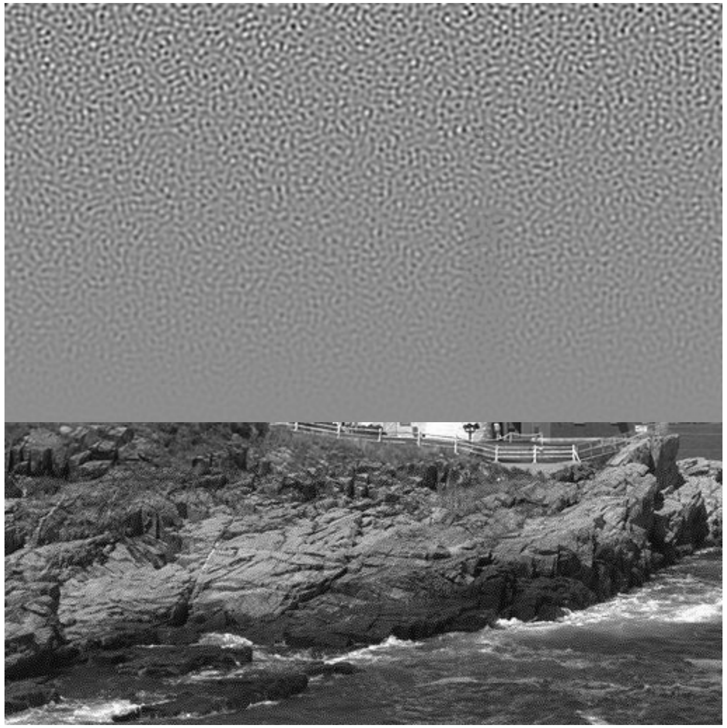
Image Quality Assessment

Outline

- Motivation
- Perceived quality
- Image distortions
- Assessment methods
  - Subjective experiments
  - Objective metrics
- Metric evaluation

Motivation

- Same amount of distortion, yet different perceived quality



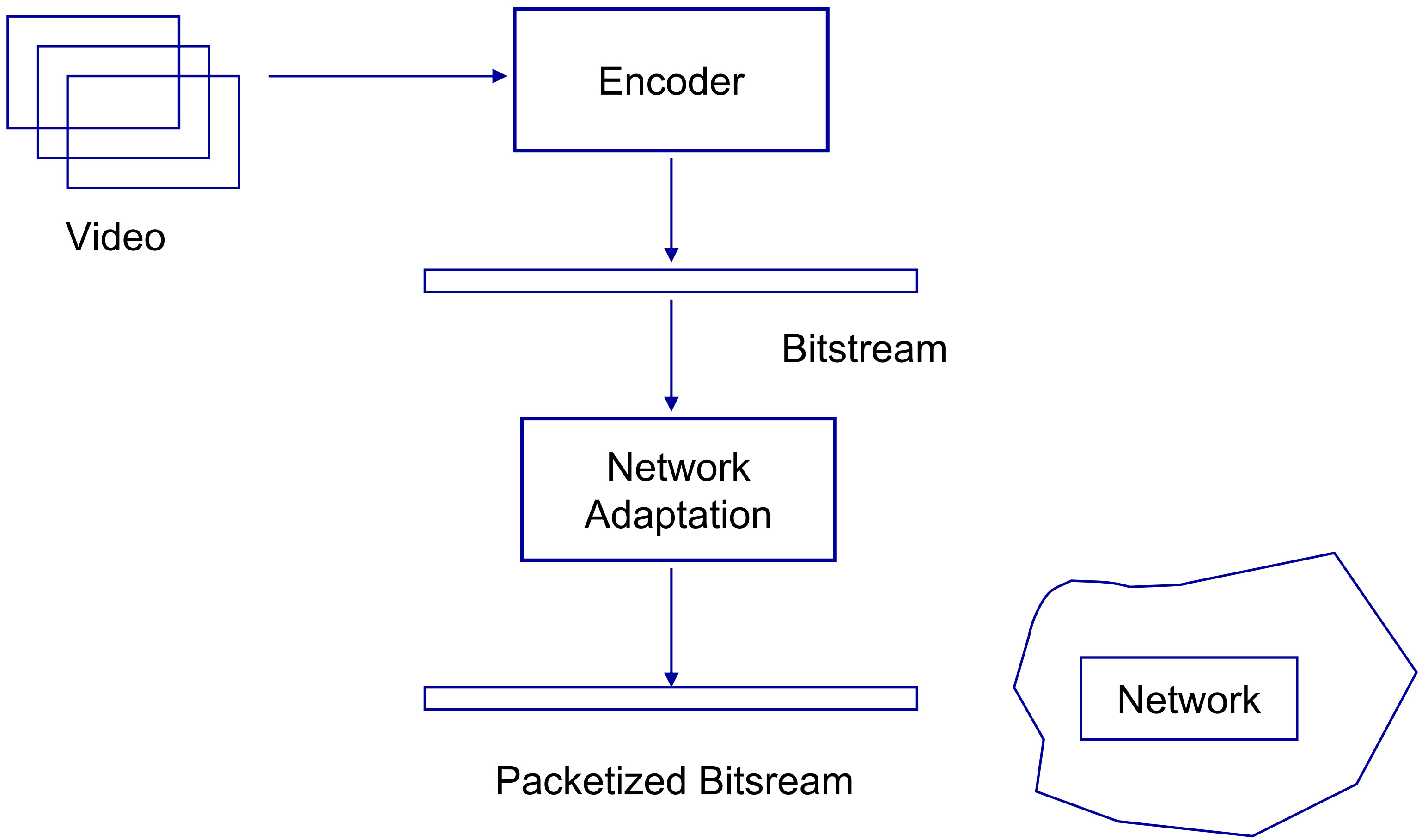
Perceived Visual Quality

- Subjective factors
  - Semantics (interest in the content)
  - Expectation
  - Experience
- Display properties
  - Type (paper, projection, CRT, LCD,...)
  - Resolution and size
- Viewing conditions
  - Distance from display
  - Lighting conditions

Perceived Visual Quality

- Visual factors
  - Fidelity of reproduction
  - Brightness
  - Contrast
  - Sharpness
  - Colorfulness
- Two-way communication
  - Delay
- Soundtrack
  - Synchrhonization
  - Quality of interactions

Transmission System



Image/Video distortions

- Pre- or post-processing
  - D/A-A/D conversion
  - De-interlacing
  - Frame rate conversion
- Lossy compression
  - Quantization, motion prediction
  - Blockiness, loss of details, noise, ...
- Transmission over noisy channels
  - Bit errors, packet loss
  - Video freeze (jerkiness)
  - Error propagation

JPEG artifacts

JPEG 2000 artifacts

Transmission Errors

JPEG/MPEG

JPEG 2000

BER 10^{-5}

BER 10^{-4}

Artifacts Summary

- Spatial effects
  - Blockiness
  - DCT basis image
  - False contours
  - Staircase effect
  - Ringing
  - Bluriness
  - Color bleeding
- Temporal effects
  - Jerkiness
  - Motion compensation mismatch
  - Mosquito noise
  - Motion blur
  - De-interlacing

Quality Assessment Methods

Objective quality metrics

- Bit-based
  - MSE, PSNR
- Models of the Human Visual System (HVS)
- Specialized artifact metrics
  - Blockiness
  - Blurriness

Subjective quality assessment

- Reference & benchmark
- Standardized procedures
- Many observers, careful setup
- Time consuming, expensive
- *Psychometric scaling*

Psychometric Scaling

- Customer perceptions: the *nesses*
  - ness : perceptual attribute, a sensation risen by an image feature (attribute)
- Image quality models
  - Link the customer's perception (nesses) with image quality *measures*
- Scaling
  - Measuring image quality based on the customer's perception of the nesses and quantify it by some indicators (numbers, labels, relative/absolute ratings)
  - Different scaling methods are suitable for different frameworks and/or evaluation tasks

Scaling

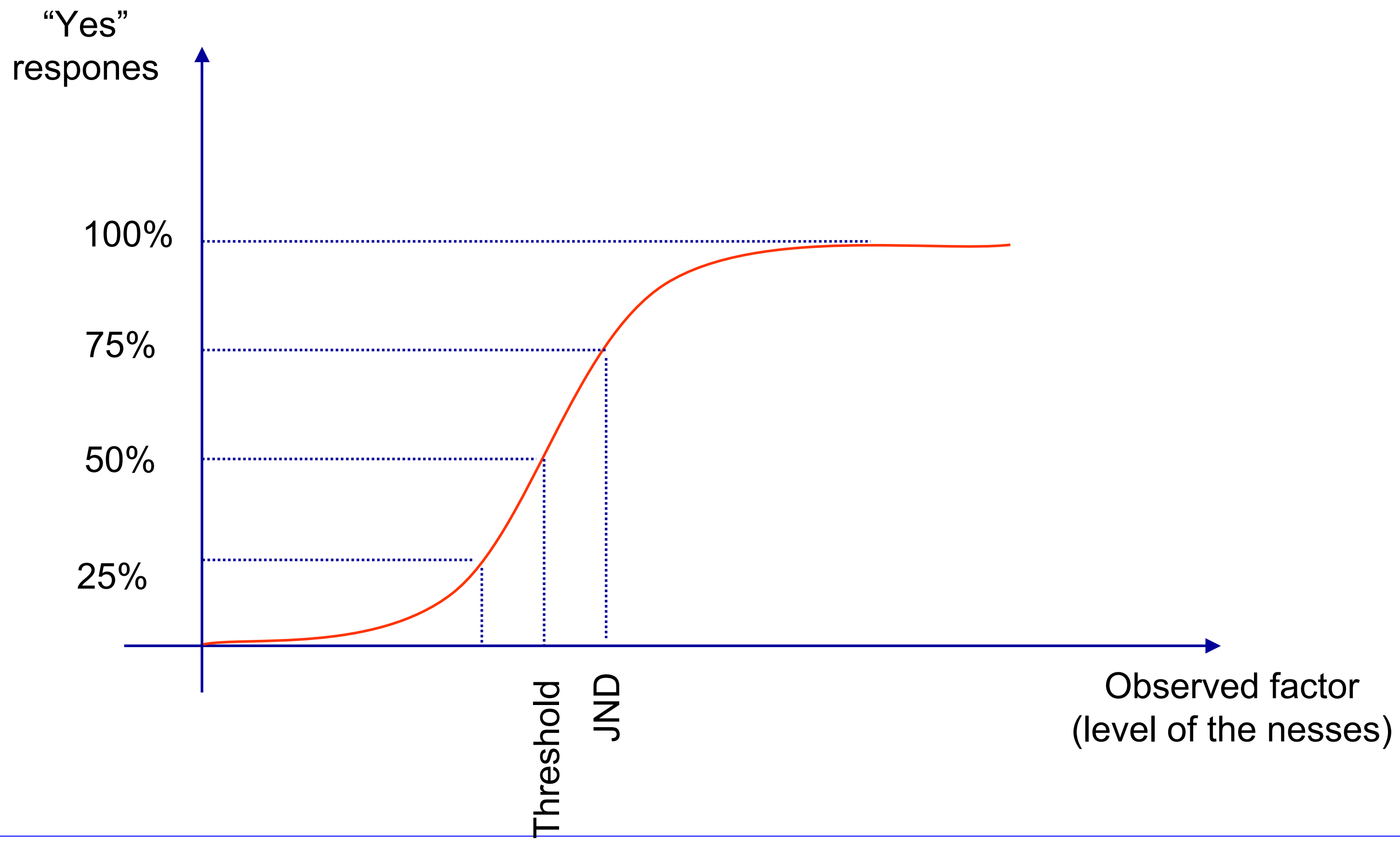
1. Select the samples
2. Prepare the samples for observer judgment
3. Select the observers
4. Determine observer judgment task or question
5. Present samples to observers
6. Collect and record observer responses
7. Analyze observer’s response data to generate the scale values

Basic concepts

- Threshold
  - “Is it visible or not?”
- Just-noticeable difference
  - “Can you distinguish them?”
- Psychometric model
  - The responses are accumulated over a number of observers
    - The observer’s responses vary even when the stimulus is held constant
  - Goal: estimation of the probability distribution of the responses
    1. Measure the empirical cumulative histogram of the responses
    2. Fit a *psychometric model* to such data
    3. Deduce some parameters
       1. Absolute thresholds
       2. Just Noticeable Differences (JND)

Psychometric Function

- Also *frequency of seeing* curve



Threshold and JND

- Stimulus threshold: smallest amount of "ness" needed to produce an awareness of the ness
  - It is usually taken as the point where 50% of the observers "see" the ness
- Stimulus JND: stimulus change required to produce a just noticeable difference in the perception of the ness. Also called difference thresholds or increment thresholds.
  - The JND depends on the stimulus level and is proportional to its value.
  - It is defined as the ness value where the 75% of the observers see a stimulus with a ness greater than the standard

Methods

- Method of limits (PEST, QUEST)
- Method of adjustment
- Method of constant stimuli
- Forced-choice methods (2AFC)

