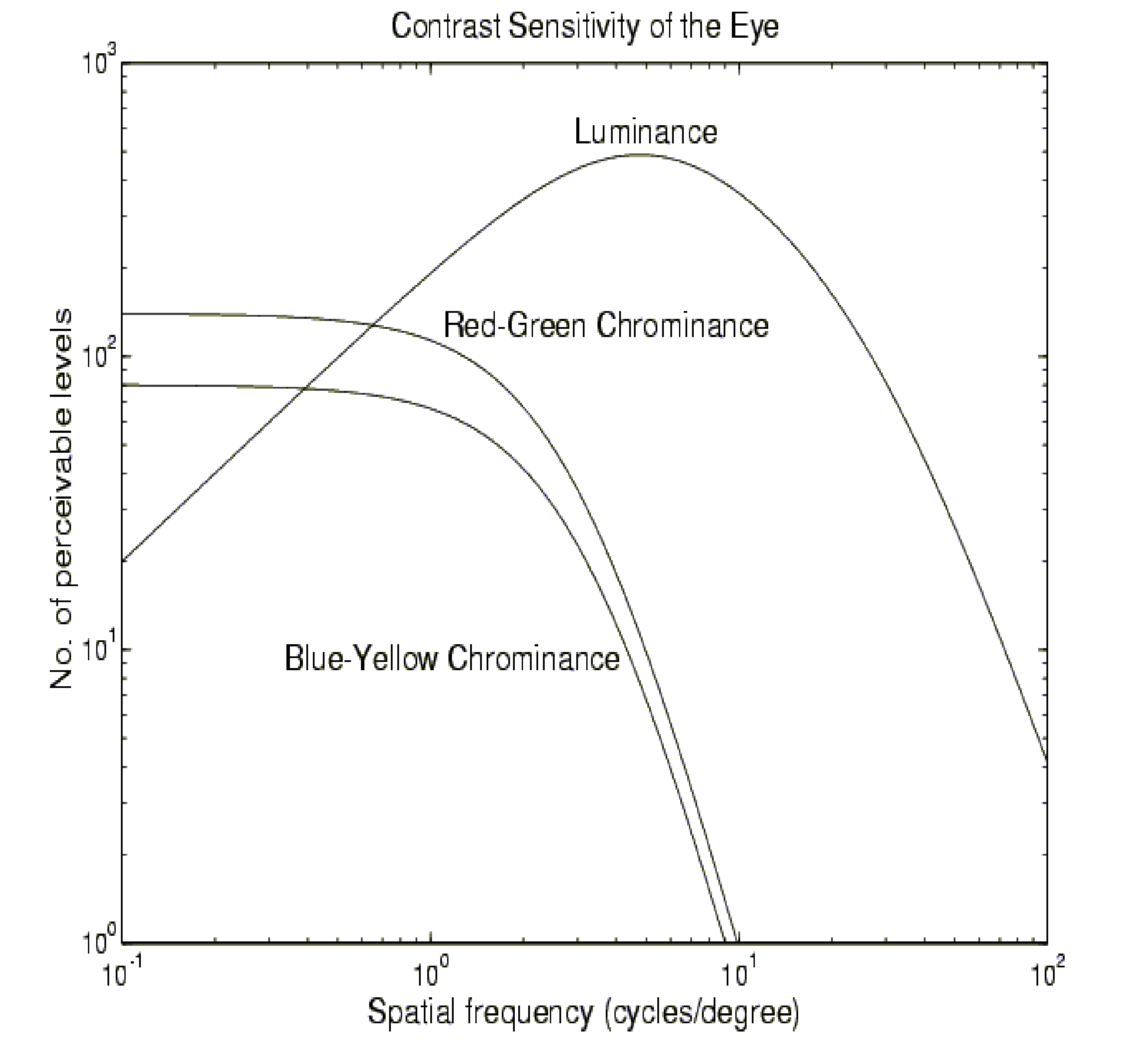
Visual acuity is 1/60 cpd, i.e. you can just resolve 2 lines separated 1/60 degrees.
Ipad is 2048x1536 .. which is at this limit when held 41cm away

Mach Bands and "Summation"

