Quality of Experience in Multimedia Communications

Dr. Anil Fernando
I-Lab Multimedia Communications Research
Centre for Vision, Speech and Signal Processing,
University of Surrey, UK

Email: W.Fernando@surrey.ac.uk

I-Lab

-Part of the Centre for Vision, Speech, and Signal processing, University of Surrey

-Concentrates in many aspects of multimedia

-A Group of 40+ researchers

-Key Player in many EU projects

I-Lab's Current Major Research

1. 2D/3D Video Coding
2. Video processing
3. Video Communications
4. Quality of Experience
5. Audio Coding/Processing
6. Speech Coding/Processing
7. Multimedia Communications

Overview

1. Introduction to video coding
2. Why Quality Assessments?
3. Classifications of Quality Assessments
4. 2D Video Quality Assessment Techniques
5. Introduction to 3D Video
6. 3D Video Quality Assessment Techniques
7. Disparity Distortion Model
8. Quality of Experience
9. Summary

1. Introduction to video coding

Why Video Coding

- Video needs high bit rates
- Bandwidth/storage
- Per-byte-cost of transmission would have to be very low

What is Video Coding

Compressed video

- A reversible conversion of data
- Requires fewer bits
- Transmitted more efficiently

An Example for Image Compression

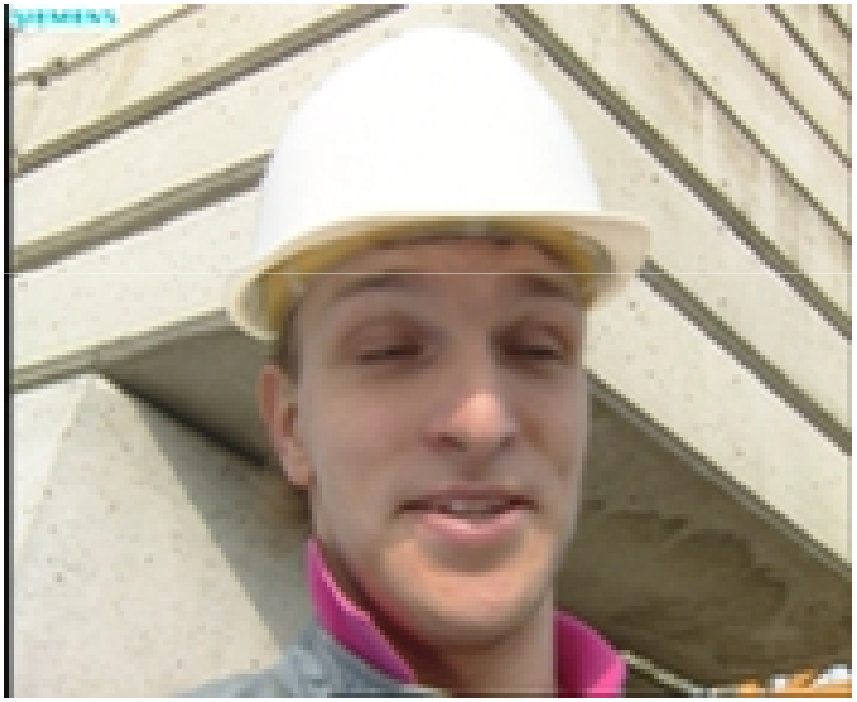
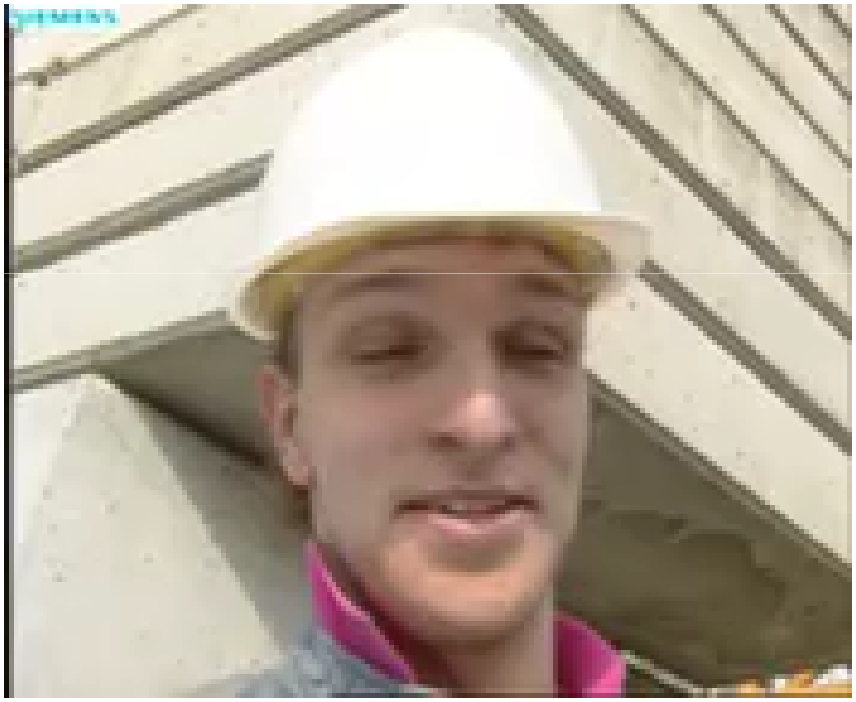
- Digital colour image: 352x288 pixel
- RGB representation: 24 bpp (8bits for red,green,blue)
- Total amount of bytes: > 300K
- JPEG: common image compression standard, < 20K, similar quality