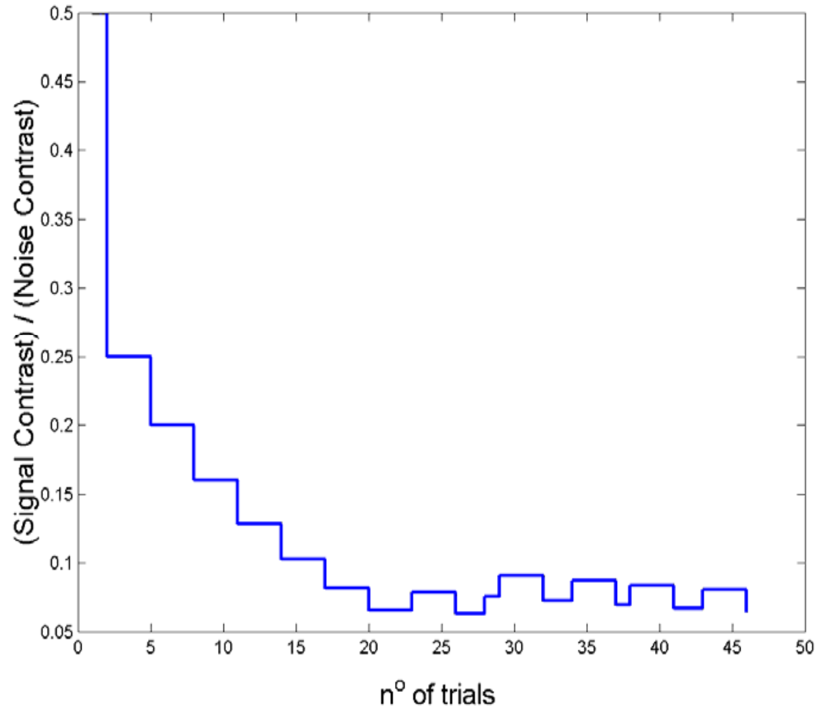
They differ in the way the stimuli are presented and the data are analyzed

Method of limits

- Guideline
  1. Start the sequence of presentation with one that does not have the ness perceptible, and keep increasing the ness until the observer detects its presence
  2. At that point the ness value is recorded and
  3. The presentations are repeated staring from a stimulus where the ness is clearly visible and keep decreasing it until it is no longer detectable
  4. After a large number of observers, the experimental proportions are estimated
- Absolute threshold
  - Do you see it?
- JND
  - Is it different from the standard?
  - Both the standard and the test stimuli must be presented simultaneously to the observer

Method of limits

- Up and down staircaise method
  - Breaks the monotonicity of the nesses
  - Double staircaise



- Issues
  - Where to start the ness sequence?
  - Initial ness size?
  - When to stop collecting data?
  - Modification of step sizes

Method of adjustment

- The observer adjusts the ness by turning a knob, moving a slider or using another control method
  - Advantage: active involvement of the subject, which improves the quality of the data
  - Disadvantage: only possible for simple continuously tunable nesses
- Guideline
  - The subject adjusts the level of the ness until it is just visible (for an absolute threshold measurement) or until it matches the standard (for JND measurements)

Method of constant stimuli

- The “contant” is a selected set of samples “stimuli” that remain fixed throughout the experiment
  - The set of samples is usually chosen such that the sample member with the lowest level of ness is never selected by the users, while the one with the highest ness level is always selected by all the observers
  - Needs a pilot experiment
  - Results in an experimental psychometric curve
- Absolute threshold
  - Stimuli are presented in random order
- JND
  - The test and reference stimuli are presented together

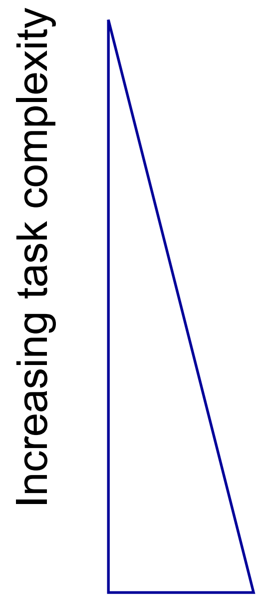
Sample ID	Ness value	pi=fi/N
A	x1	f1/N
B	x2	f2/N
C	x3	f3/N

⇩

psychometric curve

Subjective Assessment

- Nominal scales
  - Attach labels
- Ordinal scales
  - Put into order (more than or less than)
  - Problem: we don’t know how close a sample is to the adjacent one
- Interval scales
  - Add the property of distance to an ordinal scale
  - Quantify distance/level
  - Equal differences in scale values correspond to equal differences in nesses
- Ratio scales
  - Interval scale with origin (distance from zero)



Common Scaling Methods

- Ordinal Scaling
  - Rank-order
    - The subject is asked to order the stimuli according to the ness level
  - Paired comparison
    - The subject has to compare couples of stimuli (time consuming)
  - Category scaling
    - The subject is asked to gather the stimuli into categories
    - Categories can be names like “good” or “bad”, numbers....
- Direct interval scaling
  - Graphical rating scale
- Indirect interval scaling
  - Paired comparisons – Thurston’s Law of Comparative Judgement
  - Category scaling – Torgerson’s Law of Categorical Judgment

Video Quality Assessment

- ITU-R Rec. BT.500 (television)
  - Double Stimulus Impairment Scale (DSIS)
  - Double Stimulus Continuous Quality *Scales* (DSCQS)
  - Double Stimulus Continuous Quality *Evaluation* (SSCQE)

- ITU-T Rec. P.910 (multimedia)
  - Absolute category rating
  - Degradation category rating (~DSIS)
  - Pair comparison

Double Stimulus Impairment Scale (DSIS)

- Method
  - Reference & processed sequence are shown

Reference		Processed

  - Viewers rate degradation on discrete scale

- ❑ Umpercettible
- ❑ Perceptible but not annoying
- ❑ Fair
- ❑ Poor
- ❑ Bad

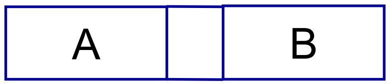
- Properties
  - Short sequences (memory effect)
  - Large degradation with respect to reference
  - Scale marks not equidistant

Double Stimulus Continuous Quality Evaluation (DSCQE)

- Method
  - No explicit reference shown
  - Viewers constantly rate *instantaneous* quality on a continuous scale using slider
  - Slider position is sampled regularly
- Properties
  - Long sequences
  - Efficient data collection
  - Captures quality variations
  - More “realistic” setup
  - Higher inter-subject variability
  - Response latency

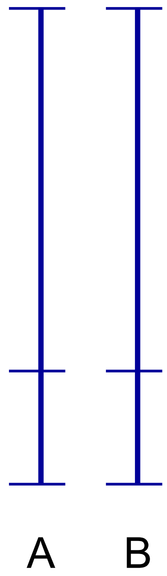
Double Stimulus Continuous Quality Scales (DSCQS)

- Method
  - Reference & processed sequence are shown



  - Viewers rate both on a continuous scale from “bad” to “excellent” (0-100)
  - Difference is recorded

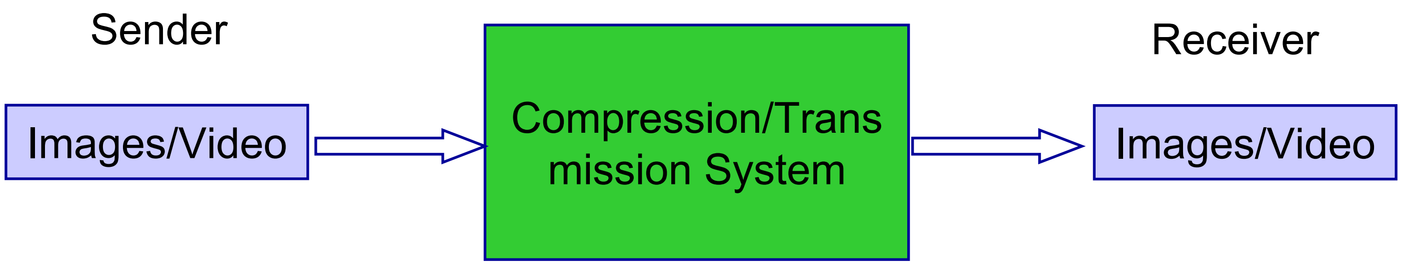
- Properties
  - Content effect reduced
  - Fine distinctions possible
  - Reference can be rated worse than processed



ITU Recommendations

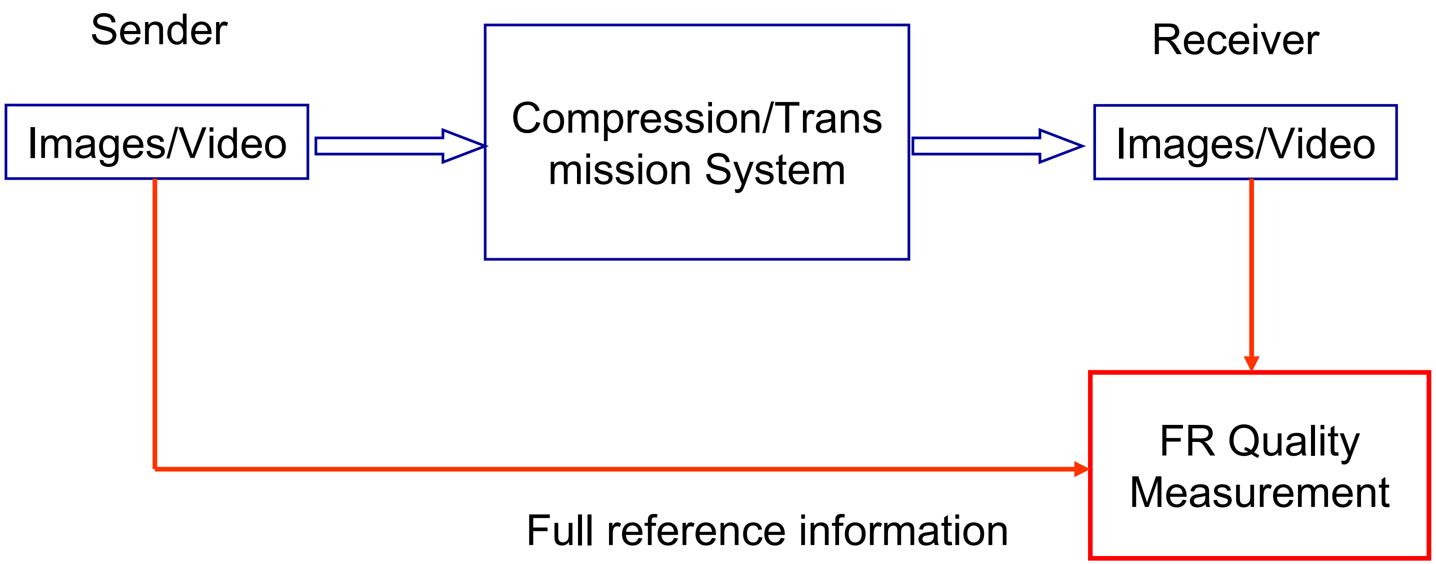
- Experimental conditions
  - Display properties and setup
  - Illumination
  - Distance from the screen
- Observers
  - >15
  - Experts vs. non-experts
  - Vision tests
  - Instructions
  - Training
- Sample selection
  - Application
  - Test method
  - Content
- Data analysis
  - Data collection
  - Data processing
  - Observer screening

Objective Quality Metrics

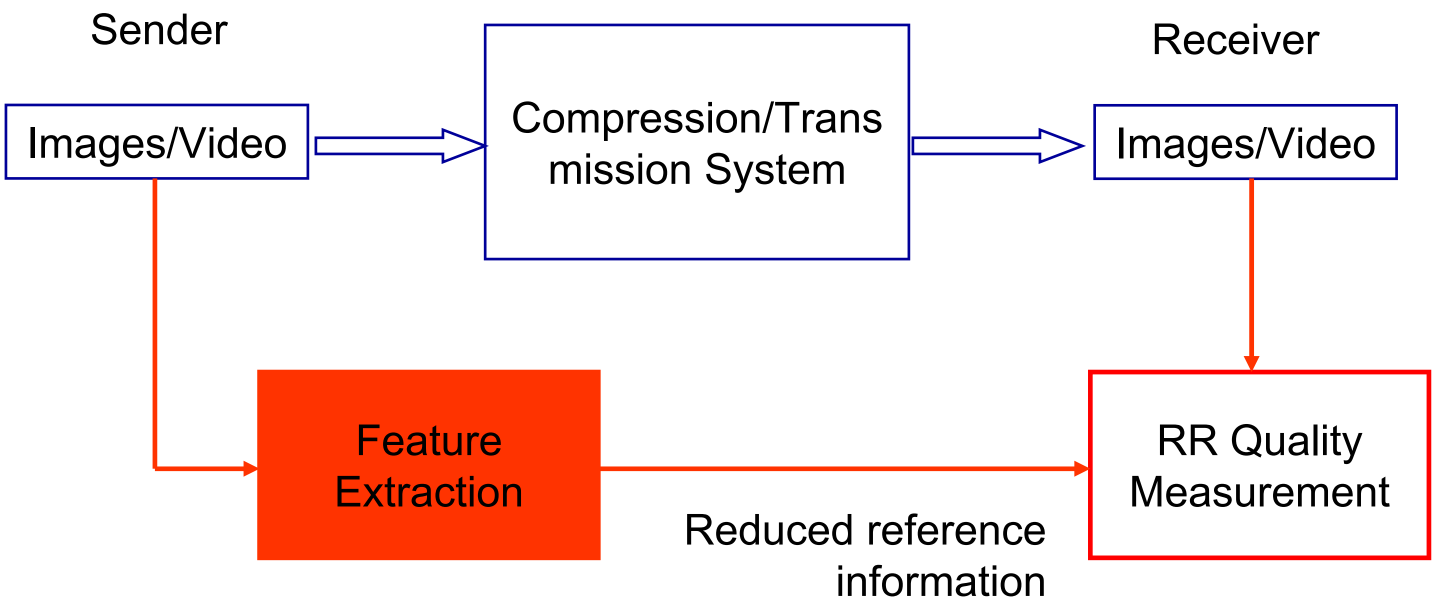


- Issues
  - Quality?
  - Relative or absolute?
  - Intrusive or not?