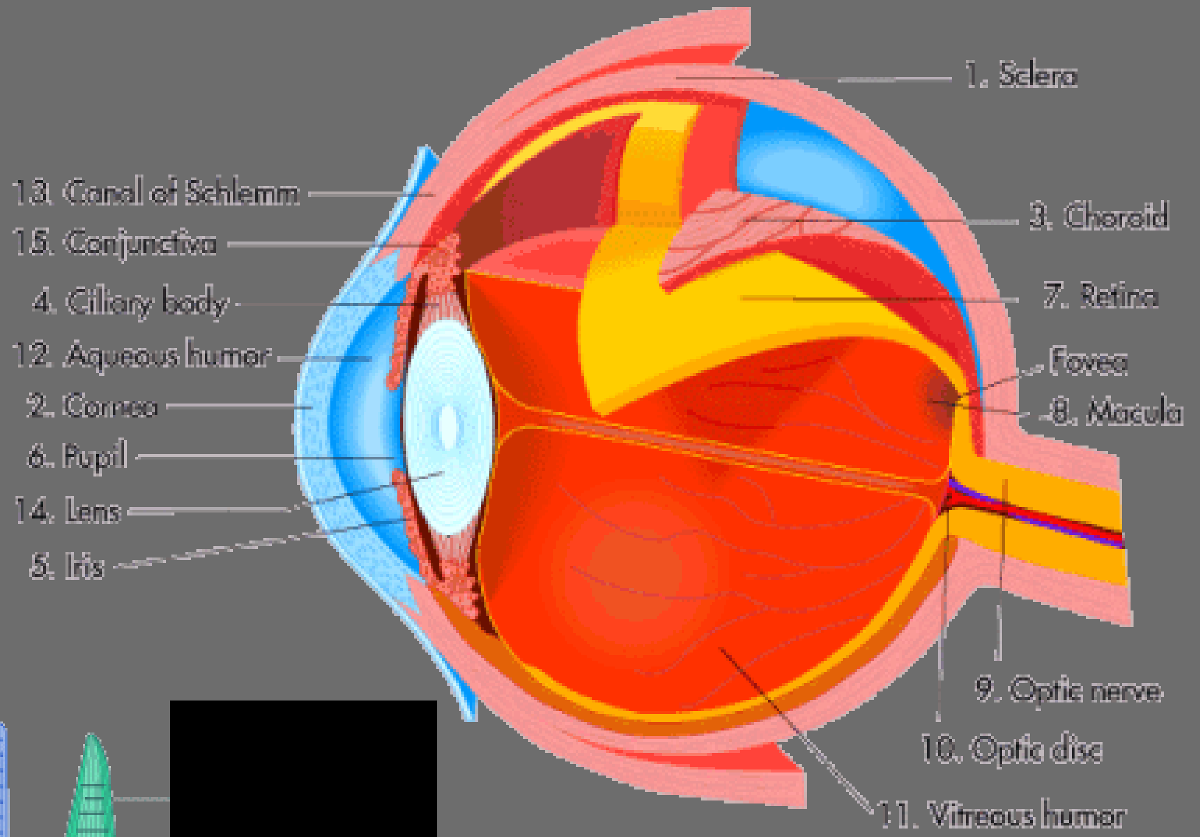


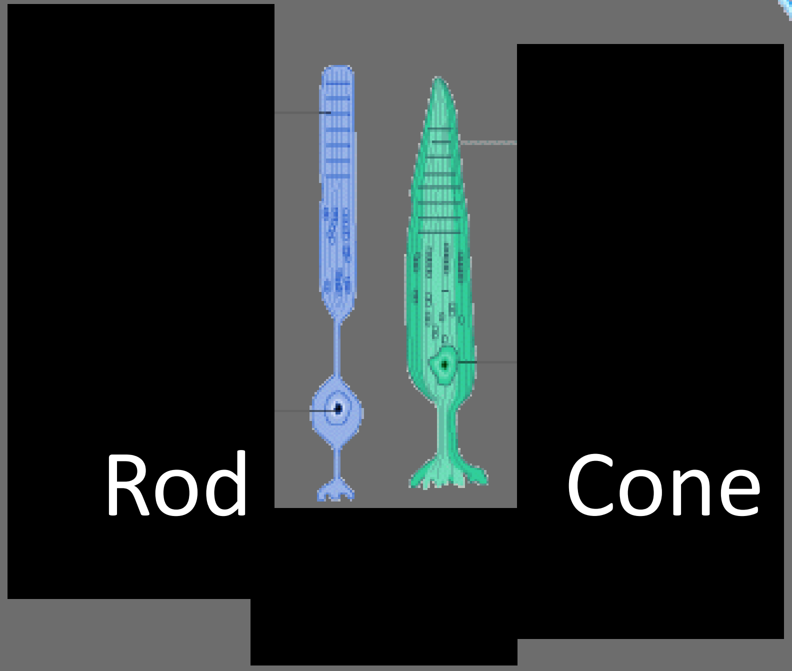


Sensitivity to Spatial Frequency



Eye and colour sensitvity





- 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

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

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

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

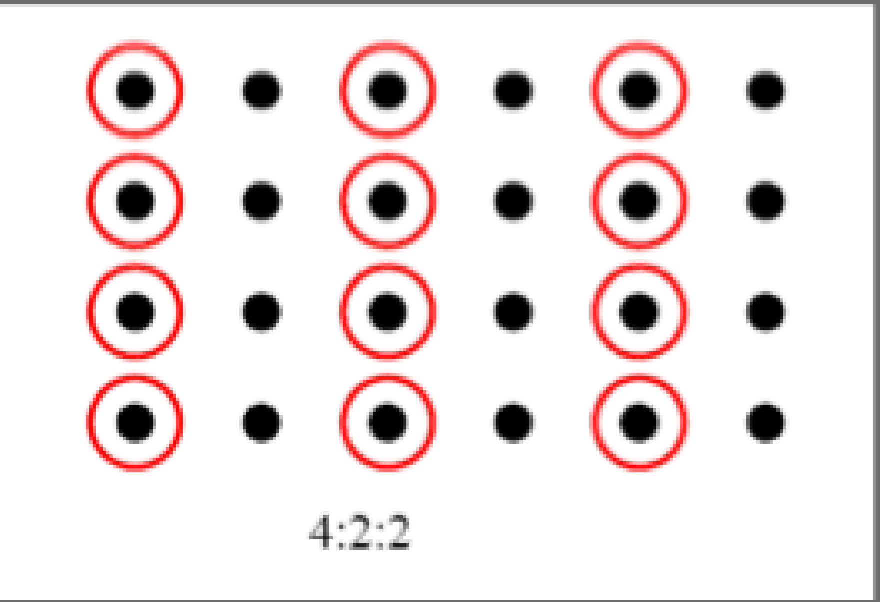
Downsample U/V 2:1, Leave Y alone

Downsample U/V 4:1, Leave Y alone

Downsample Y/U/V 4:1

Hence YUV 420, 422 etc

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



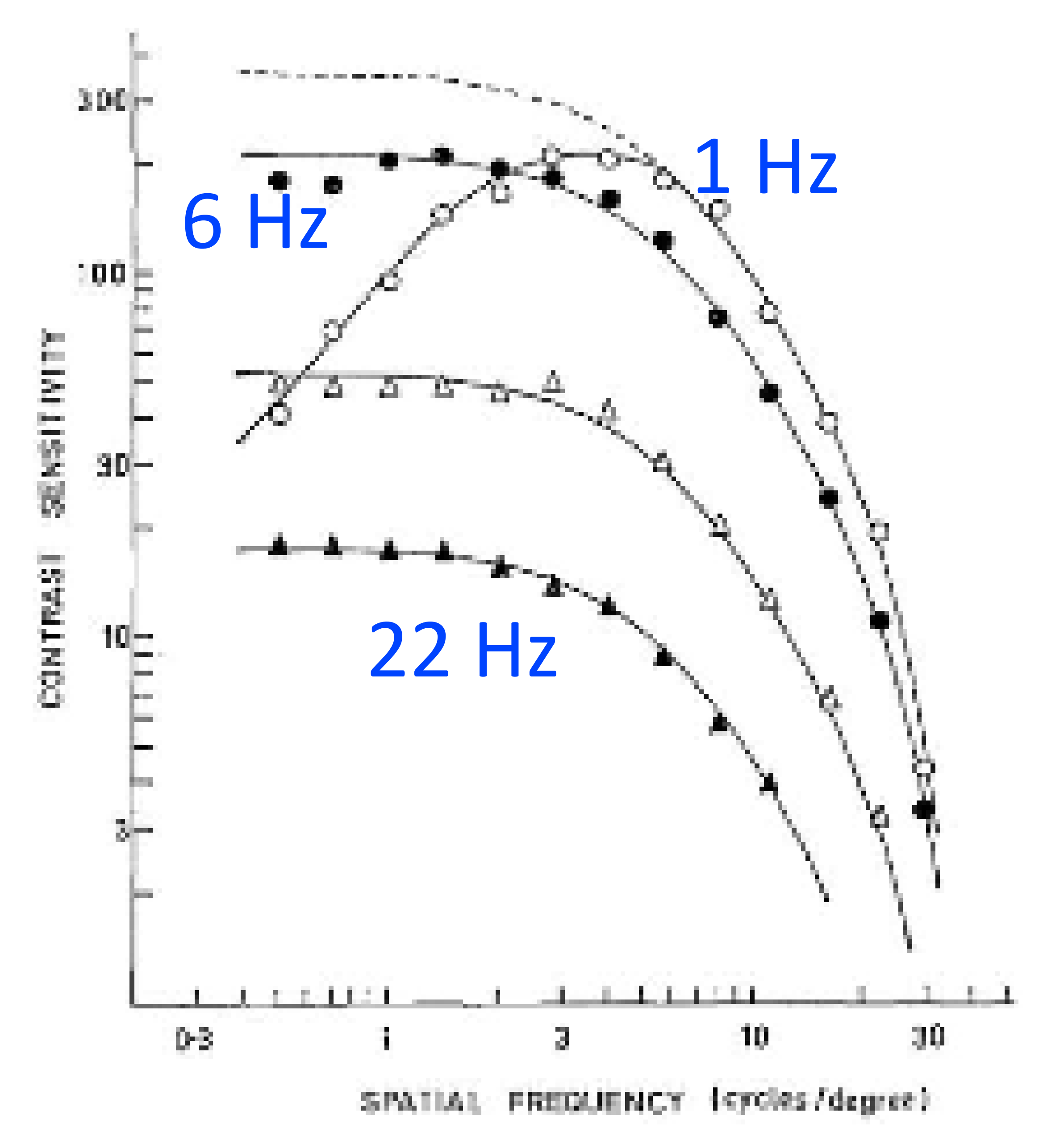
Actually this was also done in Analog TV because the bandwidth for the colour signals was 1/2 that of the Luma.

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

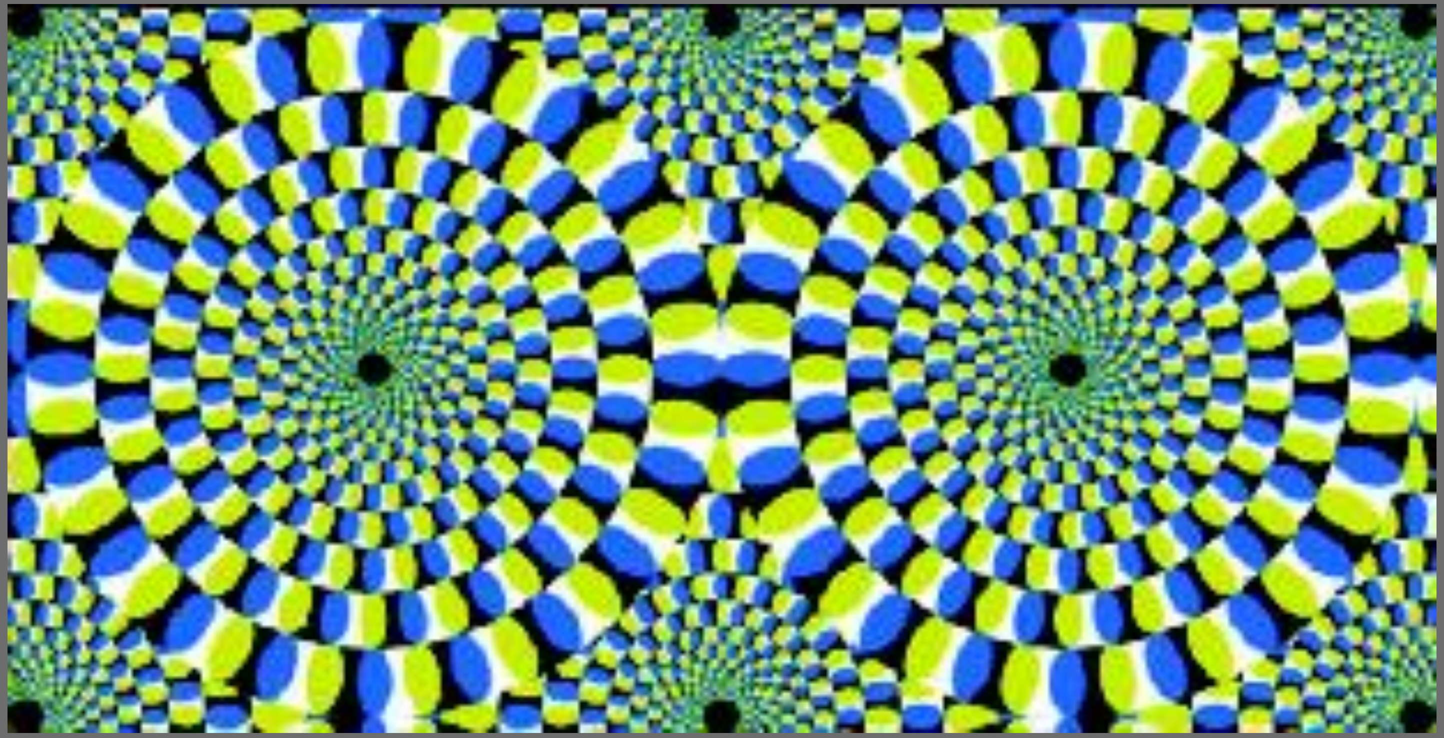
J. G. Robson
Physiological Laboratory, Cambridge, England

(Received 3 March 1966)

Journal of the Optical Society of America

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) measues 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 Testing and Predicting MOS