Original Image
Size: 300k

Compressed Image
Size: 20k
PSNR: > 30dB

An Example for Video Compression

MUSCADE
MUltimedia SCAlable 3D for Europe

Uncompressed	Compressed
	
Bit Rate= 4,562 Kbits/s	Bit Rate= 100 Kbits/s

Basics of video compression

- Identify the *redundancies* within the video signal
- Types of redundancies:
  - Spatial: adjacent pixels are often correlated
  - Temporal: adjacent frames highly correlated
- Properties of the Human Visual System

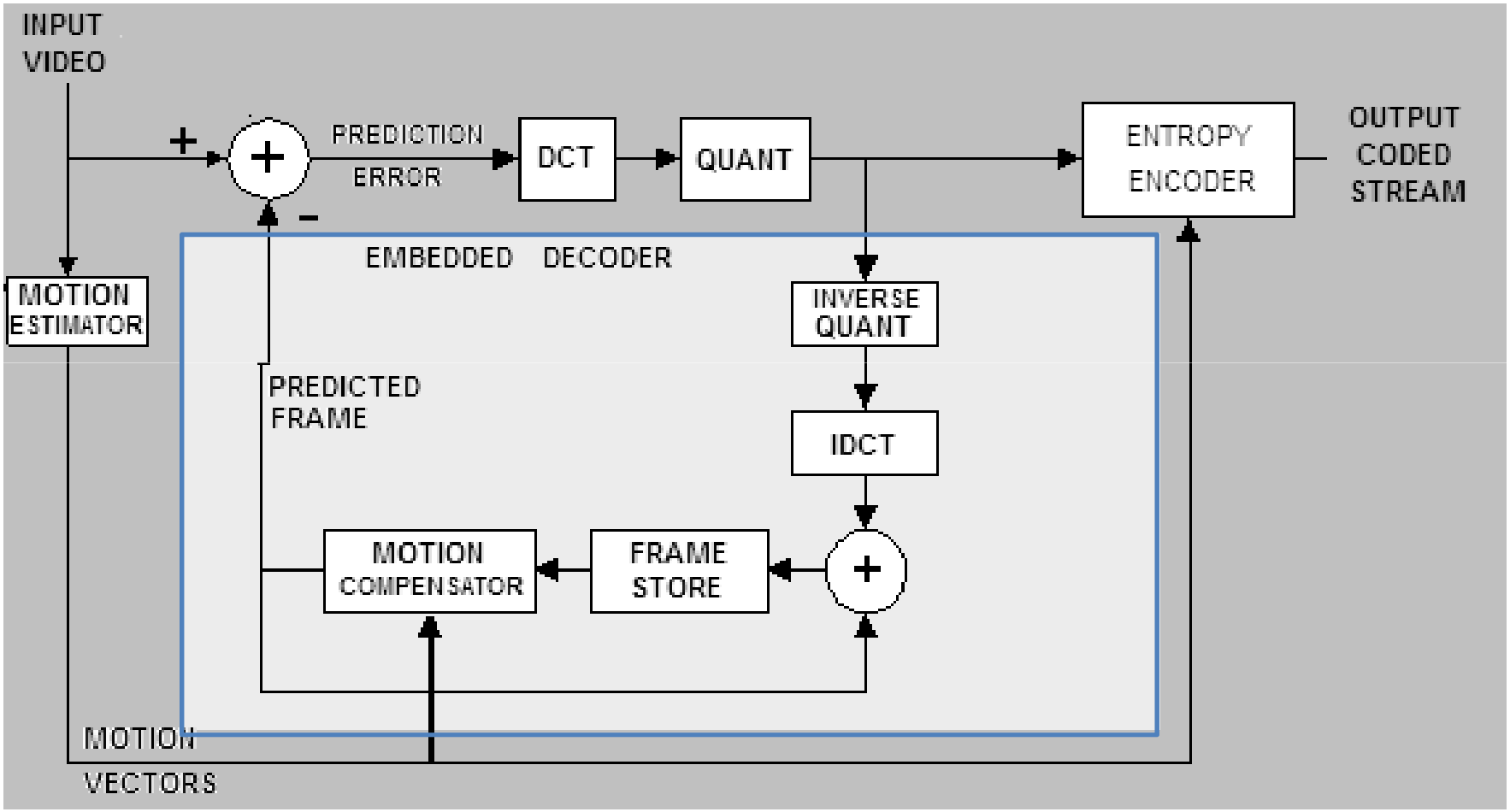
- Spatial redundancy reduction
  - Using transform coding, spatial predictive coding
- Temporal redundancy reduction
  - Motion compensation/estimation (MC/ME)
  - Temporal predictive coding