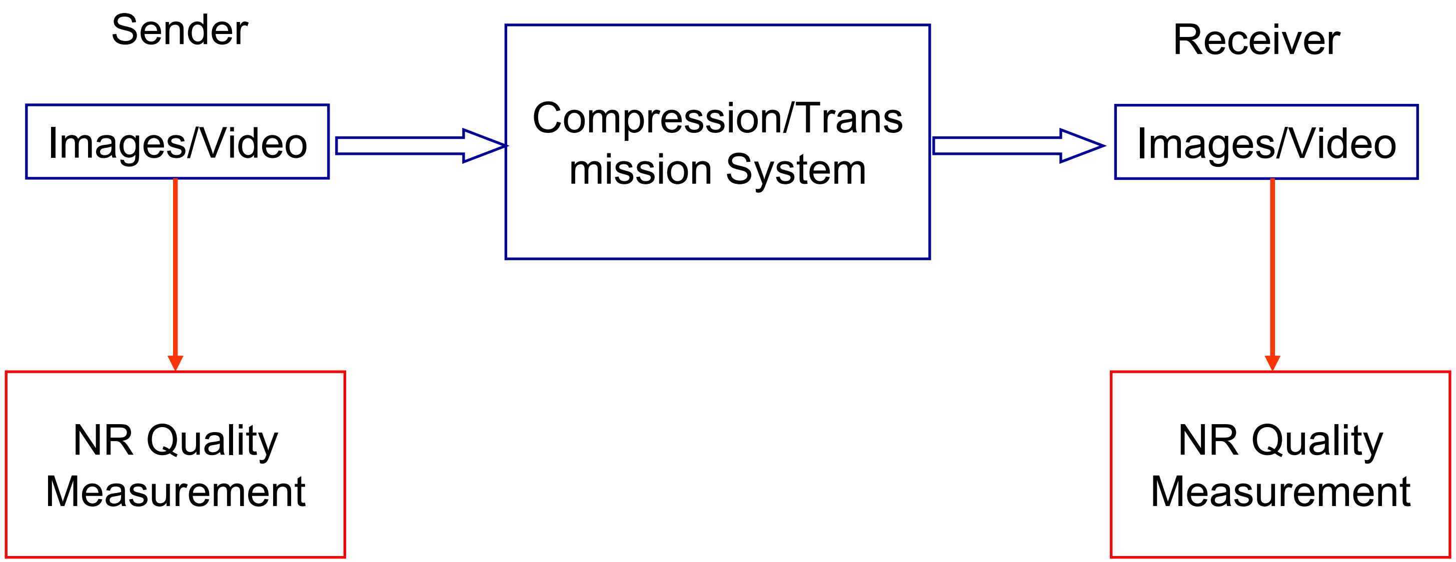
Full-Reference Metric



Reduced Reference Metric



Non-Reference Metric



Quality Metric Applications

- Automatization of all the visual evaluation tasks

- Quality monitoring (QoS for multimedia)
- Quality control

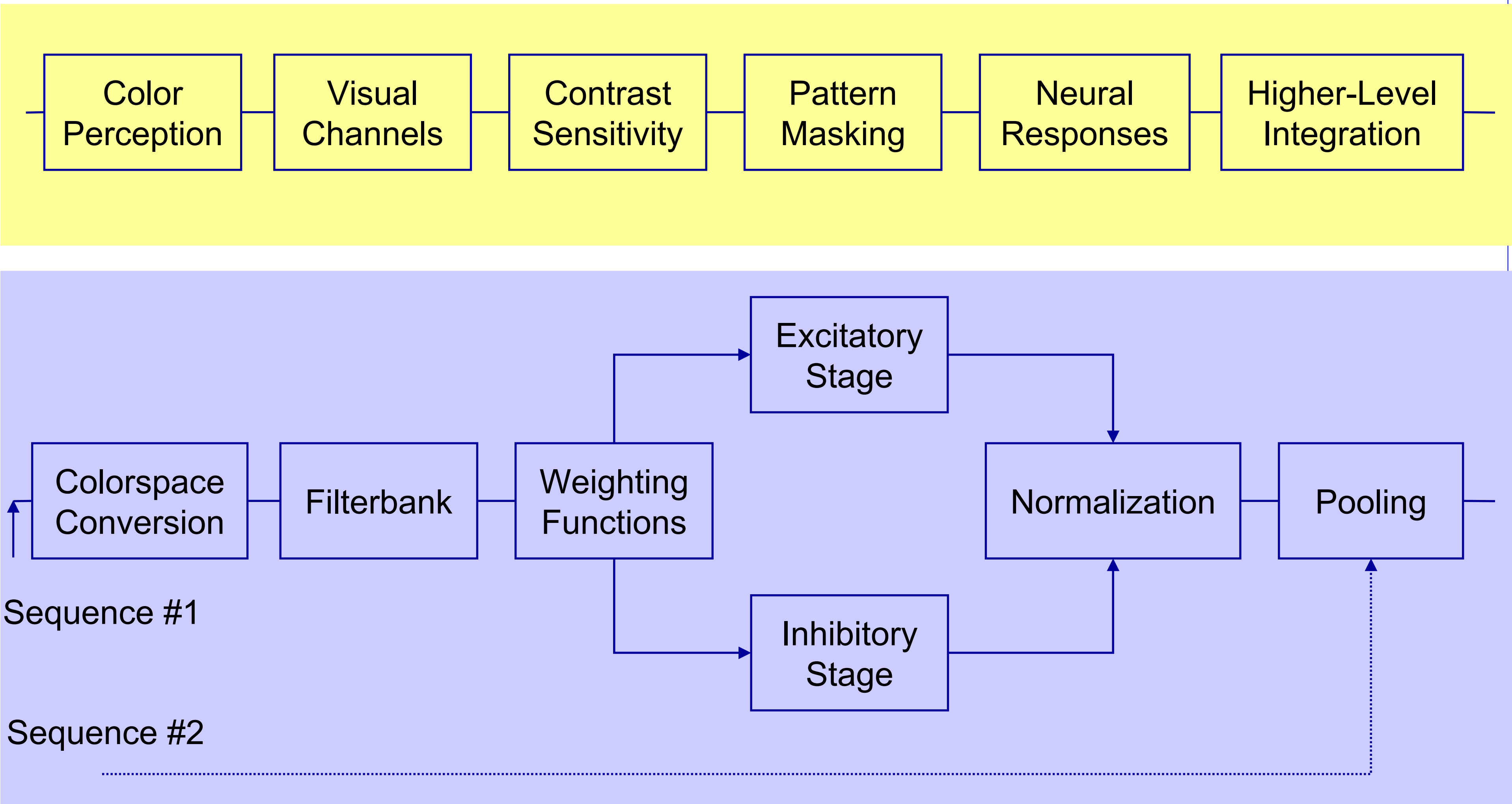
- Codecs evaluation and comparison
- Watermarking
- Restoration
- Denoising
- ...

Bit-based Metrics

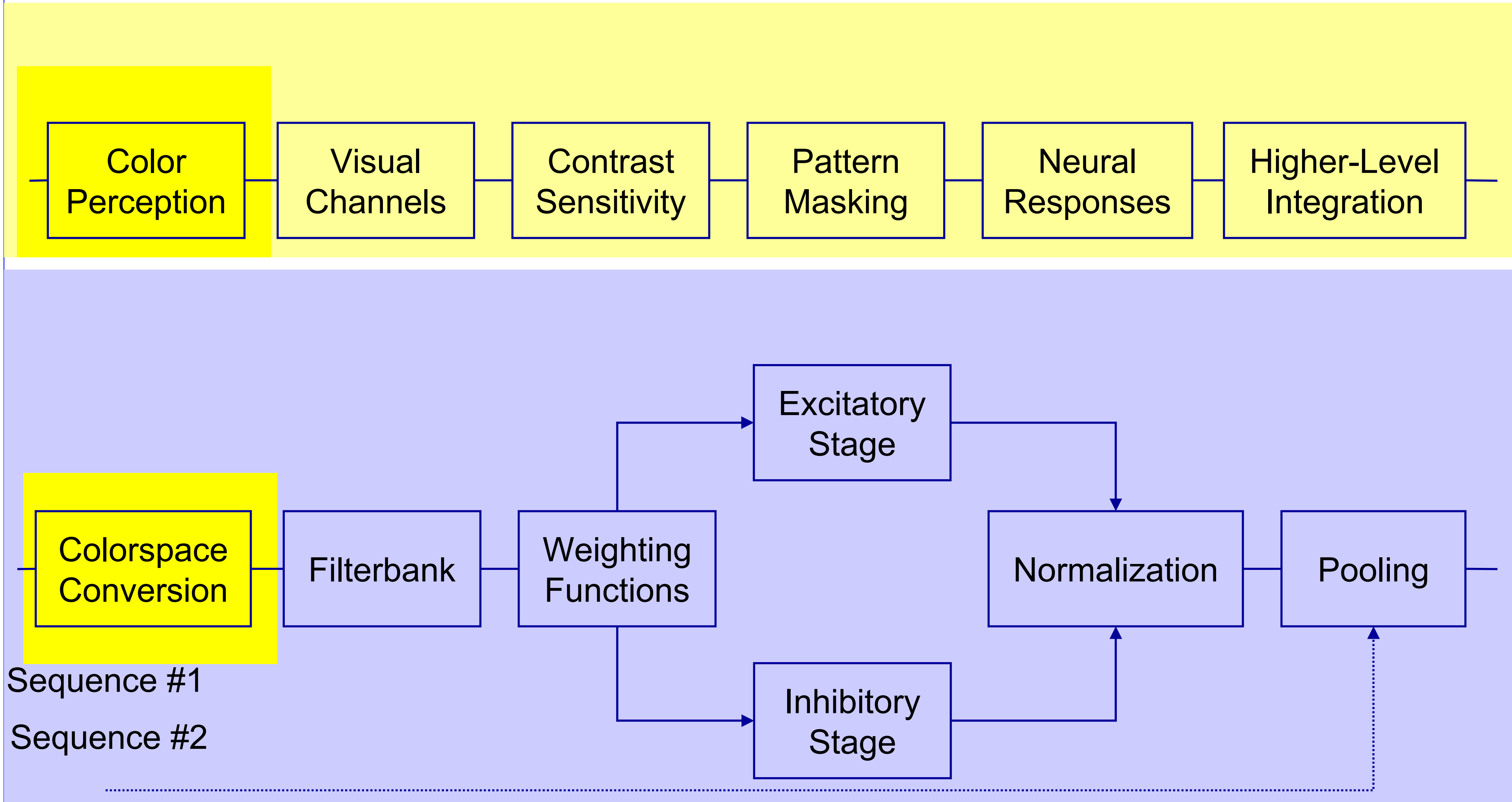
- PSNR/MSE
  - Quantify the difference to reference Images/Videos
  - Pixel-based
  - Content independent
  - Mediocre quality predictors
  - Not representative of visual perception

- Network QoS
  - Bit error rate (BER), packet loss..
  - Bit/packet-based, content independent
  - Meaningless without perception