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

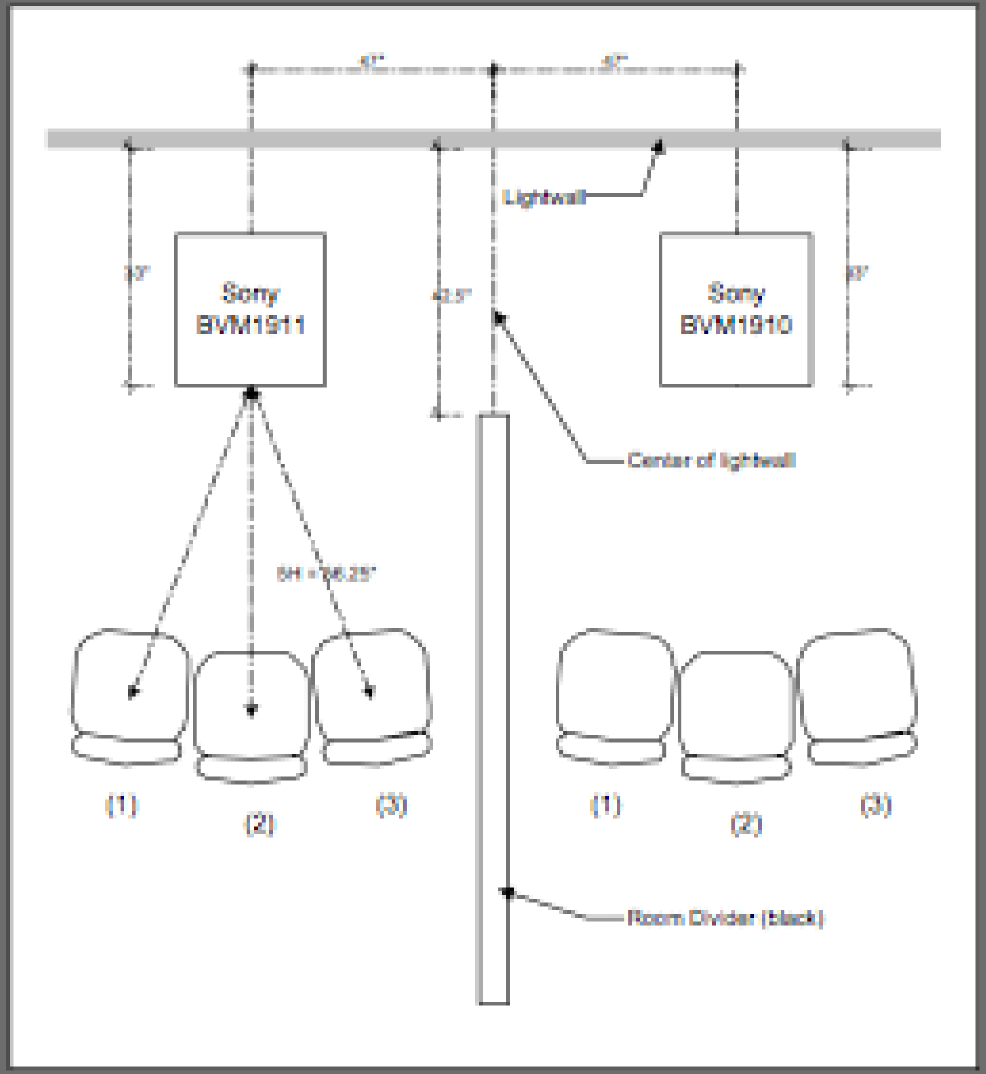
Excellent

Good

Fair

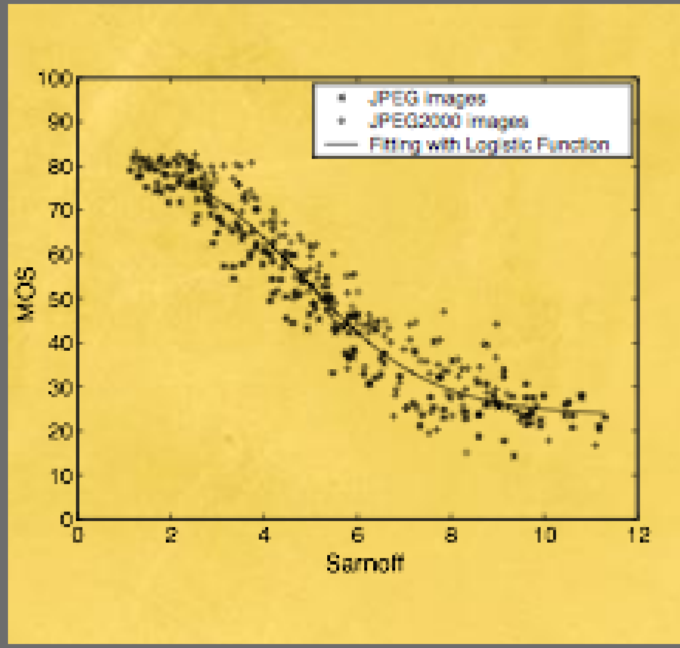
Poor

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

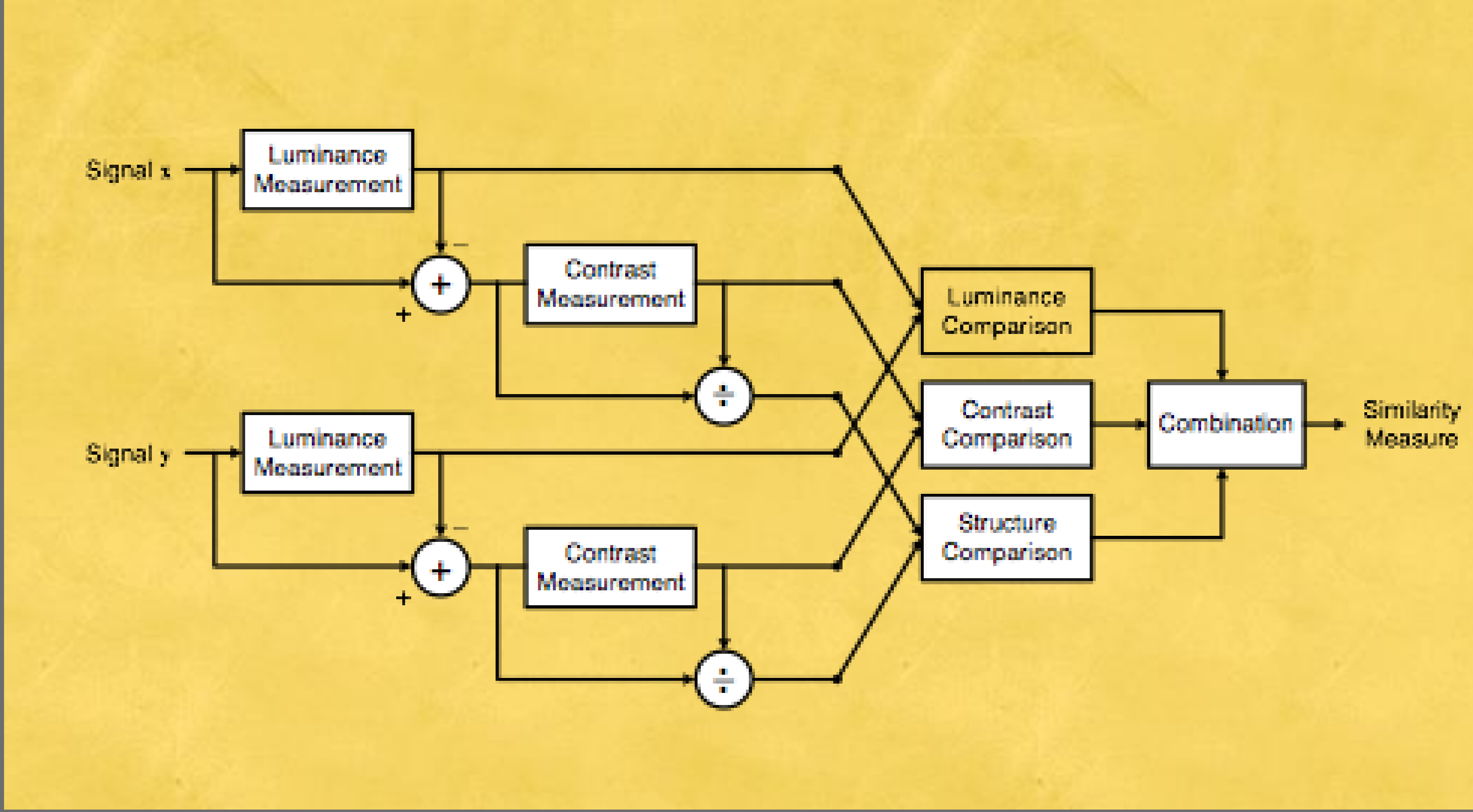


Image Quality Assessment: From Error Visibility to Structural Similarity, Zhou Wang, Alan Conrad Bovik, Hamid, Rahim Sheikh, and E. P. Simoncelli, IEEE Trans TIP, 2004.

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 alot of local variance .. which is in keeping with CSF masking

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

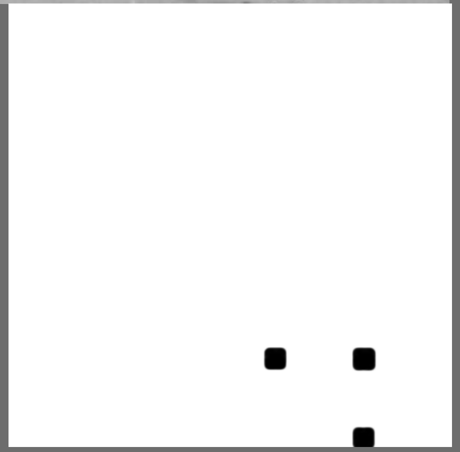
Structure is a variance measure after luma and contrast differences are compensated

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

SSIM Maps

SSIM Comparison

PSNR 28.3
SSIM 0.78



PSNR 27.5
SSIM 0.99

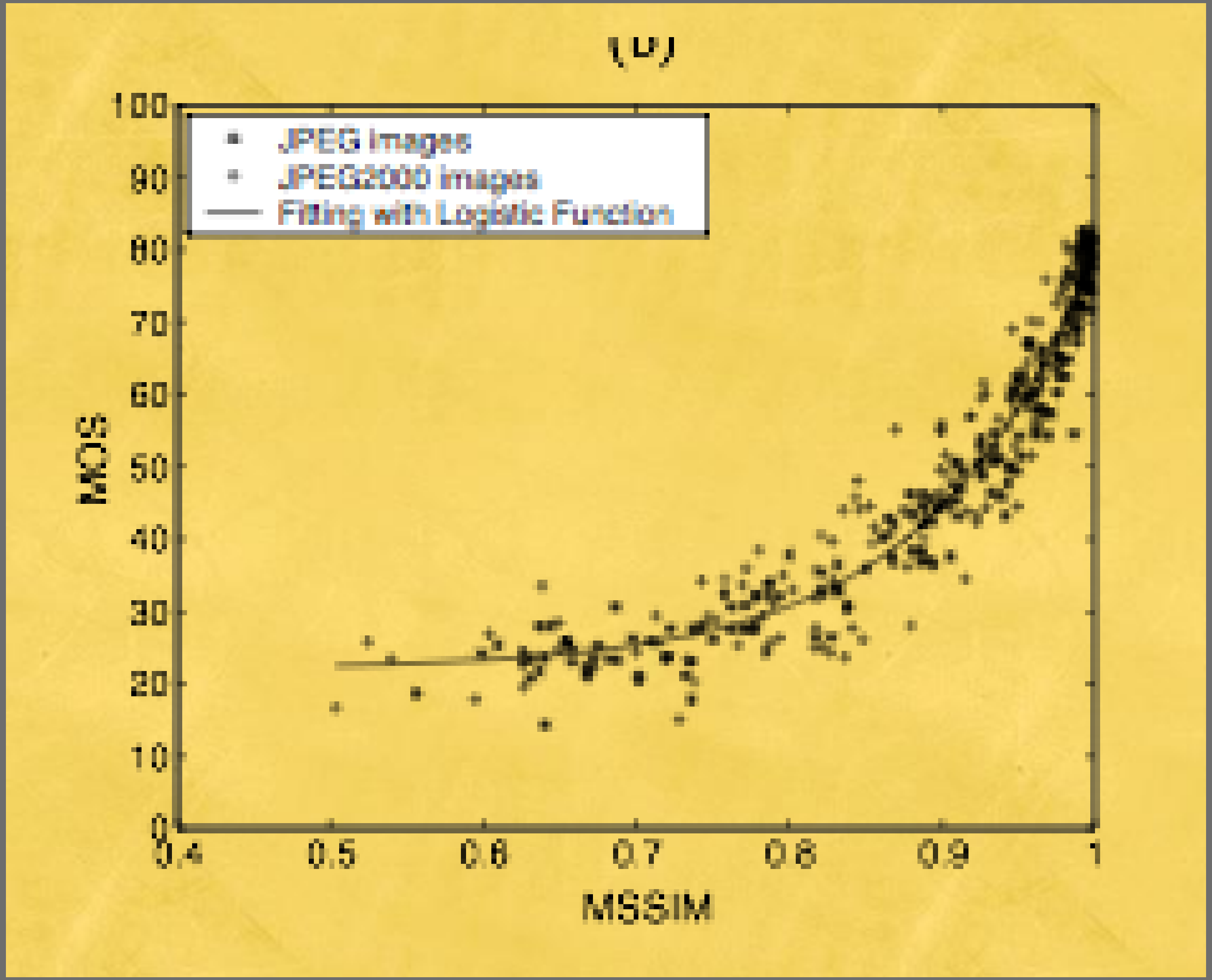
PSNR 27.5
SSIM 0.53

Predicting MOS

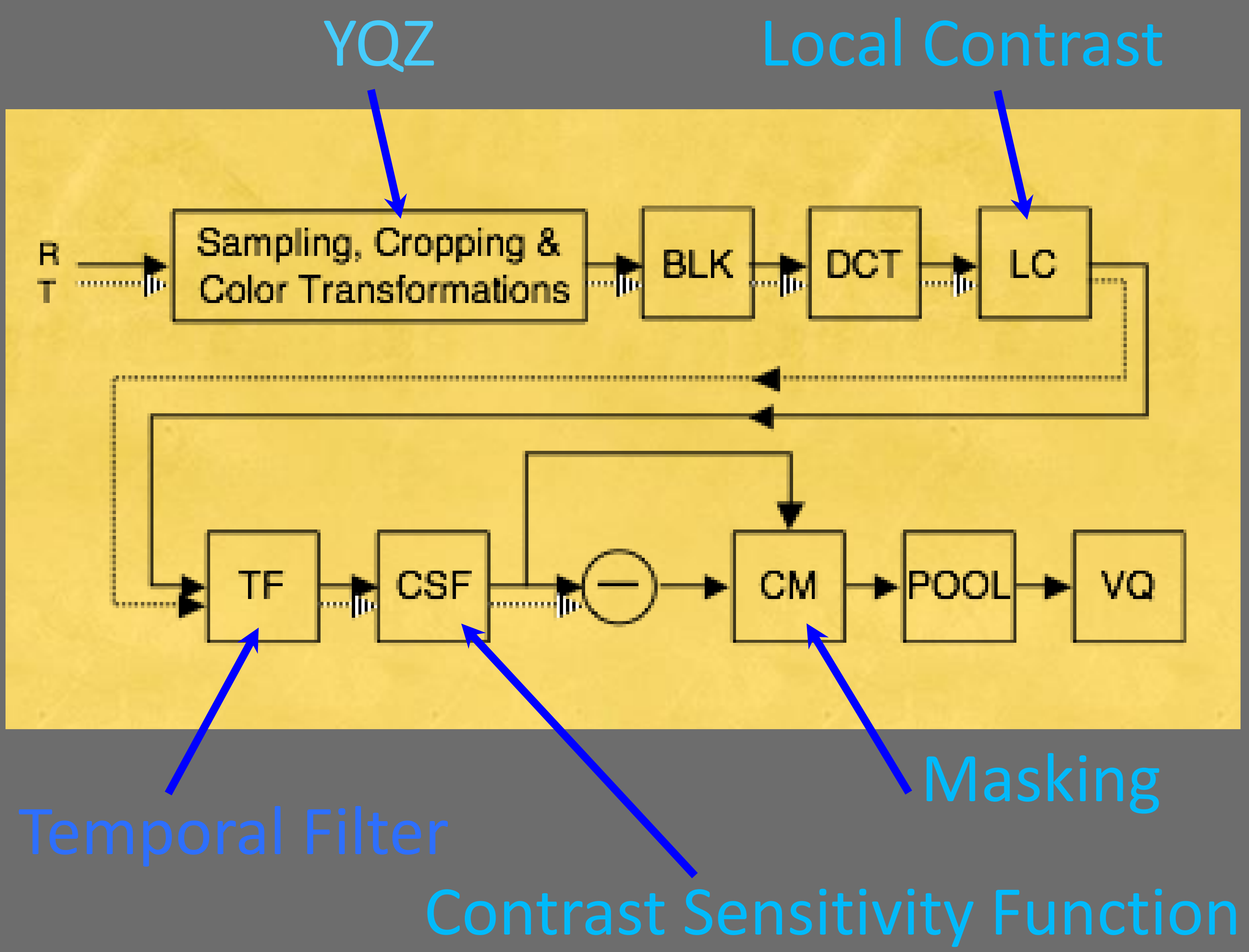
- A complaint about the SSIM is that it does not vary much numerically, and we do not know what a good number is