Basics of video compression

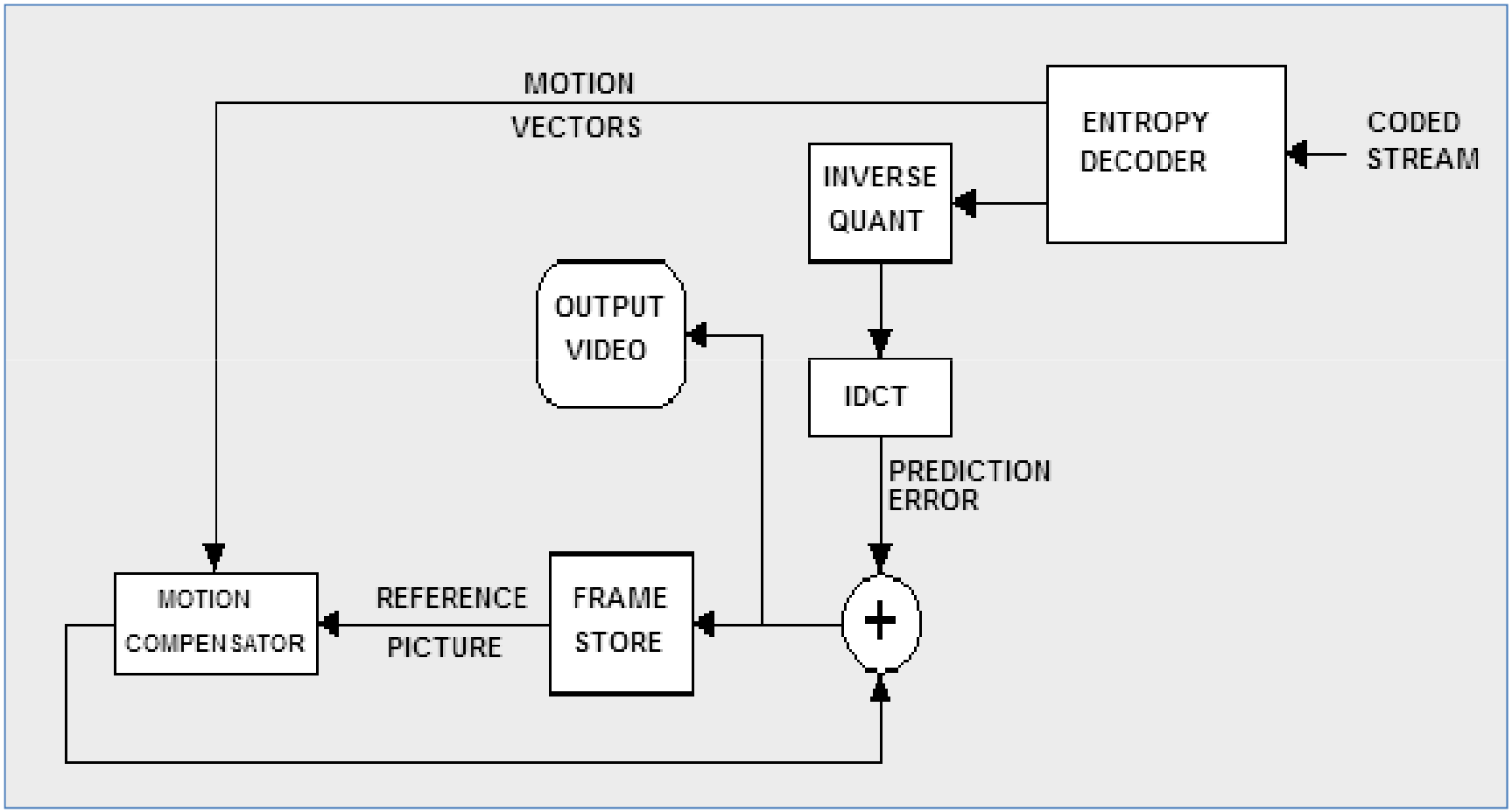
Uses two basic compression techniques:

- **Intra-frame** compression
  - Within individual frames
  - Minimize the duplication of data in each
  - picture (Spatial Redundancy)
- **Inter-frame** compression
  - Consider adjacent frames
  - Minimize data redundancy in successive pictures (Temporal redundancy)

Typical Video Encoder

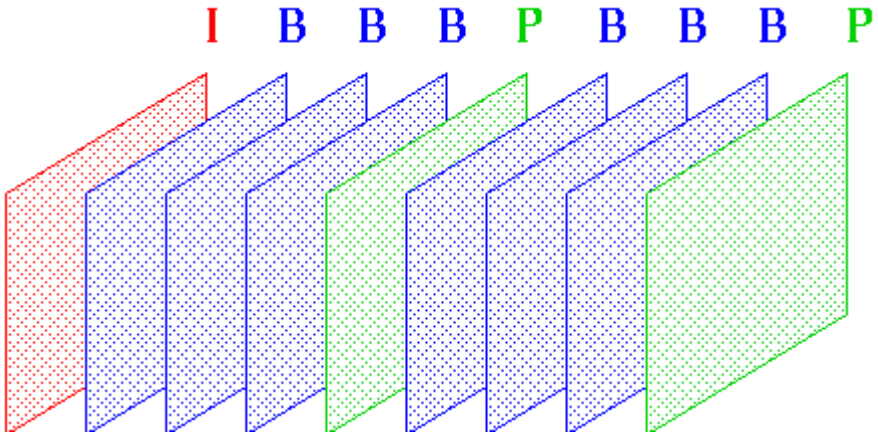


Typical Video Decoder



Moving Picture Types

To achieve the requirement of random access, a set of pictures can be defined to form a Group of Picture (GOP), consisting of a minimum of one I-frame, which is the first frame, together with some P-frames and/or B-frames.



Three types of pictures:

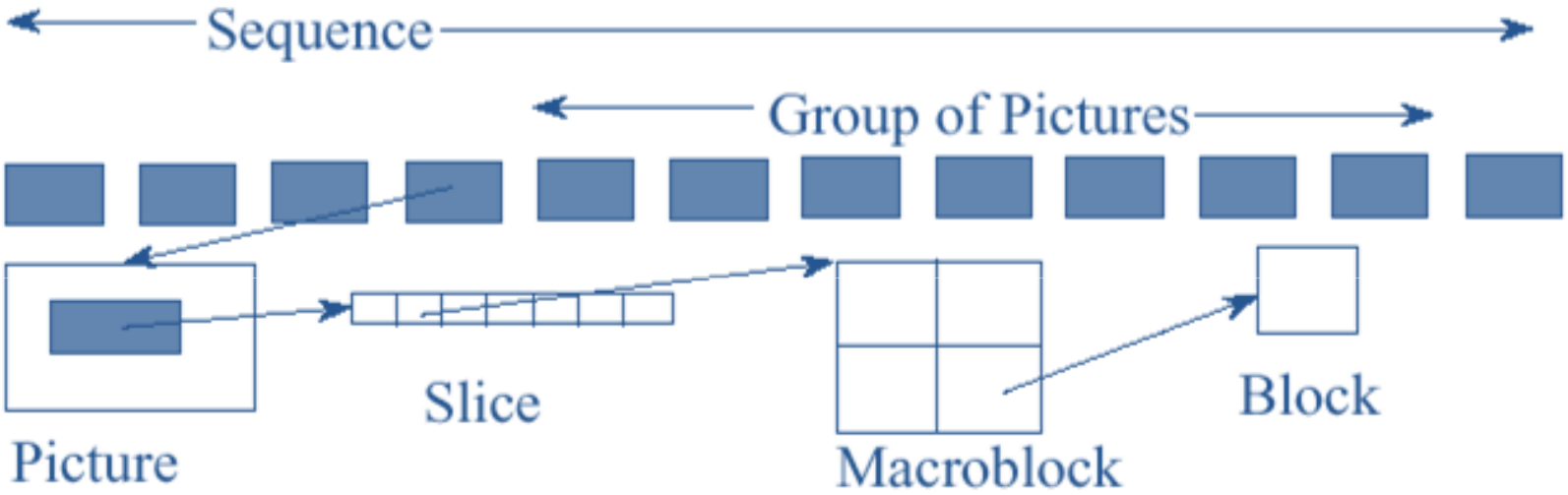
- Intra-pictures (I)
- Unidirectional predicted pictures (P)
- Bidirectional predicted pictures (B)

Grouped together (typically 12 pictures) in GOPs

Quantization

- The human eye responds differently to different spatial frequency coefficients.
- If the magnitude of higher frequency coefficients is below a certain threshold the human eye will not detect it.
- Quantization is a process that attempts to determine what information can be safely discarded without a significant loss in visual fidelity
- Based on a set of quantization tables derived from empirical experimentation

Typical structure of the compressed video sequence (MPEG)