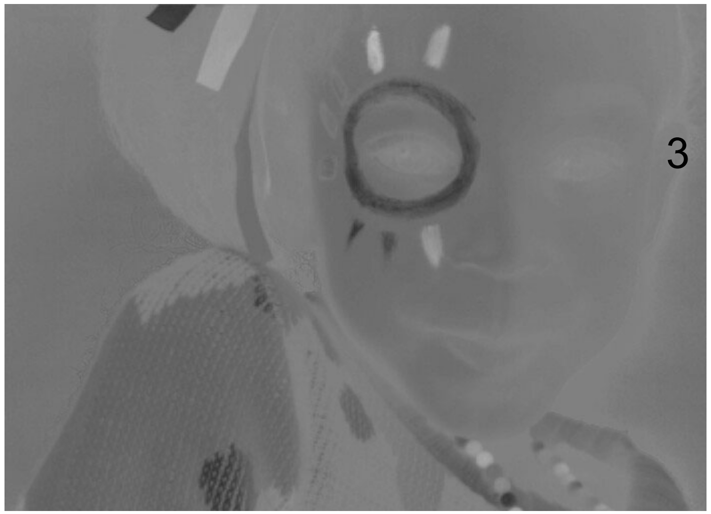
Vision-based metrics



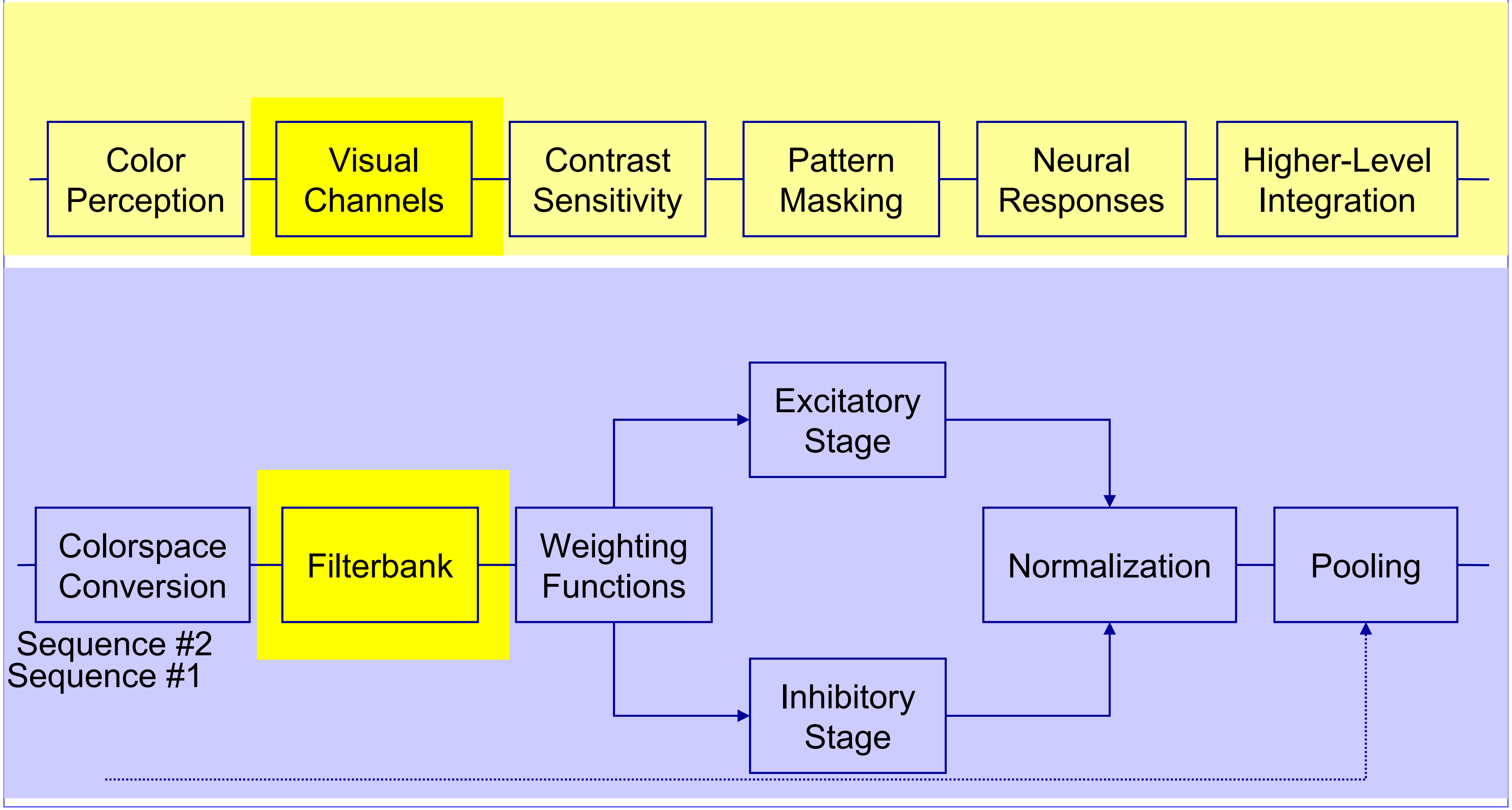
Typical Vision Model



Opponent Colors



Typical Vision Model



Visual Channels

Issues	Number of mechanisms	Position	Bandwidth
Temporal frequency	2-3	0 Hz 8 Hz	8 Hz 2 Hz
Spatial frequency	4-6	1-15 cpd	1-2 octaves
Orientation	4-8		20 °-60°

DB scales

In every kind of dB, a *factor of 10* in amplitude increase corresponds to a *20 dB boost* (increase by 20 dB):

$$20\log_{10}\left(\frac{10\cdot A}{A_{\mathbf{ref}}}\right) = \underbrace{20\log_{10}(10)}_{20\ \mathrm{dB}} + 20\log_{10}\left(\frac{A}{A_{\mathbf{ref}}}\right)$$

and $20\log_{10}(10) = 20$

A function f(x) which is proportional to 1/x is said to ``fall off'' (or ``roll off'') at the rate of 20 *dB per decade*. That is, for every factor of 10 in x (every ``decade''), the amplitude drops 20 dB.
Similarly, a factor of 2 in x amplitude gain corresponds to a 6 dB boost:

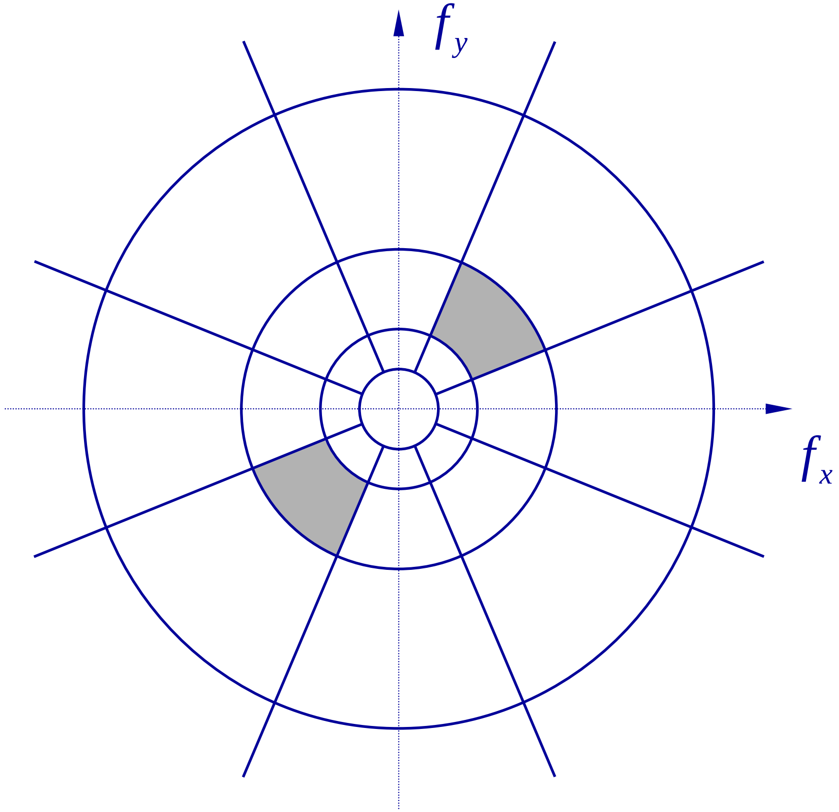
$$20\log_{10}\left(\frac{2\cdot A}{A_{\mathbf{ref}}}\right) = \underbrace{20\log_{10}(2)}_{6\ \mathrm{dB}} + 20\log_{10}\left(\frac{A}{A_{\mathbf{ref}}}\right)$$

$$20\log_{10}(2) = 6.0205999\ldots \approx 6$$

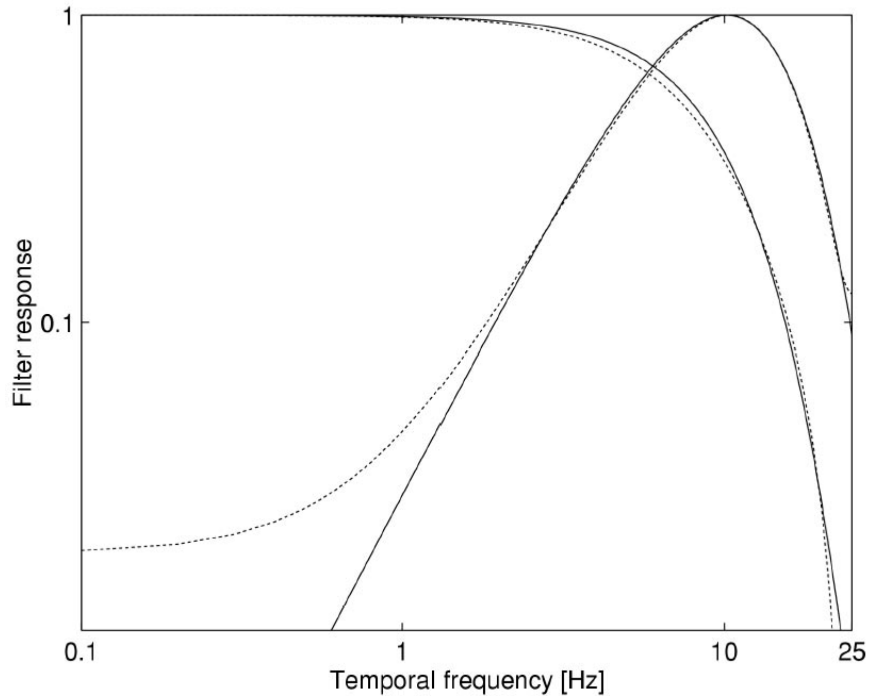
A function f(x) which is proportional to 1/x is said to fall off 6 *dB per octave*. That is, for every factor of 2 in x (every ``octave''), the amplitude drops close to 6 dB. Thus, 6 dB per octave is the same thing as 20 dB per decade.