- We know 40dB PSNR in coding is good .. but not what 0.998 means in SSIM

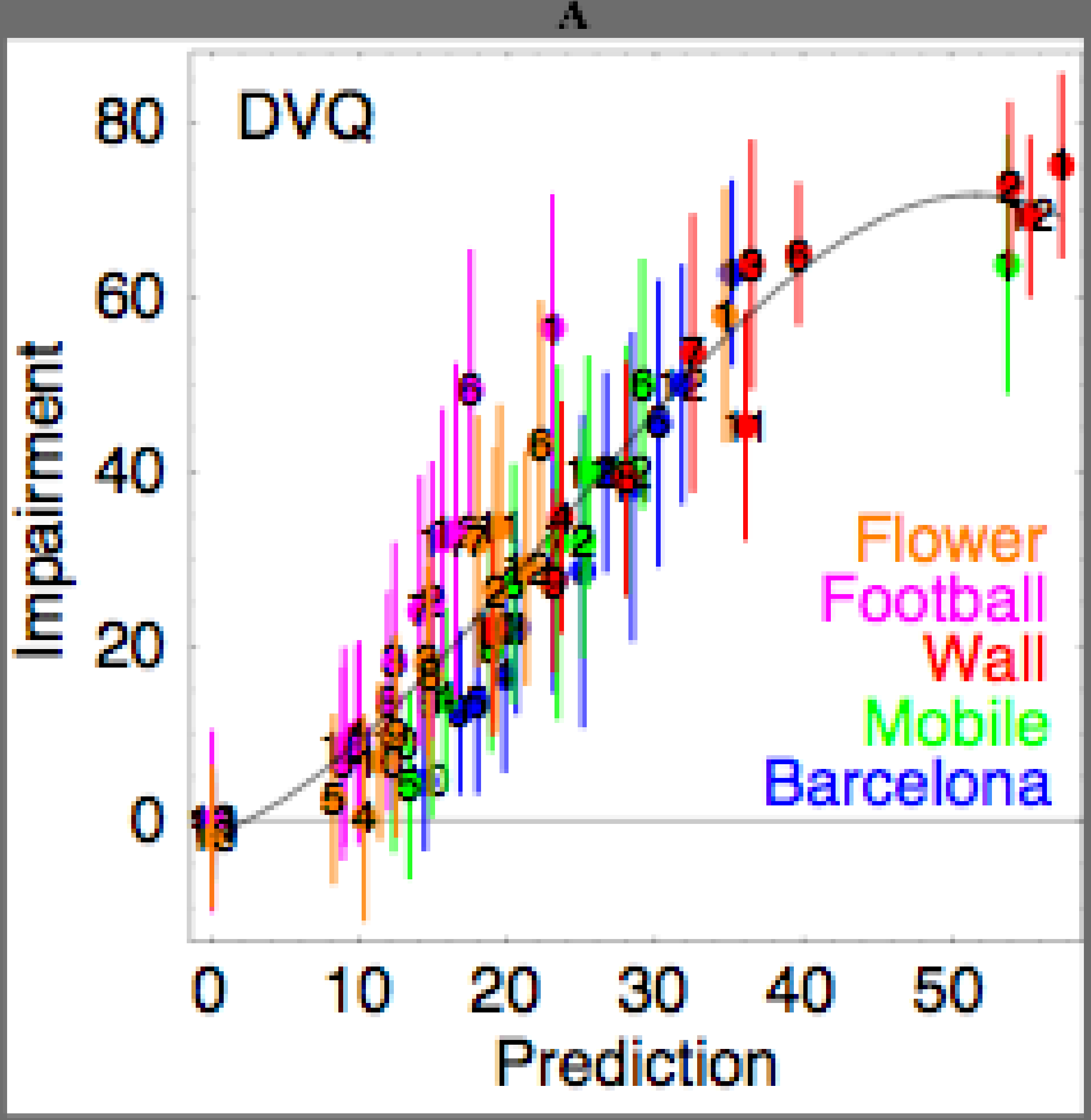
- In their original paper they actually provide a mapping to MOS.



Temporal Video Quality Metric



4th order polynomial fit



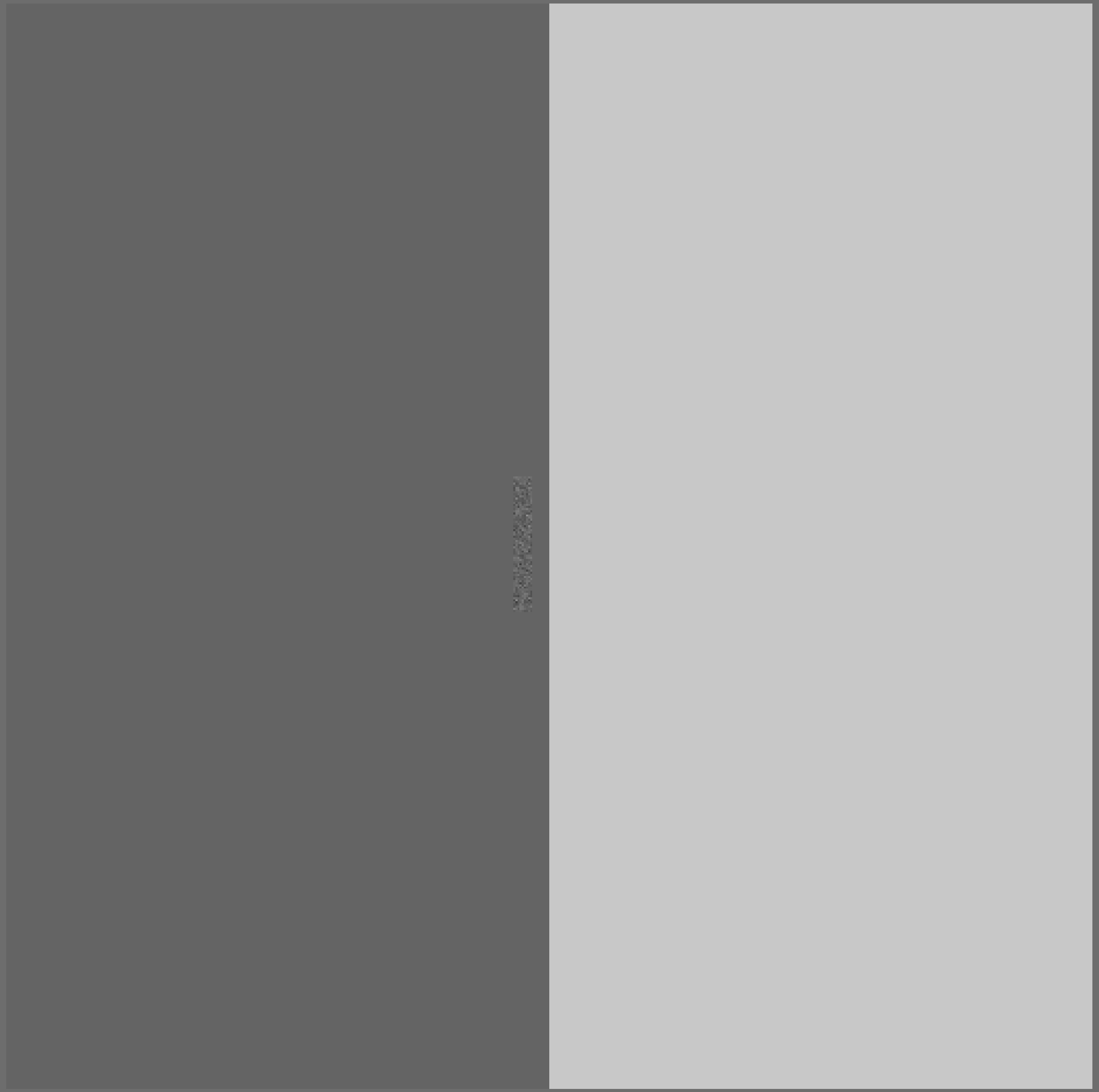
Watson @ NASA Ames 2001

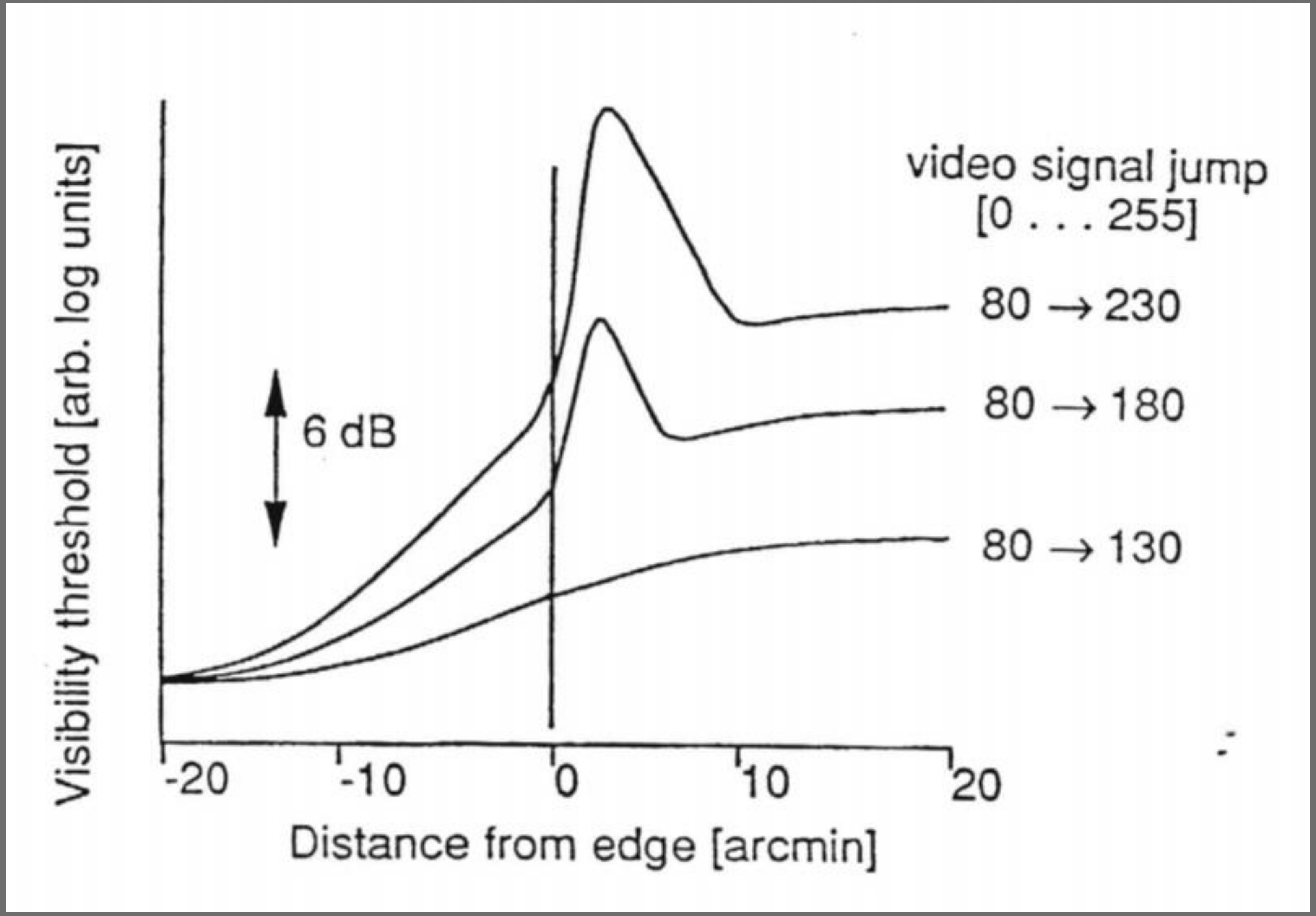
VMAF (Video Multimethod assessment Fusion)

- VMAF is a reference metric developed at USC sponsored by Netflix
- It varies between 0 – 100 (Ioannis will mention it)
- Now fast becoming an industry standard : implemented in ffmpeg too
- No explict 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