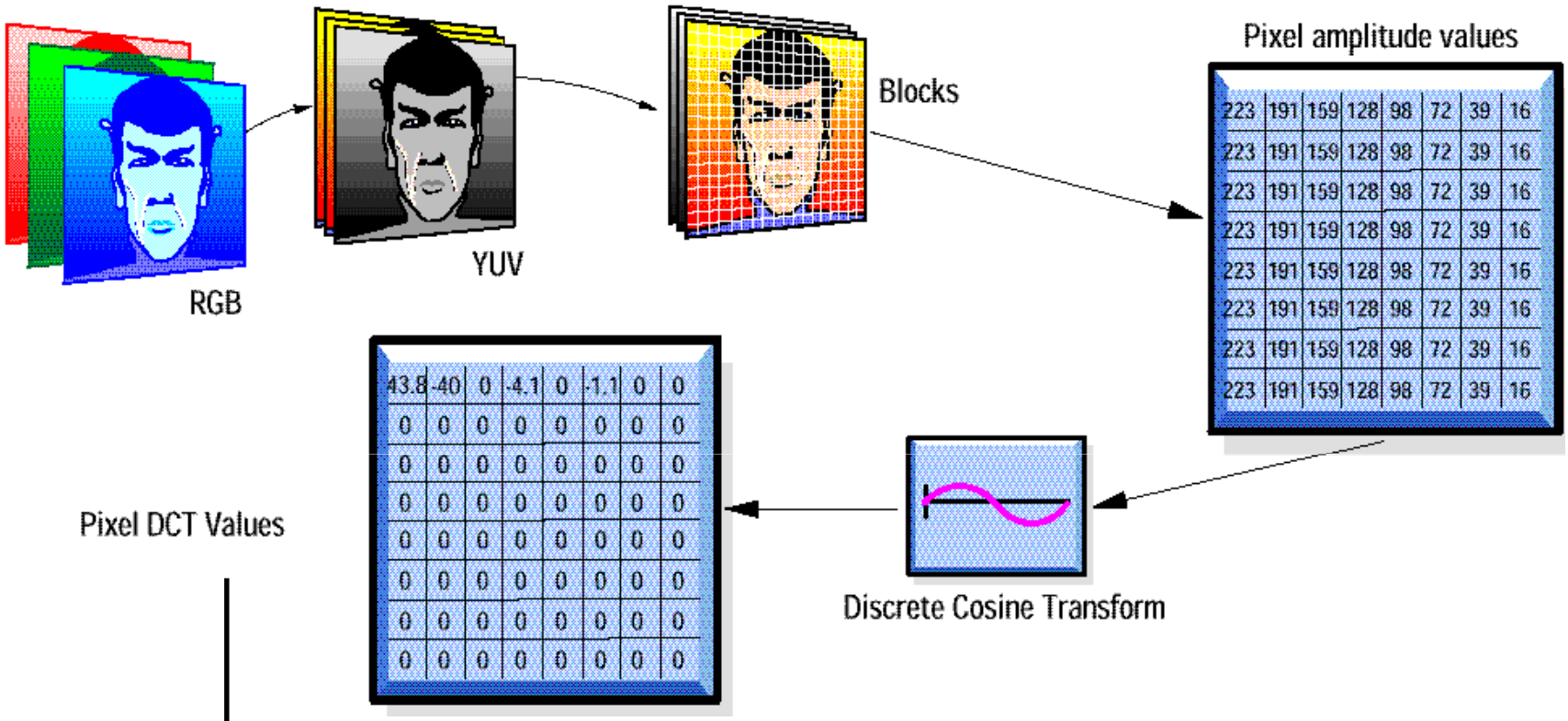
The Bit Stream

Sequence 1 Level	GOP Level	Picture Level	Slice Level	MB Level	Block Level	Sequence 2 Level	

I-frame

- A picture coded without reference to any picture except itself.
- It is a still image encoded in JPEG in real-time.
- Often, I pictures (I-frames) are used for random access and are used as references for the decoding of other pictures.

Intra-frame Coding (I)