Perceptual Decomposition

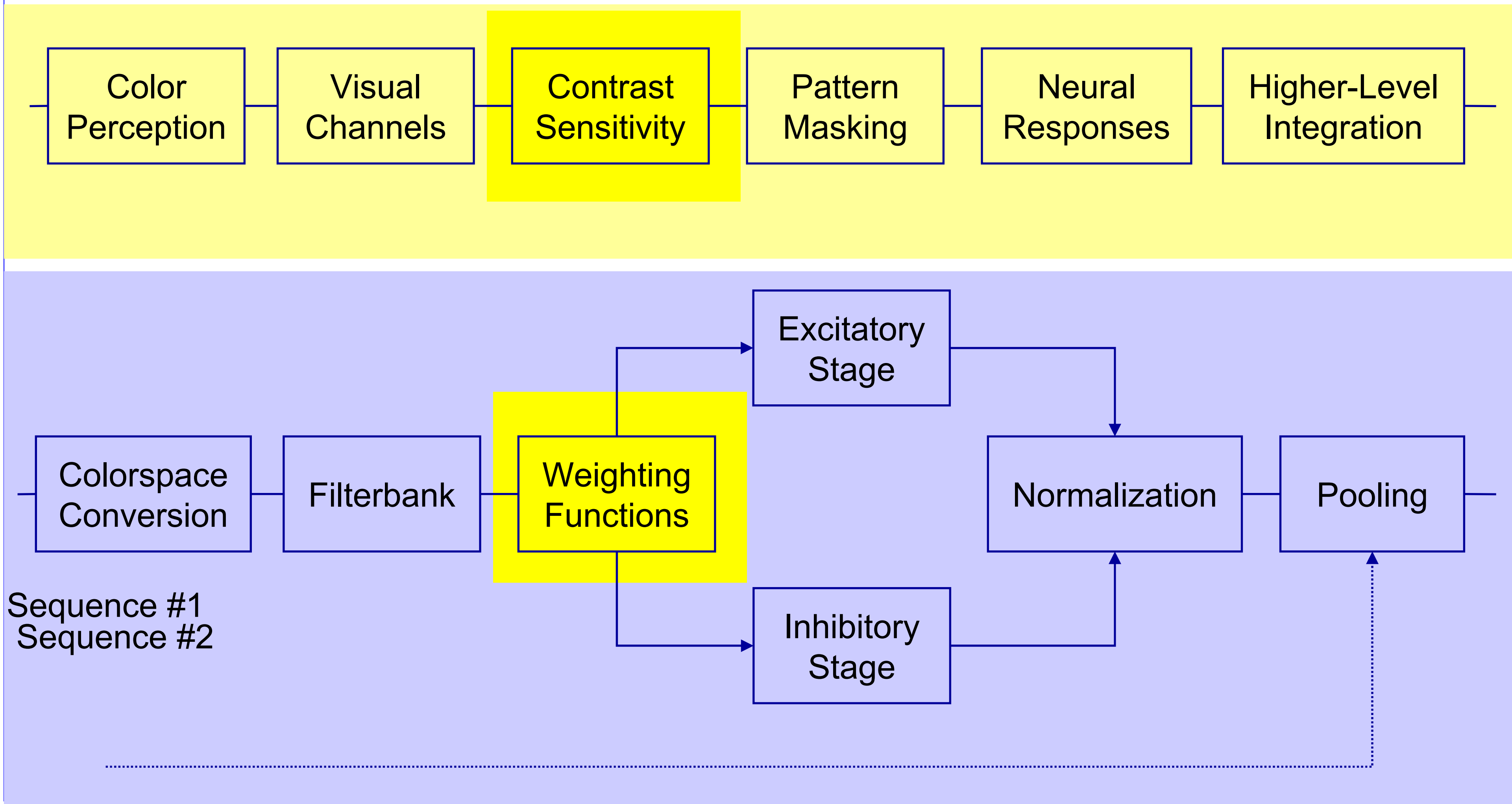
- Spatial mechanisms



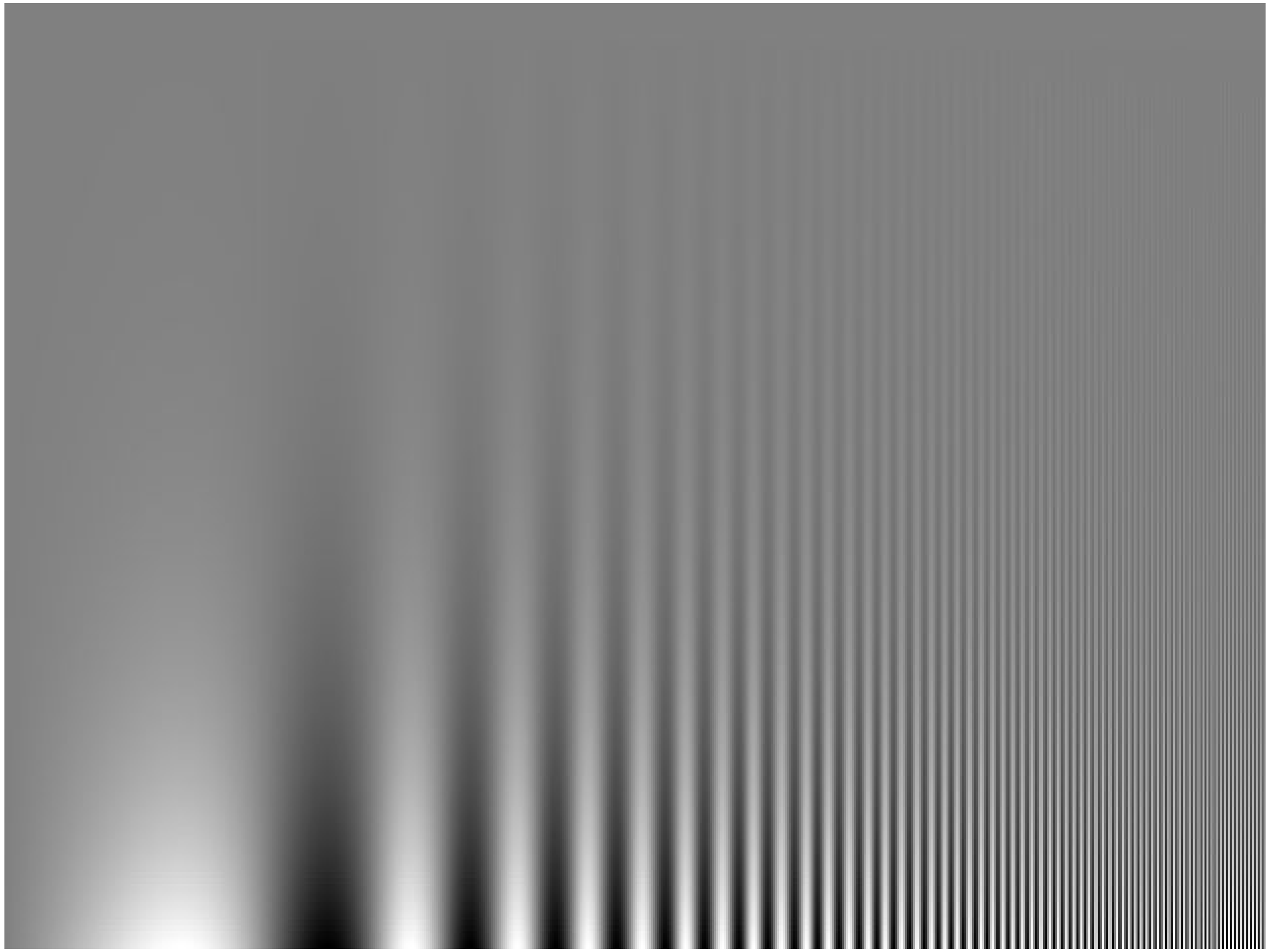
- Temporal mechanisms



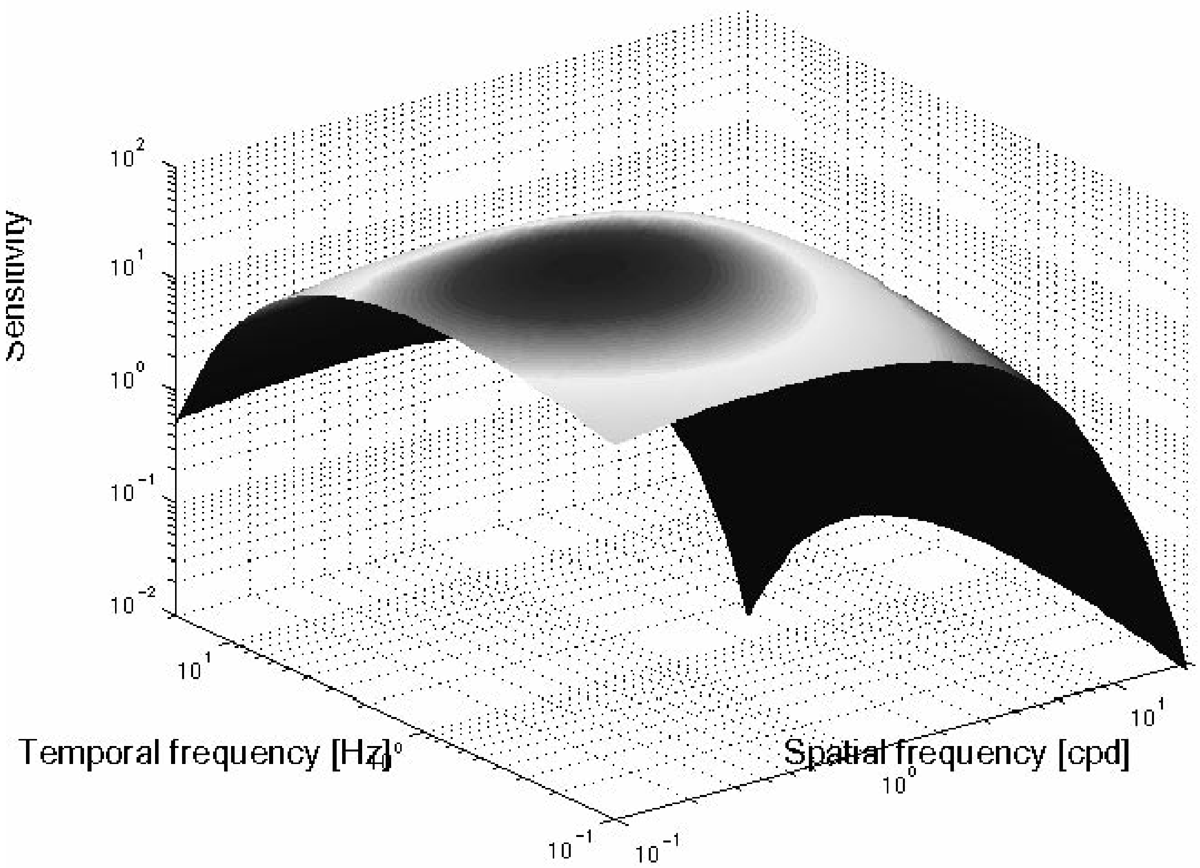
Typical Vision Model



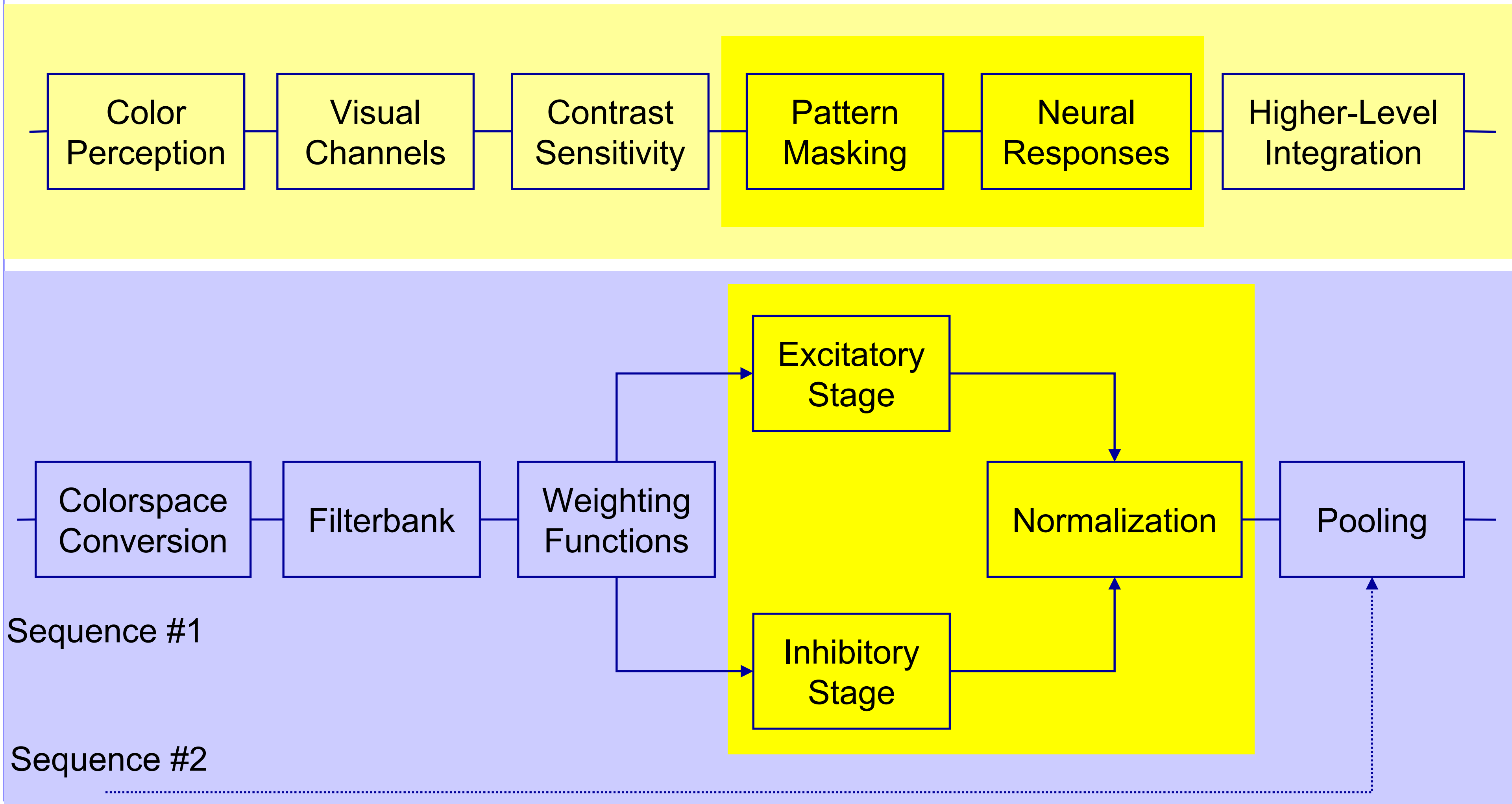
Contrast Sensitivity



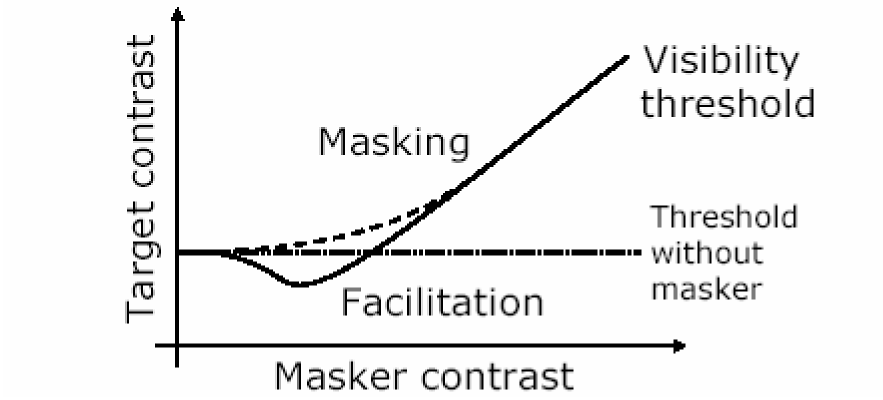
Contrast Sensitivity Function

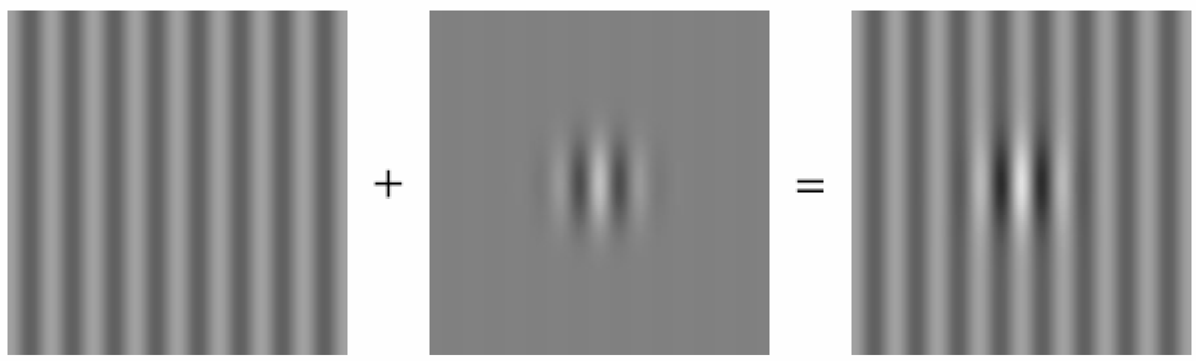


Typical Vision Model



Pattern Masking



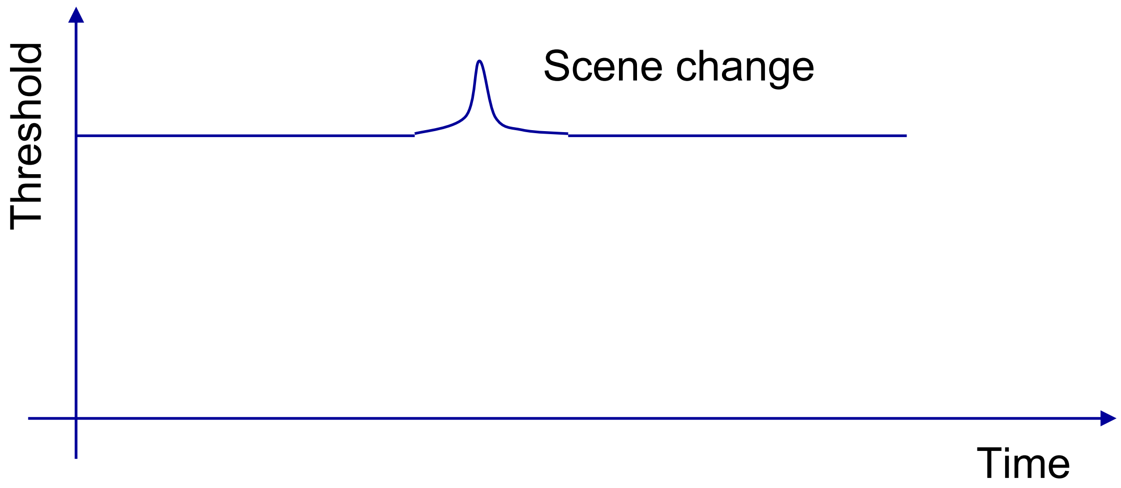


Cosine Masker Gabor Target Stimulus

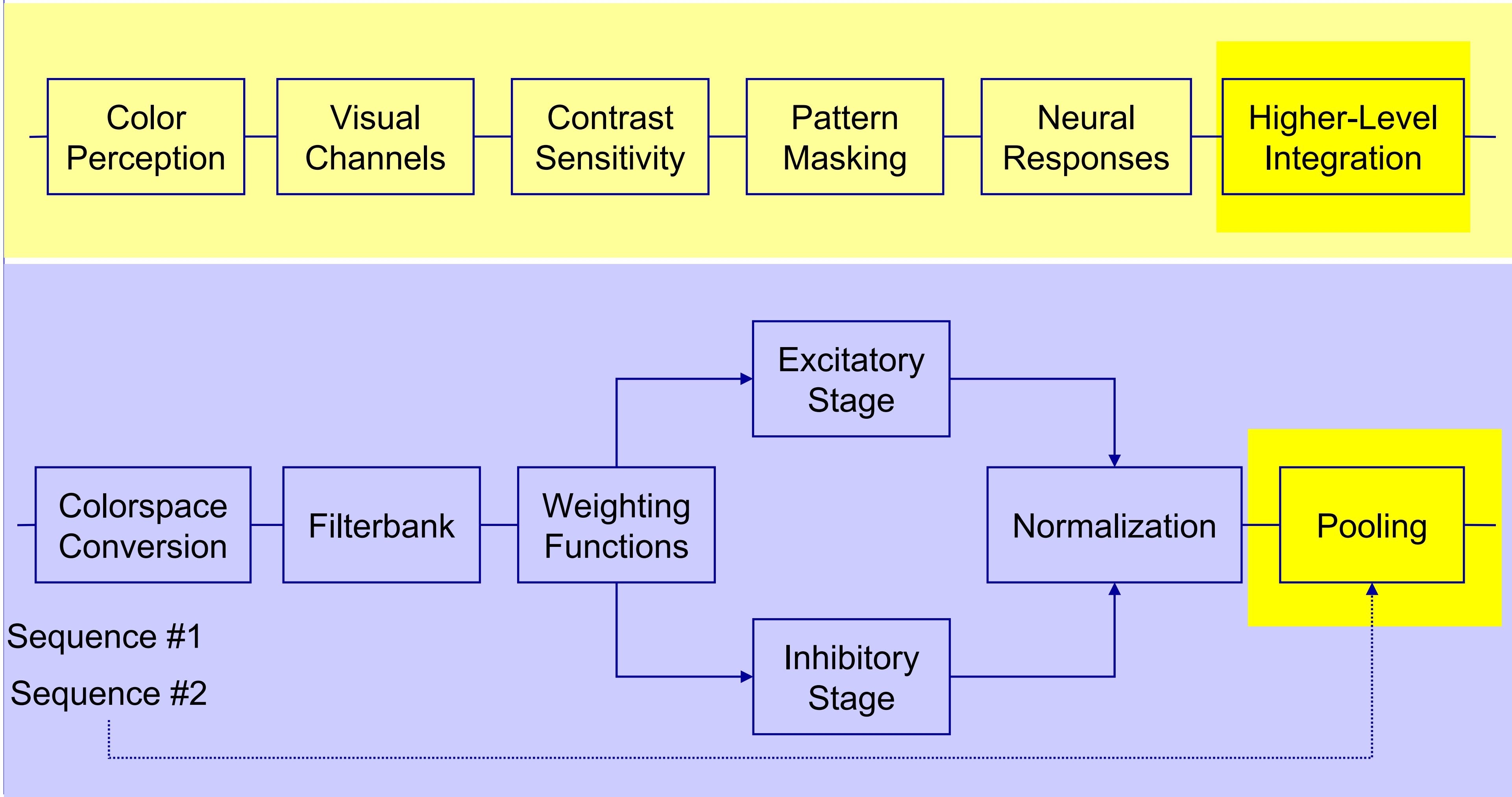
Masking

- Masking behavior depends on
  - Stimulus type (grating/noise)
  - Orientation, frequency, color,....

- Temporal masking
  - Sensitivity drop around scene changes



Typical Vision Model