Interlude on earlier work

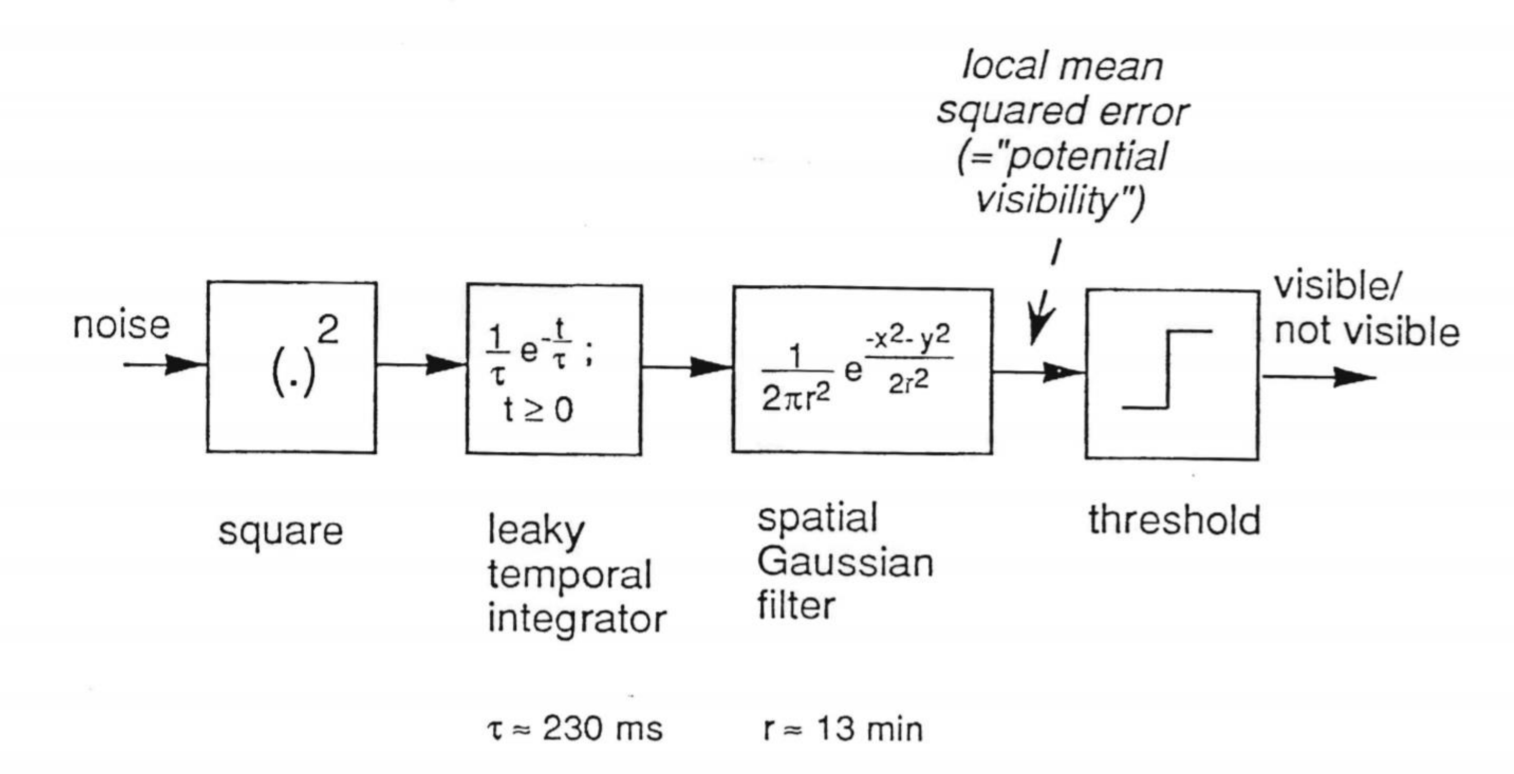
Early work in perceptual quality prediction





B. Girod, "What's Wrong With Mean Squared Error?" in A. B. Watson (ed.), Visual Factors of Electronic Image Communications, MIT Press, pp. 207-220, 1993.

Model



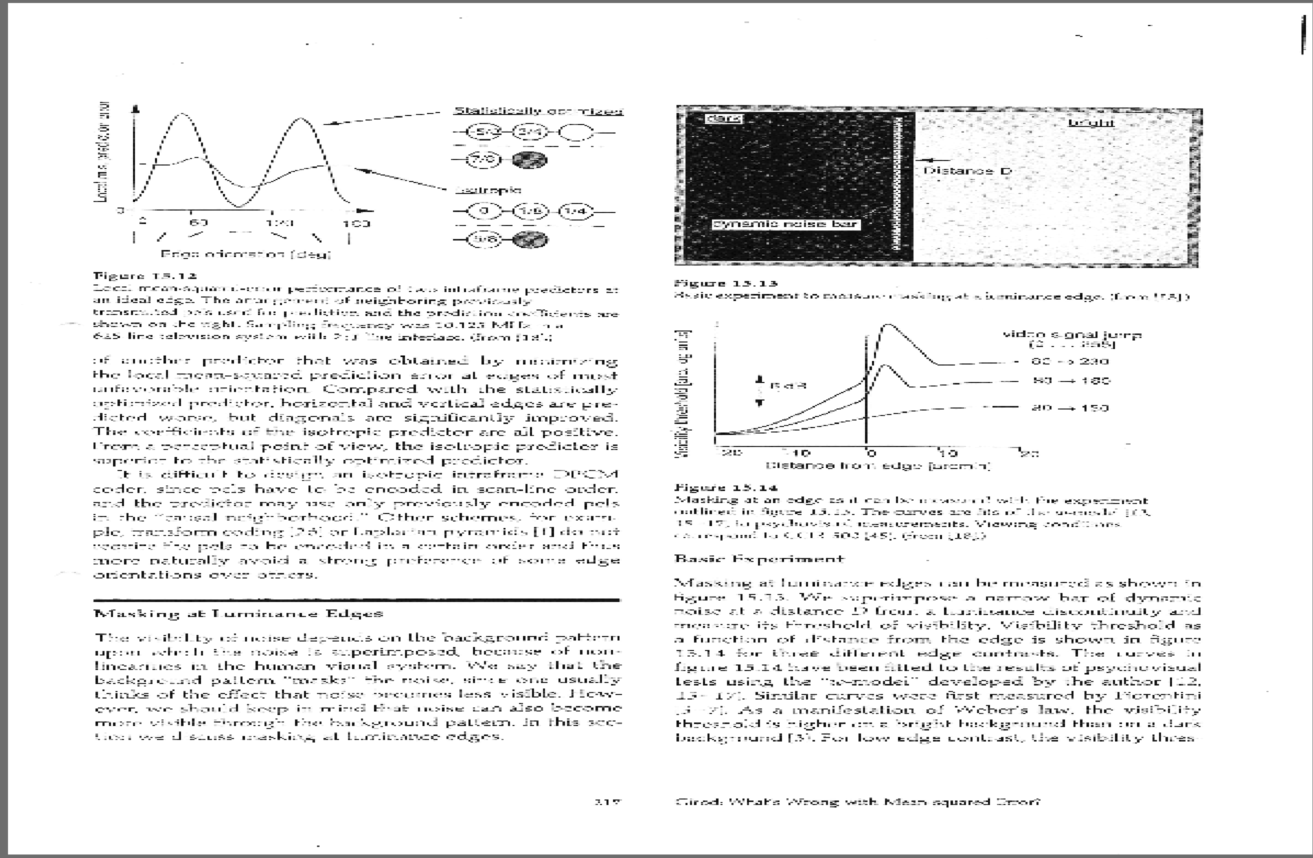


Figure 15.12
Local mean-squared-error performance of two intraframe predictors at an ideal edge. The arrangement of neighboring previously transmitted pels used for prediction and the prediction coefficients are shown on the right. Sampling frequency was 10.125 MHz in a 625-line television system with 2:1 line interlace. (From [18].)

of another predictor that was obtained by minimizing the local mean-squared prediction error at edges of most unfavorable orientation. Compared with the statistically optimized predictor, horizontal and vertical edges are predicted worse, but diagonals are significantly improved. The coefficients of the isotropic predictor are all positive. From a perceptual point of view, the isotropic predictor is superior to the statistically optimized predictor.

It is difficult to design an isotropic intraframe DPCM coder, since pels have to be encoded in scan-line order, and the predictor may use only previously encoded pels in the "causal neighborhood." Other schemes, for example, transform coding [26] or Laplacian pyramids [1] do not require the pels to be encoded in a certain order and thus more naturally avoid a strong preference of some edge orientations over others.

Masking at Luminance Edges

The visibility of noise depends on the background pattern upon which the noise is superimposed, because of nonlinearities in the human visual system. We say that the background pattern "masks" the noise, since one usually thinks of the effect that noise becomes less visible. However, we should keep in mind that noise can also become more visible through the background pattern. In this section we discuss masking at luminance edges.

Figure 15.13
Basic experiment to measure masking at a luminance edge. (From [18].)

Figure 15.14
Masking at an edge as it can be measured with the experiment outlined in figure 15.13. The curves are fits of the w-model [13, 15–17] to psychovisual measurements. Viewing conditions correspond to CCIR 500 [45]. (From [18].)

Basic Experiment

Masking at luminance edges can be measured as shown in figure 15.13. We superimpose a narrow bar of dynamic noise at a distance D from a luminance discontinuity and measure its threshold of visibility. Visibility threshold as a function of distance from the edge is shown in figure 15.14 for three different edge contrasts. The curves in figure 15.14 have been fitted to the results of psychovisual tests using the "w-model" developed by the author [12, 13–17]. Similar curves were first measured by Fiorentini [5–7]. As a manifestation of Weber's law, the visibility threshold is higher on a bright background than on a dark background [3]. For low edge contrast, the visibility thres-

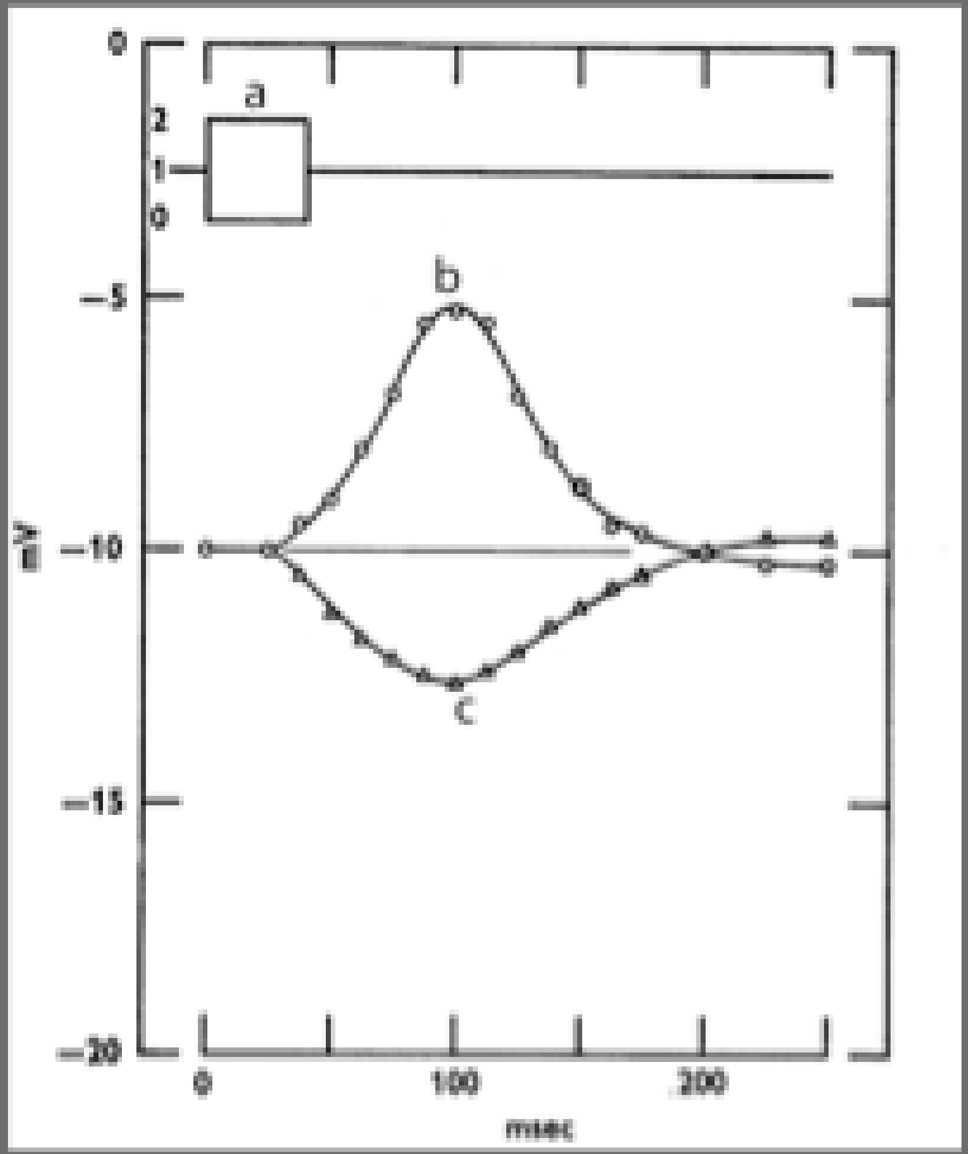
317

Girod: What's Wrong with Mean-squared Error?

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 that turning the lights up.

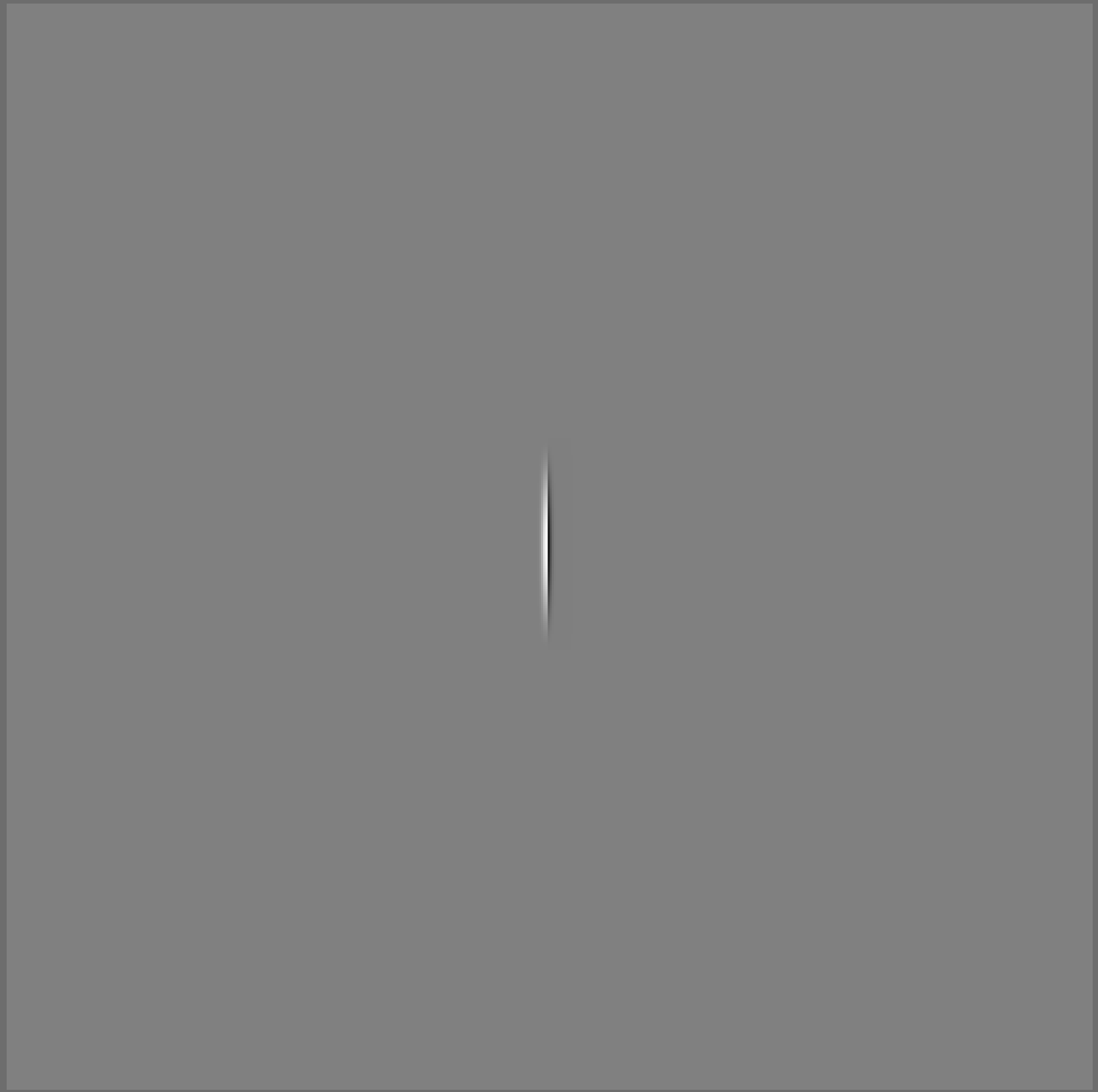


Experiment on turtle Cones. Light on/off for 40 ms and then measured mV.

Black-White Asymmetry in Visual Perception, Zhong Lin Lu and George Spirling, Journal of Vision 2005

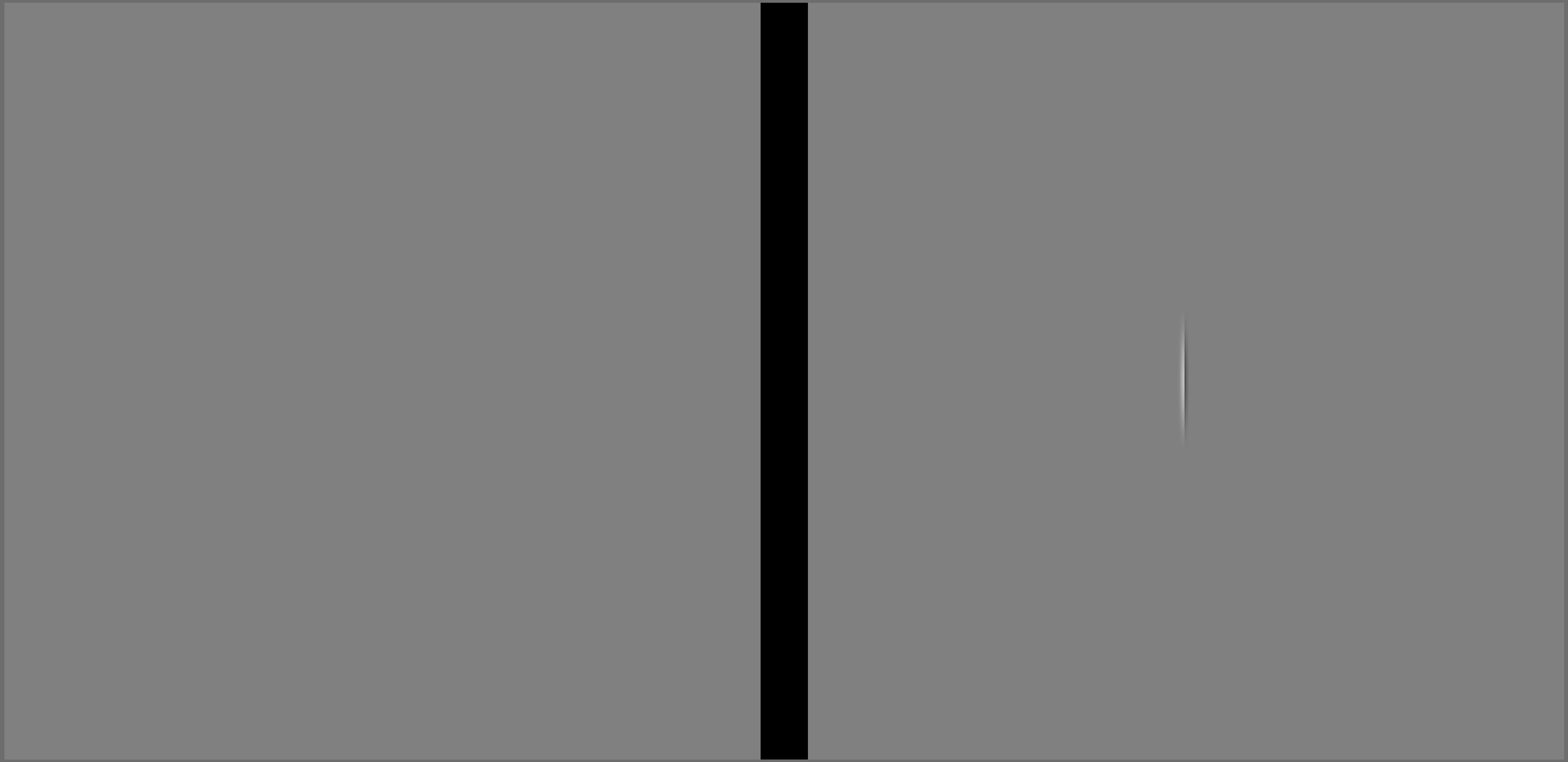
In 90‘s some experimented with direct assessment of artefacts

- 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.