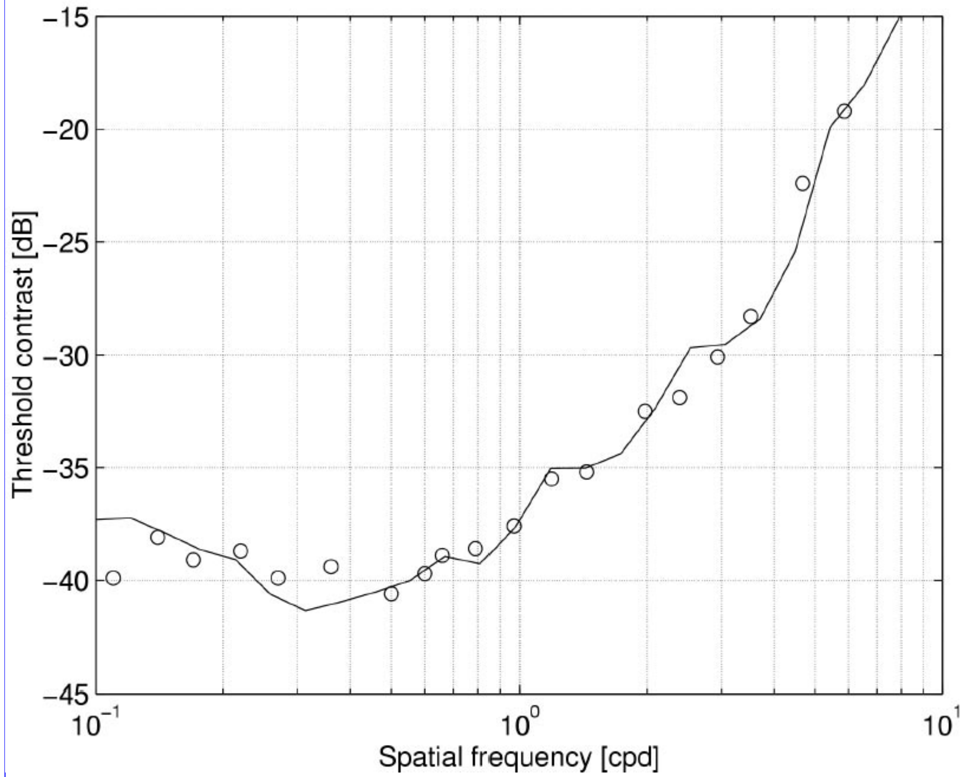
Pooling

- Pooling of “sensor” responses
  - Collect data from all channels
  - Visibility map

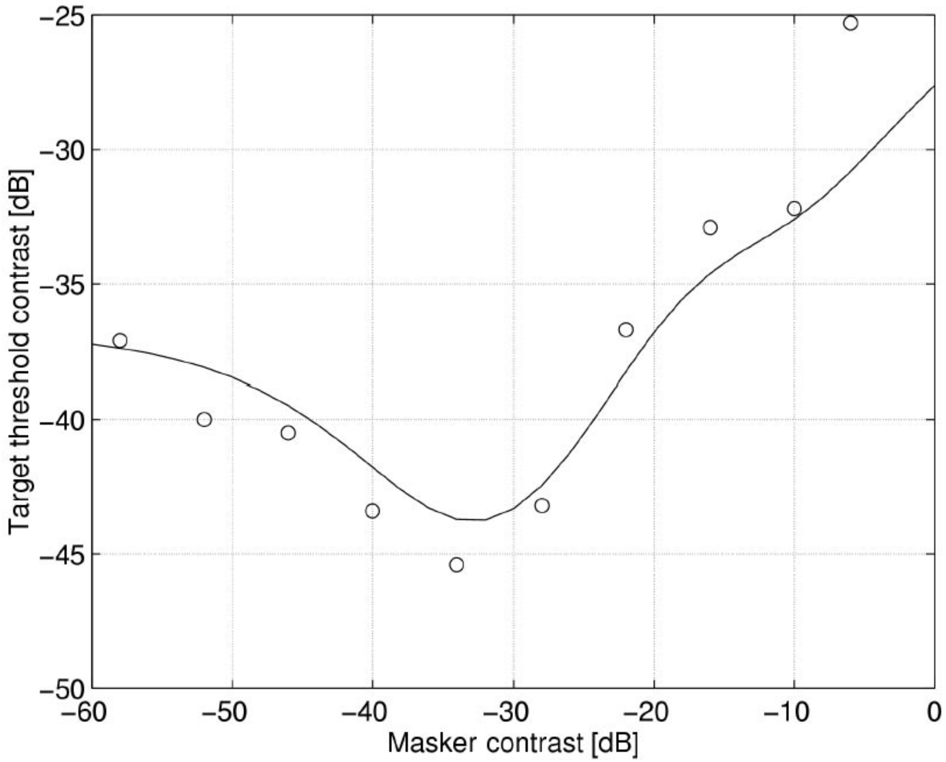
- Parameter tuning
  - Threshold data from psychophysics
  - Quality MOS data from subjective experiments

Model Fitting

- Contrast sensitivity: channel weights



- Pattern masking: contrast gain control

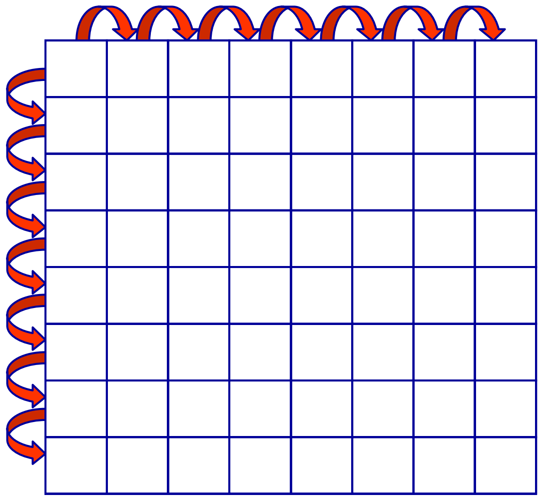


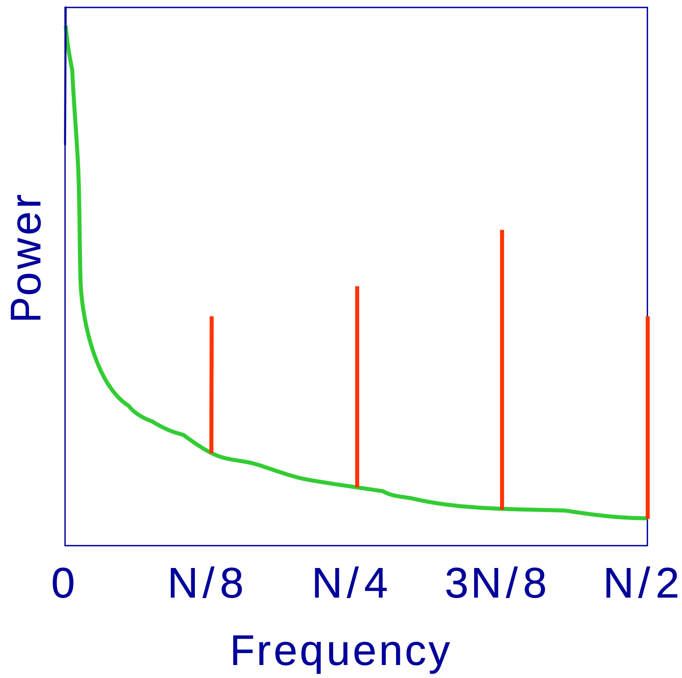
Artifact Metrics

- Blockiness
  - Block structure, block boundaries
- Blurriness
  - Reduction of high frequencies
- Jerkiness
  - Frame rate reduction (if motion)
- Noise
  - Addition of high frequencies
- Assumptions on codec/artifacts
  - Quality assessment in compressed domain