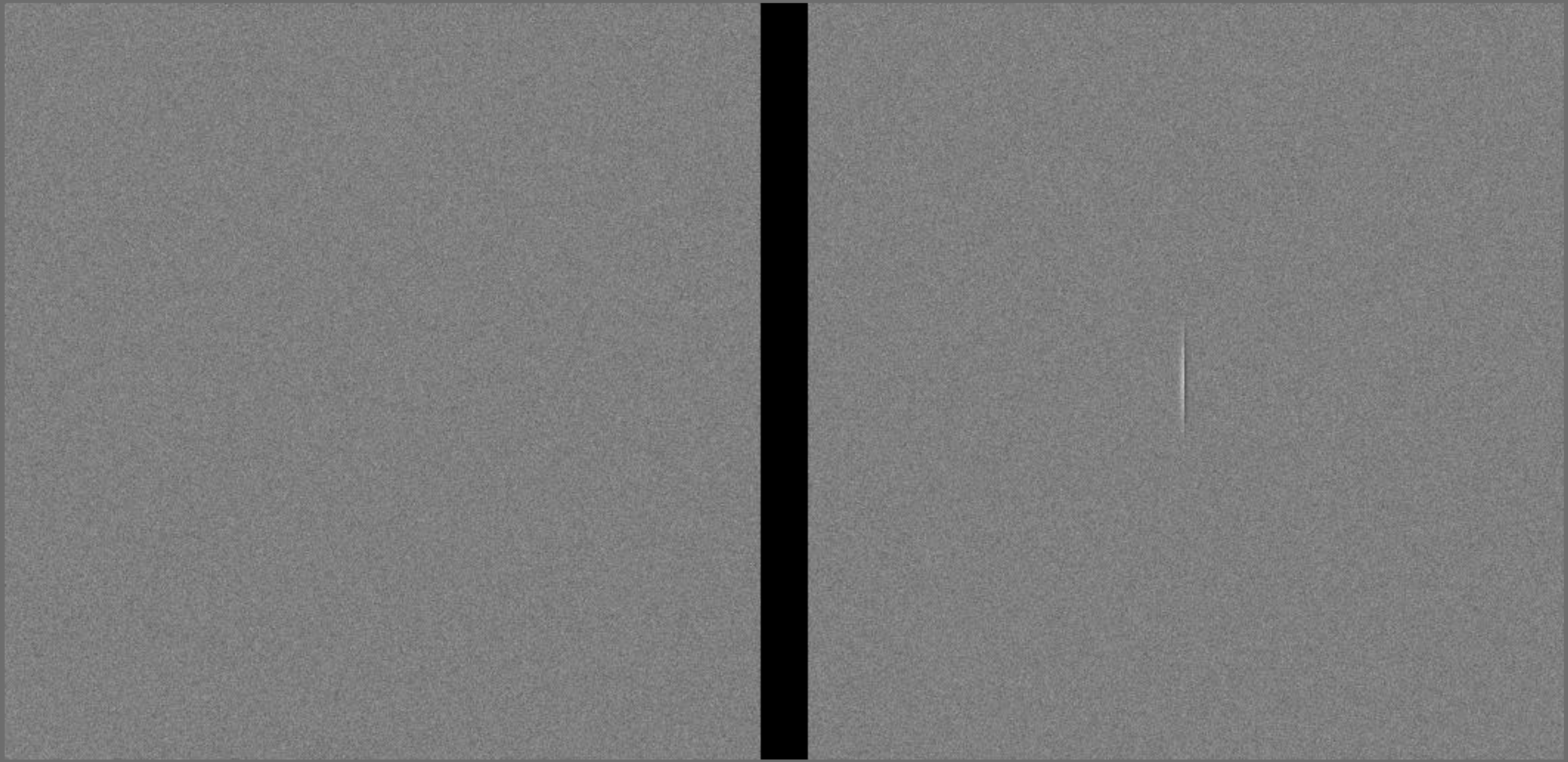


Kingsbury, Karunasekera, Yeh, Kokaram, Cambridge University

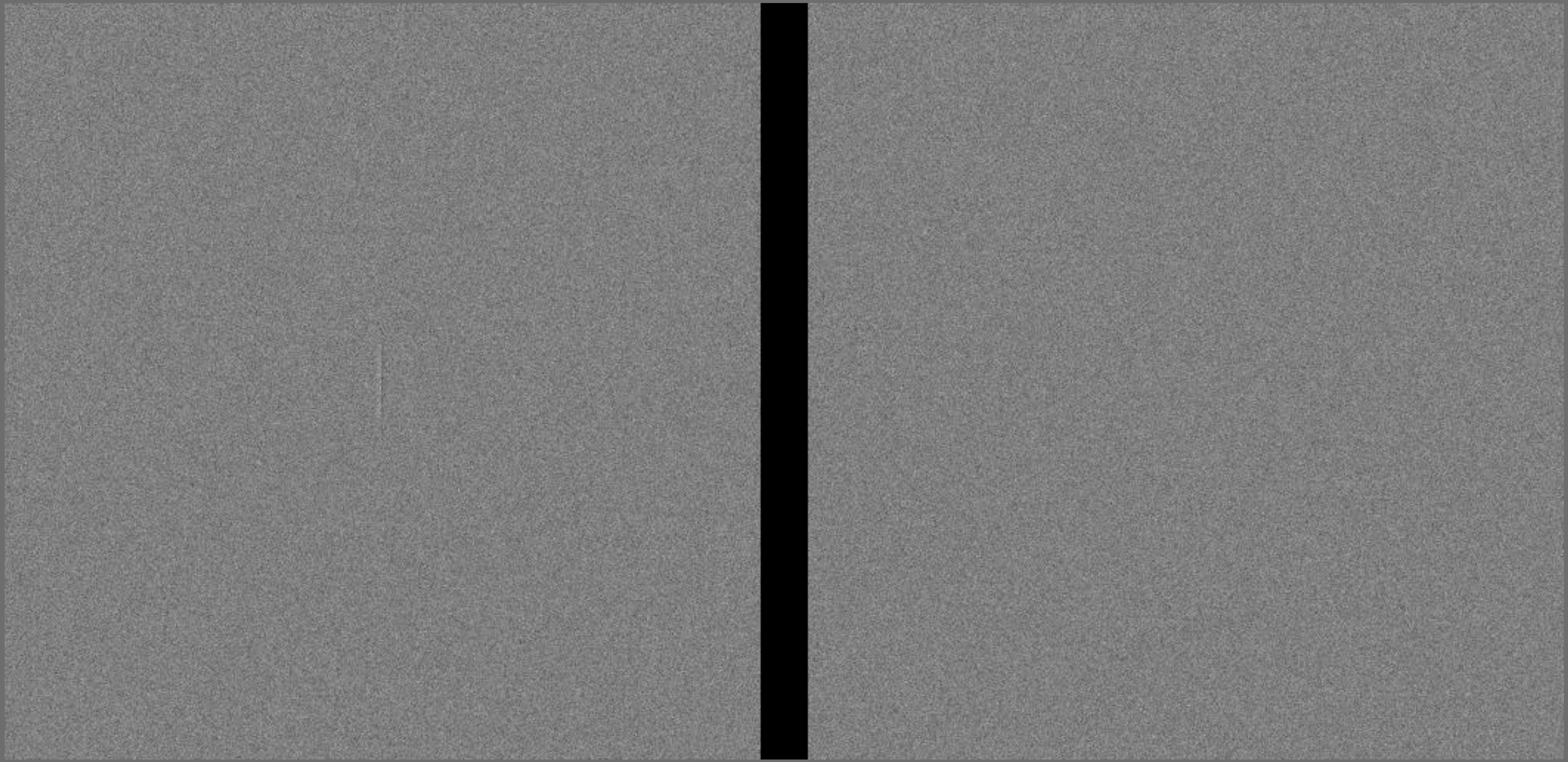
Reaction time experiments



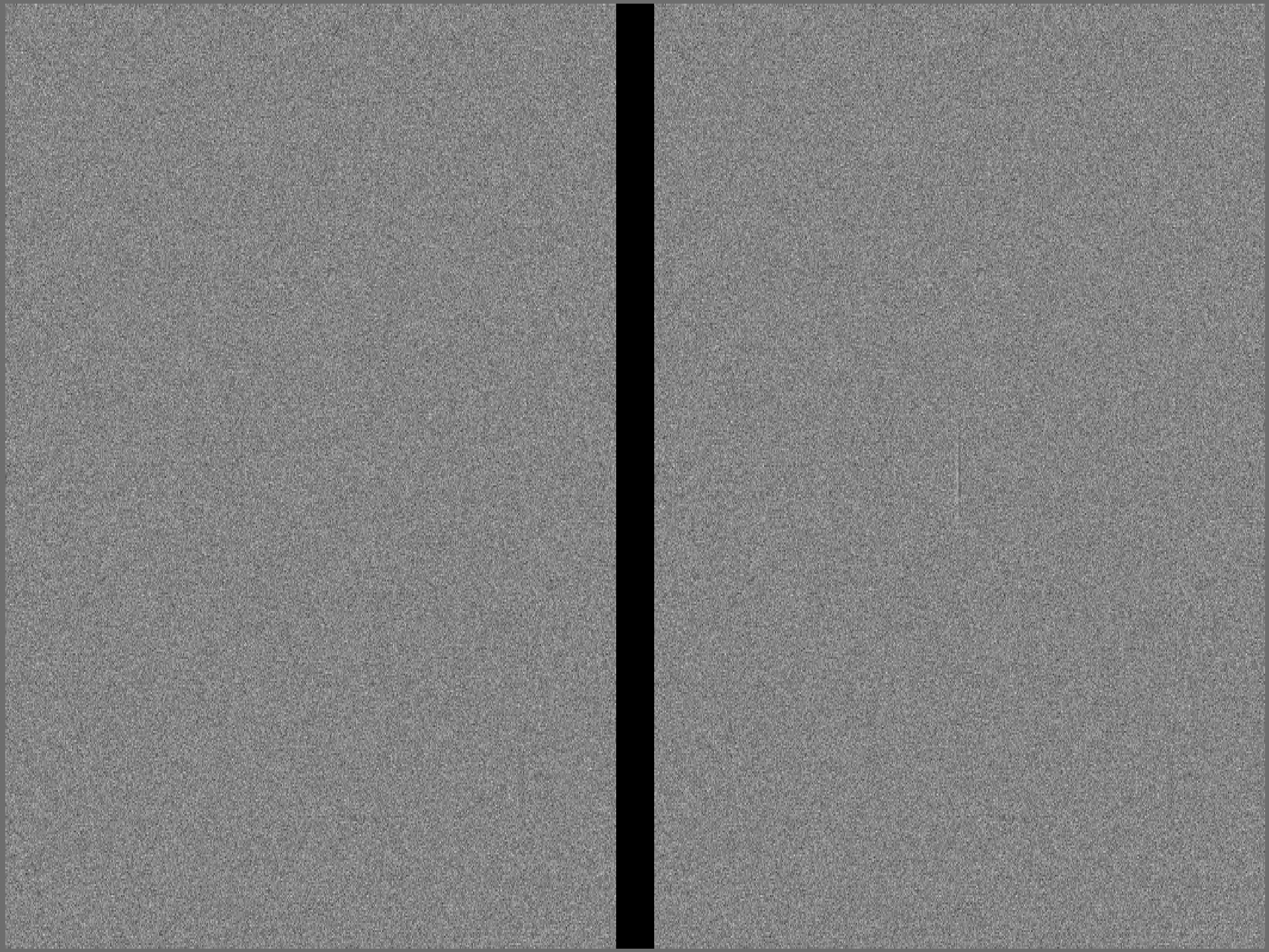
Edge Masking with noise



Which side? (10 sec timeout)