NR Blockiness Metric

- Average 1D power spectra of horizontal and vertical differences

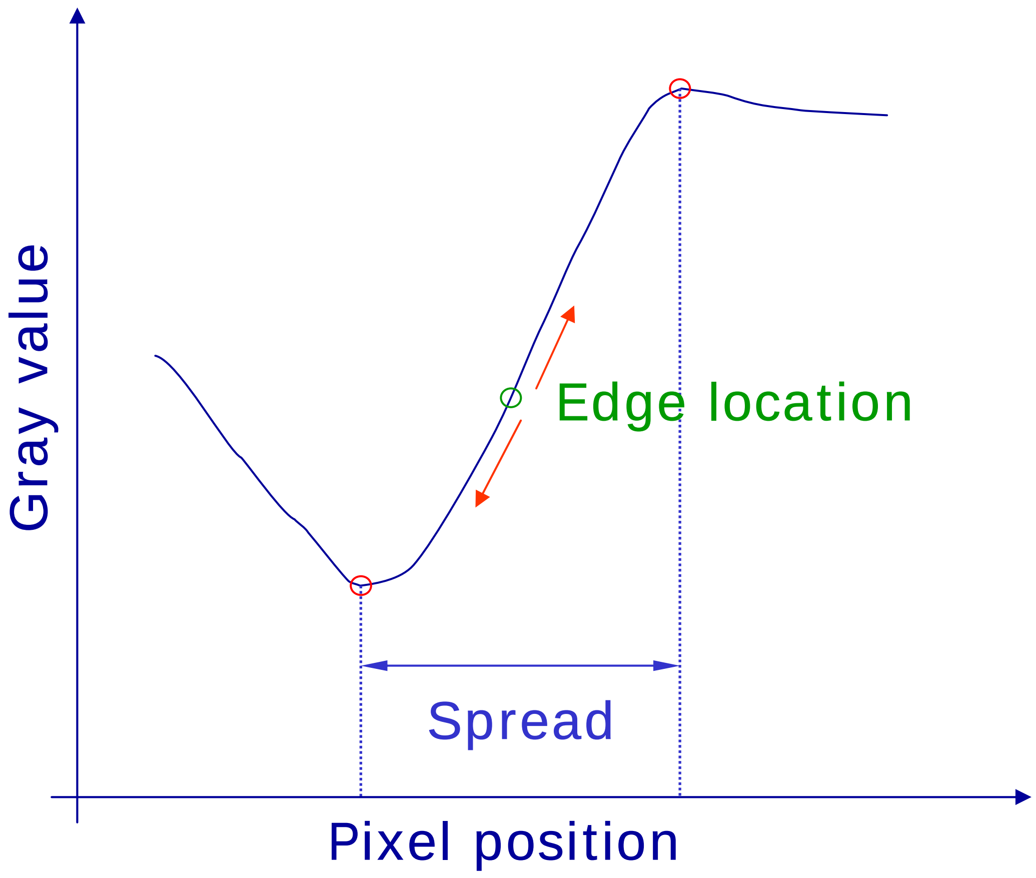




Peaks at multiples of N/8

NR Blurriness Metric

- Average spread of significant edges



Metric Extensions

- Image appeal
  - Fidelity ≠ perceived quality
- Region of interest
  - Foveal vision
  - Object tracking
- Cognitive aspects

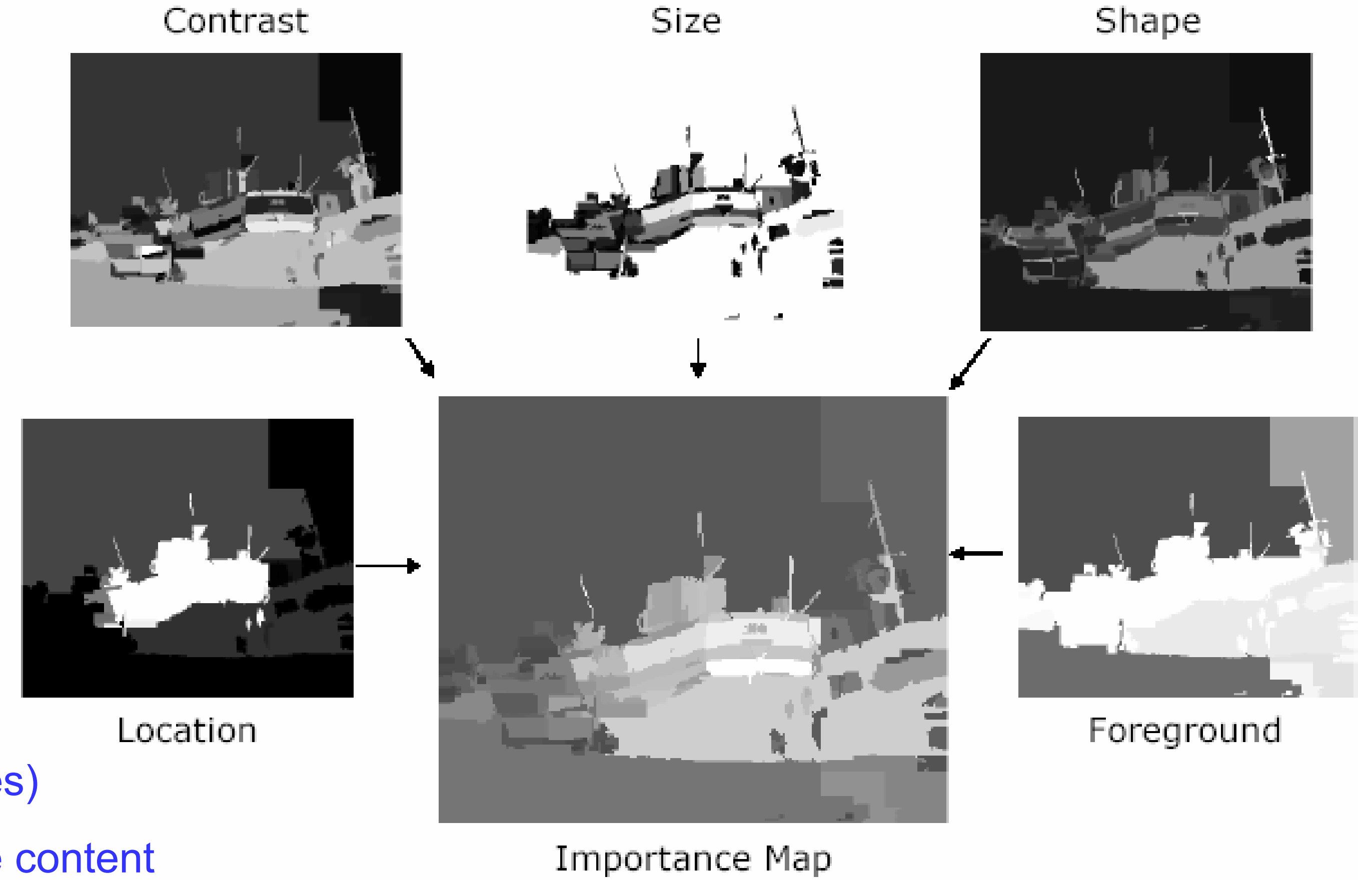
Object-based approach

Low-level features

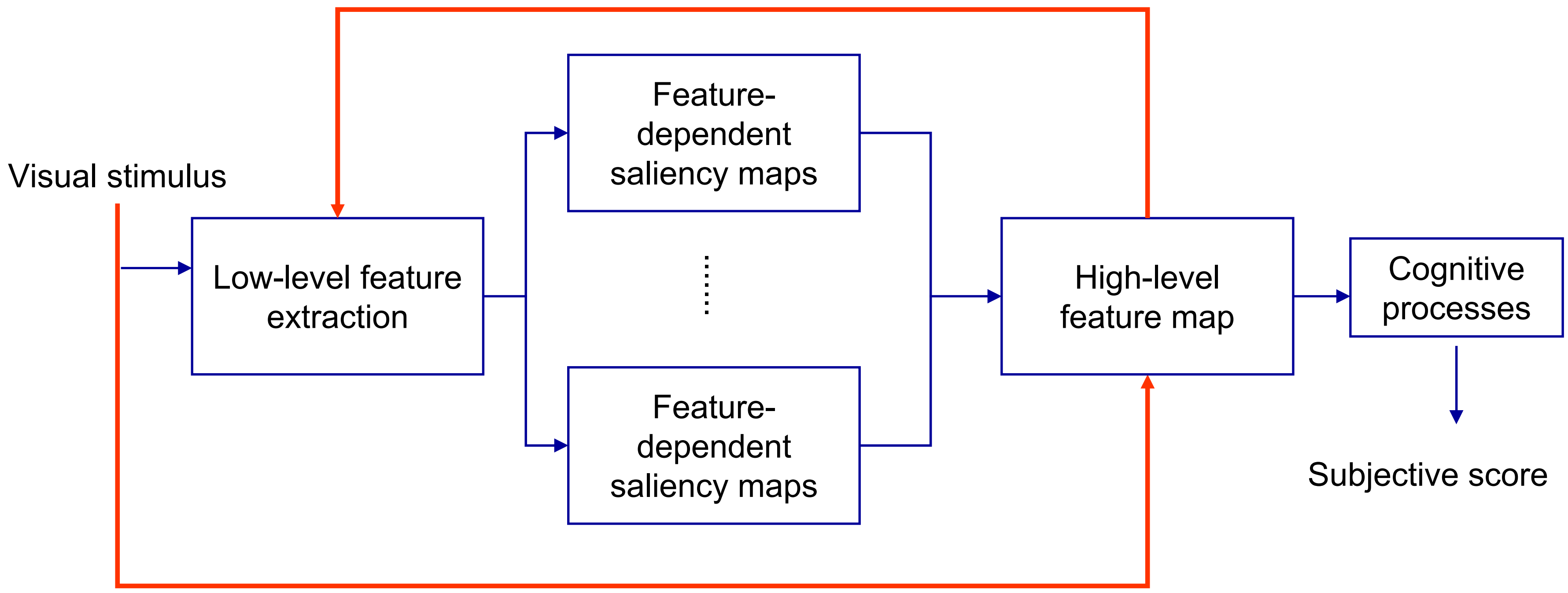
- Motion
- Location (central)
- Contrast
- Size differences
- Shape differences
- Color differences

High-level features

- Semantic objects (faces)
- Expectations on image content



Closed-loop metric



Metric Evaluation

- Reference: subjective experiments
  - Map metric predictions to subjective ratings
- Statistical analysis of prediction performance
- Performance attributes
  - Mean Opinion Score (MOS) curves
    - Measures vs predictions
  - Accuracy
    - Ability of a metric to predict subjective ratings with minimum average error
  - Monotonicity
    - Monotonicity measures if increments (decrements) in one variable are associated with increments (decrements) in the other variable, independently on the magnitude of the increment (decrement)
  - Consistency
    - Number of outliers with respect to the number of data points

VQEG Evaluation

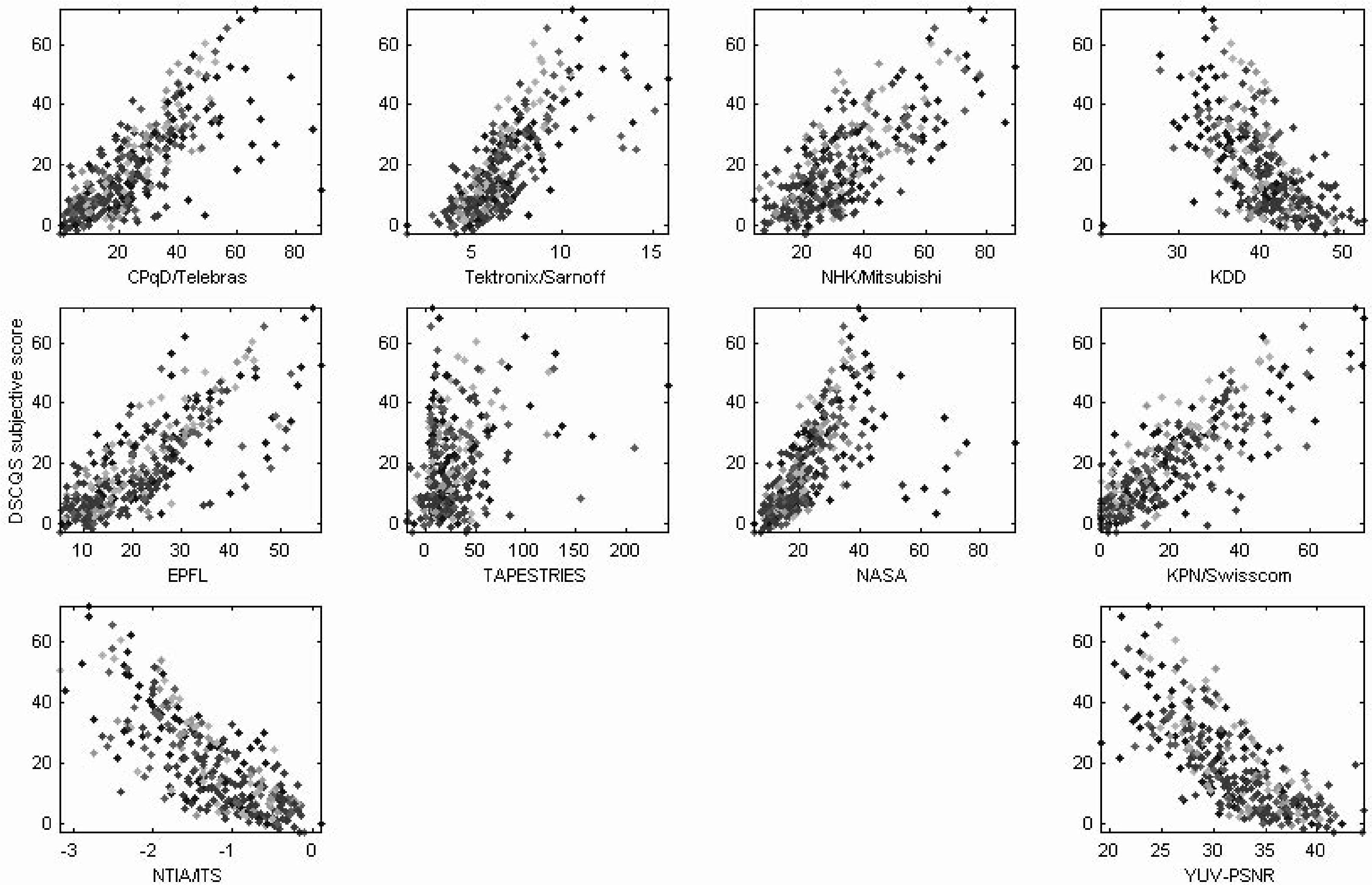
- Video Quality Experts Group (VQEG)
  - Quality metric evaluation
  - Test sequence generation
  - Subjective experiments

- Scope (Phase I)
  - Television/broadcast applications
  - Short sequences, single rating
  - Full-reference metrics

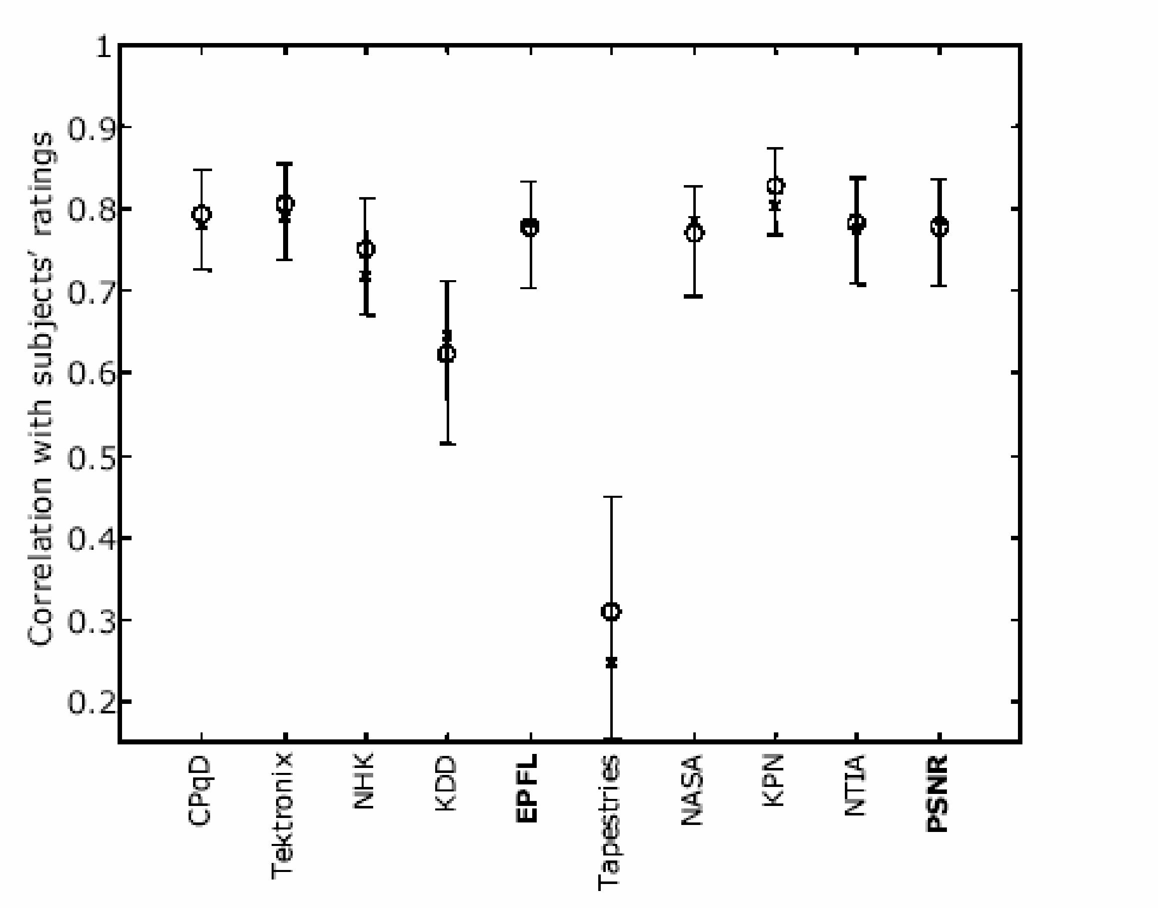
- Setup
  - 20 test scenes, 8 sec each, PAL&NTSC
  - 16 test conditions
    - MPEG2 compression (750kb/s-50Mb/s)
    - Transmission errors
    - D/A conversion
  - 320 test sequences

  - Subjective tests
  - DSCQS: 4 hours
  - 8 labs
  - 300 viewers
  - ~26.000 ratings

Metrics Performance



Metric Comparison

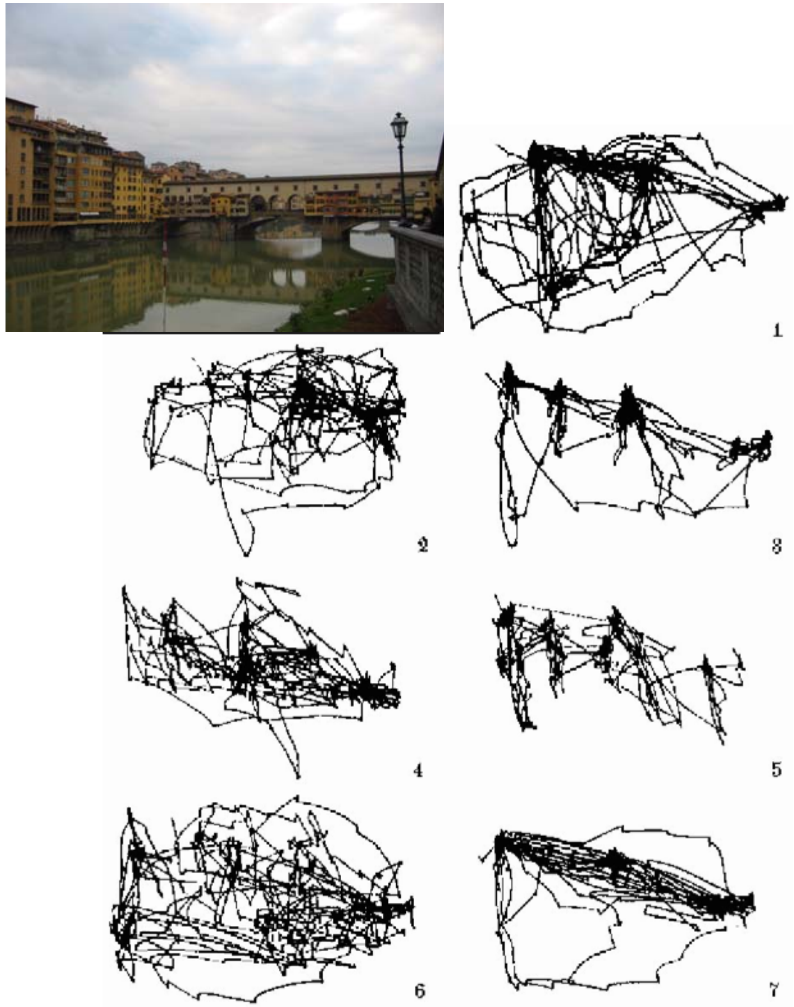


VQEG Conclusions

- Valuable set of data
- No single best metric
  - Under investigation
- No metric outperforms clearly PSNR
  - Large quality range
  - Sequence normalization
- No metric can replace subjective tests
- VQEG restrictions
  - Single rating
  - Availability of full reference
  - Offline metrics
- Work in progress

Metric Extensions

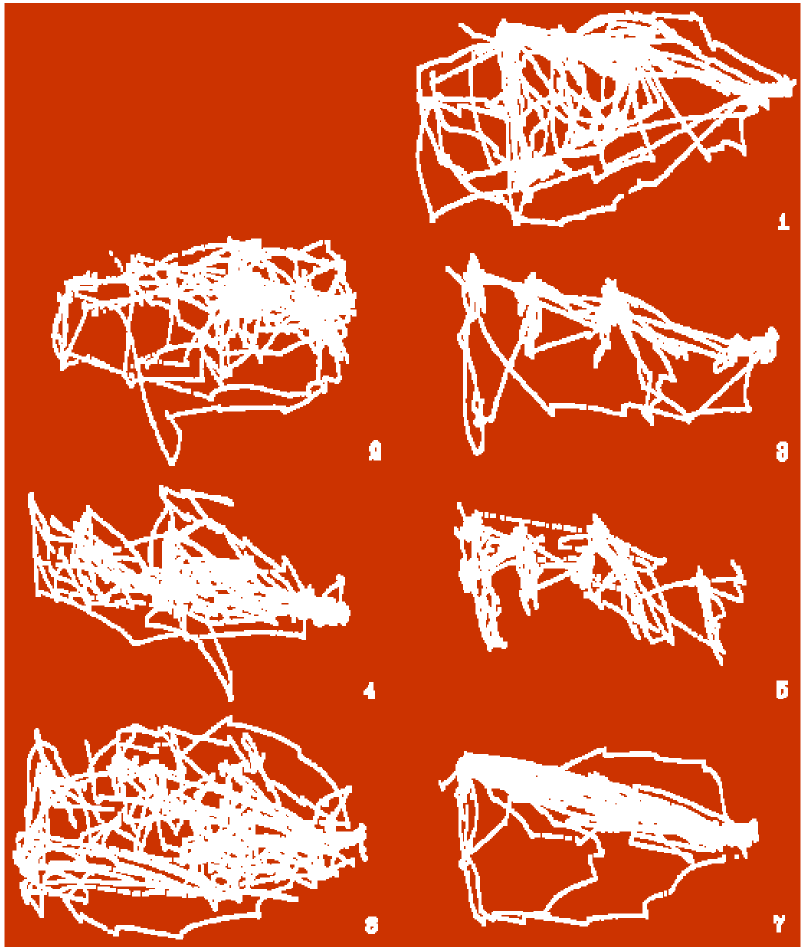
- Image *appeal*
  - Fidelity vs perceived quality
  - Sharpness (average contrast)
  - Colorfulness (spatial distribution of chroma and saturation)
- Region of interest
  - Foveal vision
  - Object tracking
  - Investigation by tracking eye movements
- Cognitive aspects



Colorfulness

Sharpness

Eye Movements



[Yarbus, 1967]

Conclusions

- State of the art
  - Full-reference
  - Out of service
  - Complex, dedicated hardware (DSP)
  - TV studio applications
- Challenges
  - Reduced-reference, no-reference
  - In service, real-time
  - Software implementation
  - Multimedia applications

Further Reading

- S. Winkler: Vision Models and Quality Metrics for Image Processing Applications. Ph.D. Thesis, 2000. (chapters 3&4)
  http://stefan.winkler.net/publications.html
- M. Yuen, H.R. Wu: "A survey of hybrid MC/DPCM/DCT video coding distortions." Signal Processing 70(3):247–278, 1998.
- P.G. Engeldrum: Psychometric Scaling. Imcotek Press, 2000.
- ITU-R Rec. BT.500-11: Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU, 2002.
- ITU-T Rec. P.910: Subjective Video Quality Assessment Methods for Multimedia Applications. ITU, 1996.
- VQEG: http://www.vqeg.org
- Visual illusions:
  http://www.ritsumei.ac.jp/~akitaoka/index-e.html

Summary

- State of the art
  - Full-reference
  - Out of service
  - Complex, dedicated hardware (DSP)
  - TV studio applications
- Challenges
  - Reduced-reference, no-reference
  - In service, real-time
  - Software implementation
  - Multimedia applications
- Perspectives
  - QoS, no-reference, real-time
  - Investigation of perceptual aspects (low level and cognitive)