Motion perception/masking



Reaction time results

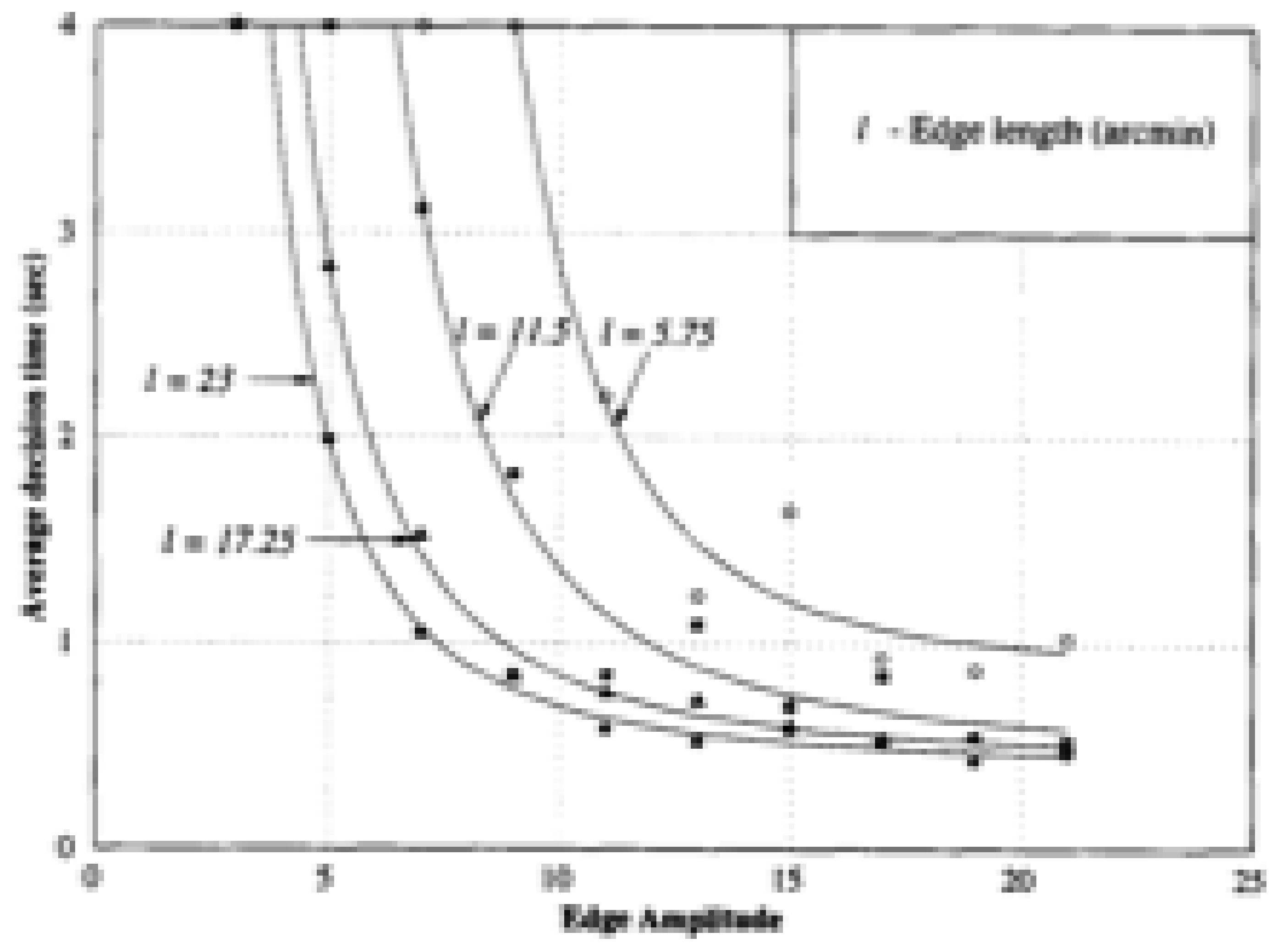
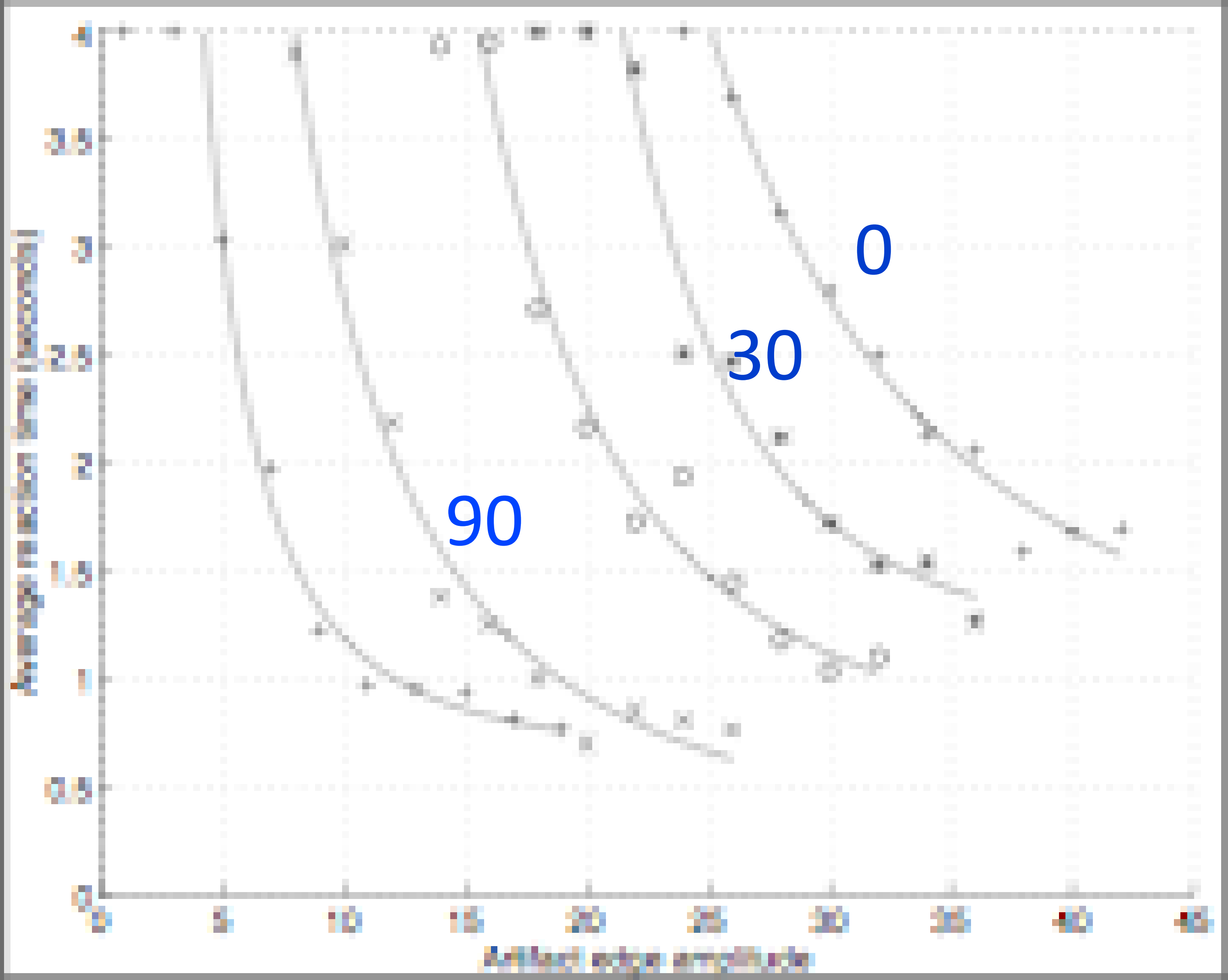
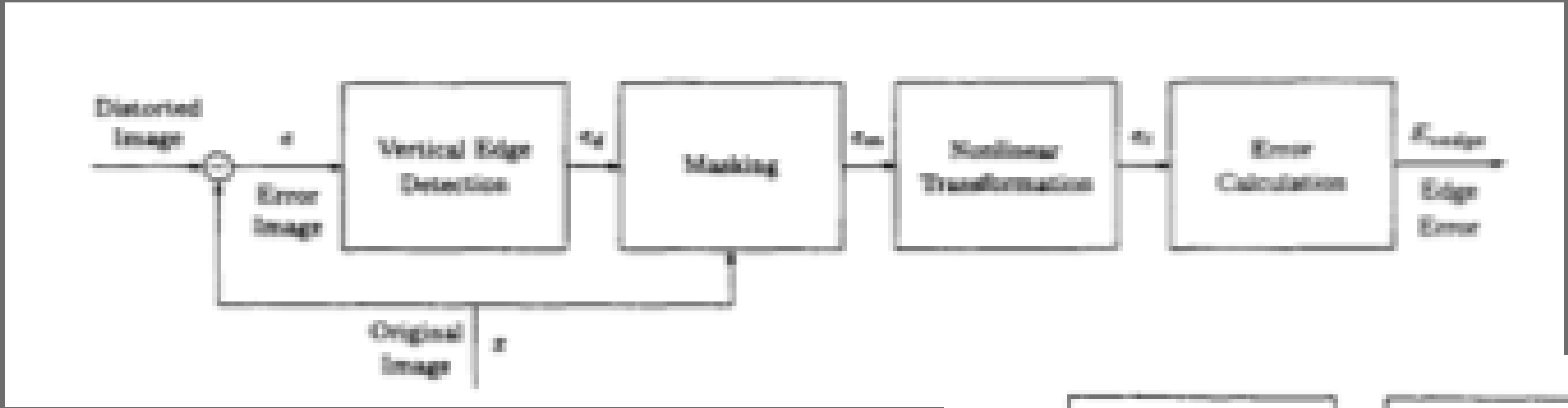


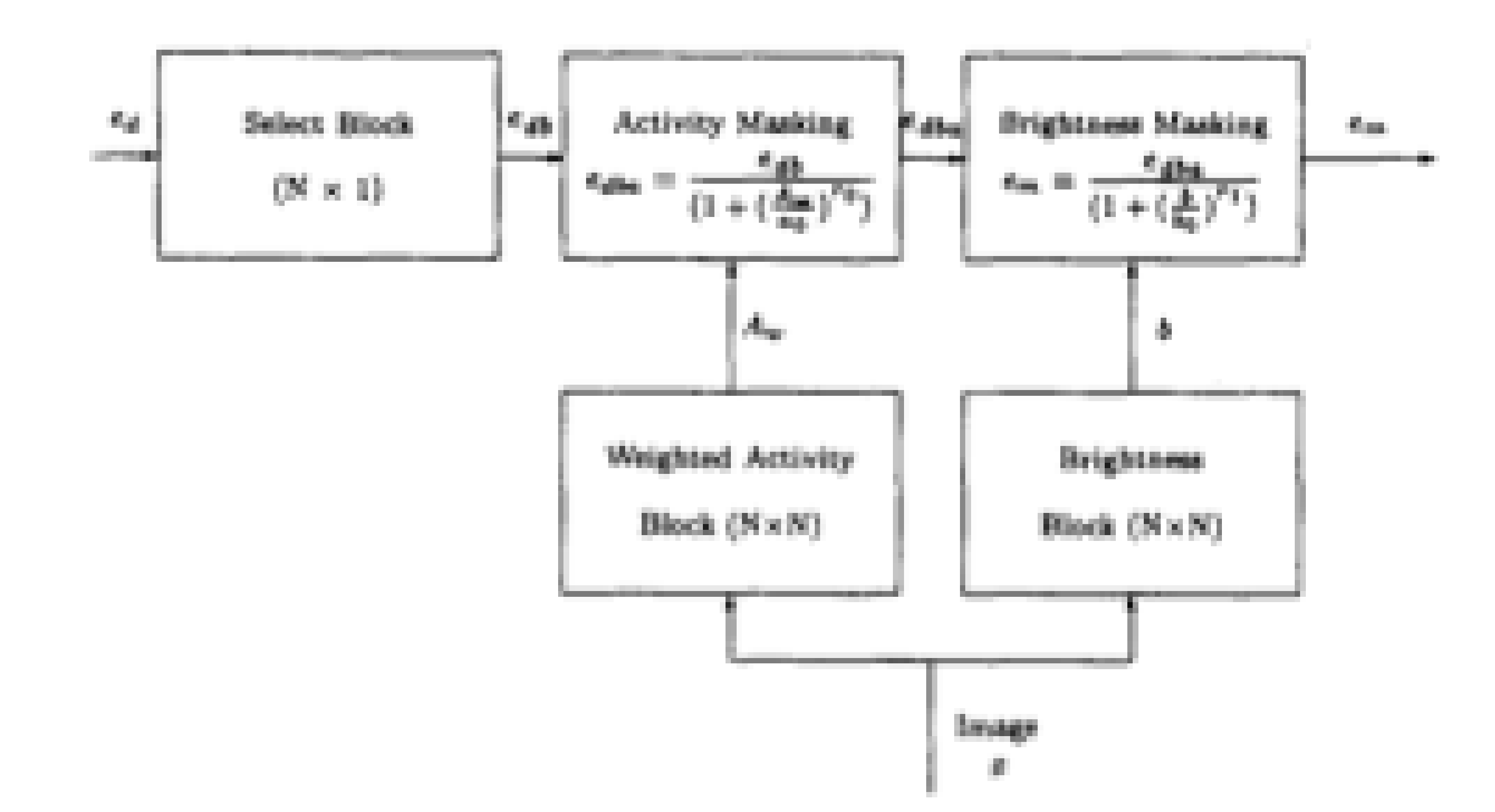
Fig. 6. Decision time versus edge amplitude for constant edge length.



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

TABLE II
RESULTS OF SUBJECTIVE TESTING

Picture	MSE	$E_{blocking}$	Subjective Ranking							
1	1.50	0.19	1	1	1	1	1	1	1	1
2	1.50	0.33	2	3	3	3	3	3	2	4
3	1.50	0.35	3	2	4	2	4	4	3	3
4	1.51	0.68	5	4	2	4	2	2	7	2
5	1.50	0.78	4	5	7	7	8	7	4	5
6	1.50	0.80	6	6	5	5	5	6	5	7
7	1.51	0.82	7	7	6	6	6	5	6	6
8	1.50	0.90	8	8	8	8	7	8	8	8
9	1.50	0.93	9	9	9	9	9	9	9	9
10	1.50	0.99	10	10	10	10	10	10	10	10
11	1.50	1.08	11	11	11	11	11	11	12	11
12	1.51	1.24	12	12	12	12	12	12	11	12

Done Interlude on earlier work

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 experiments
- How can YT records help?

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)

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.

Reading pictures in your mind

[Jack Gallant at Berkely]

Reference Material

1. “Digital Video and HD”, Charles Poynton, Morgan Kaufman Publishers, 2012
2. “Video Demystified”, Keith Jack, 2011
3. “The Art of Digital Video”, John Watkinson, Focal Press, 2000
4. Digital Image Quality and Perceptual Coding, Taylor and Francis Pub., 2006
5. Fundamentals of Image Processing, Anil Jain, Prentice Hall
6. Two Dimensional Image and Signal Processing, Jae Lim, Prentice Hall
7. http://www.poynton.com/notes/misc/Poynton-square-pixels.html

- http://www.fxphd.com/