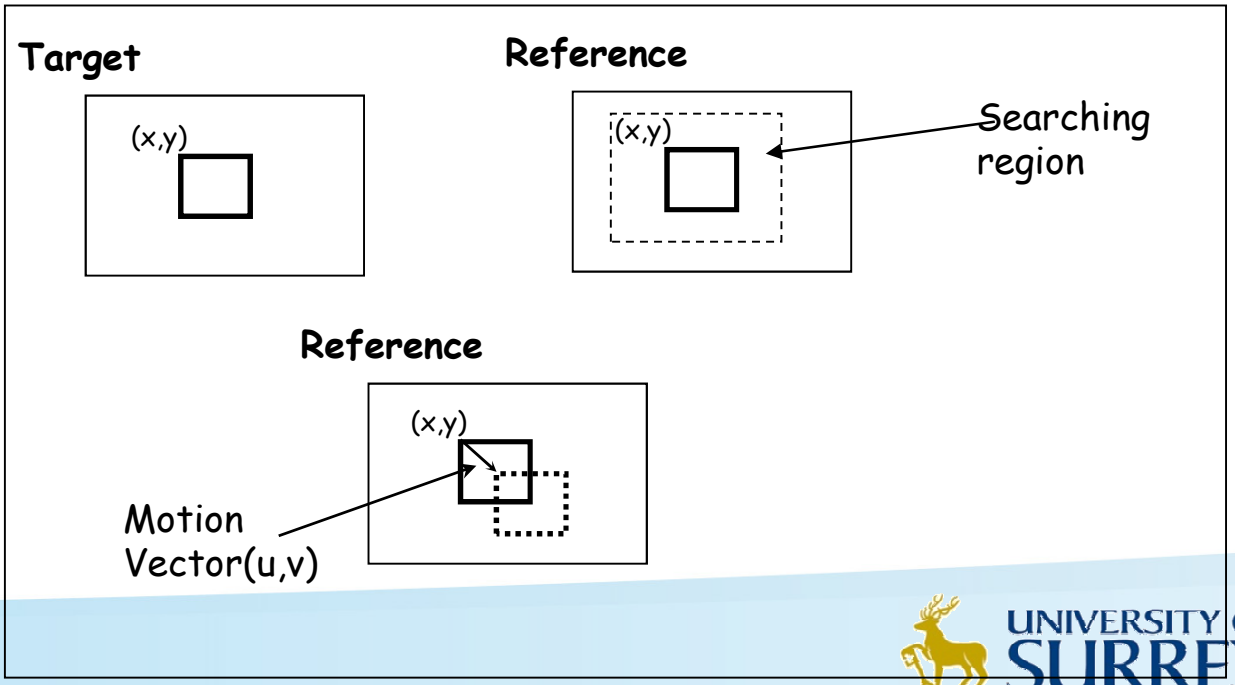


Inter-frame Coding

Motion Estimation: Find the best match for each MB from reference frame(s)

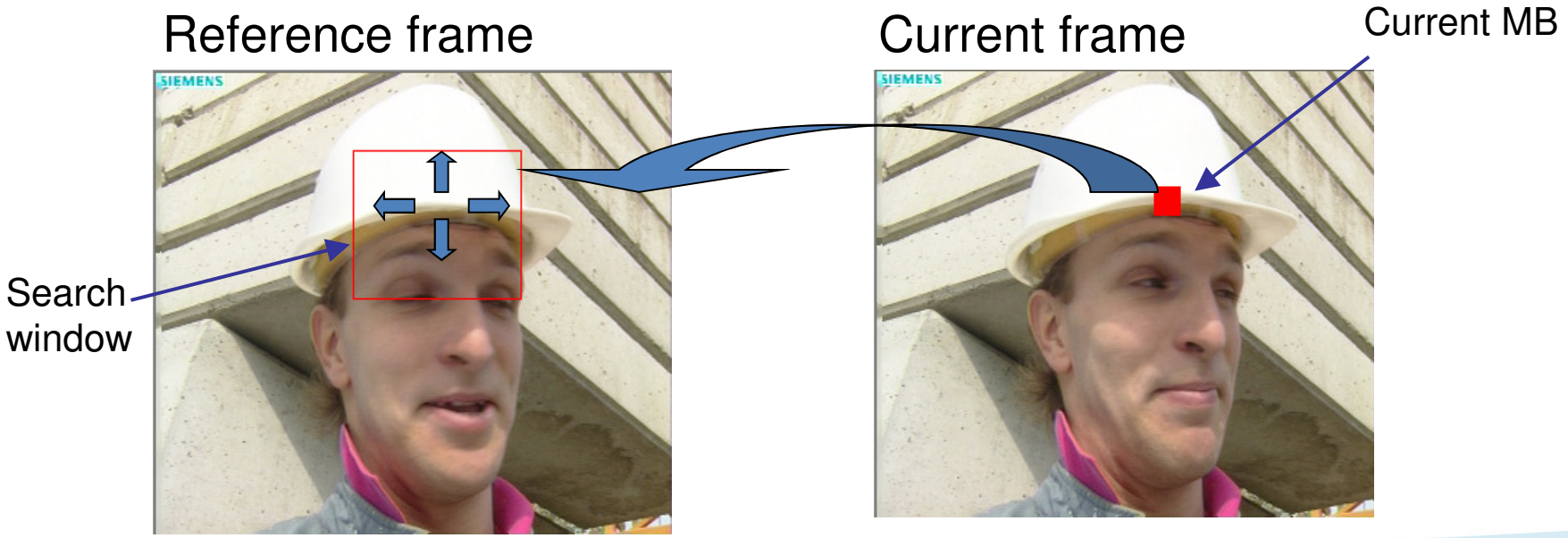
Error Signal: Coded using DCT

Motion Estimation



Motion Estimation

- Find the MB in the reference frame which is the most similar to the current MB
- Error metric: MSE (*Mean Square Error*) or SAD (*Sum of Absolute Differences*)

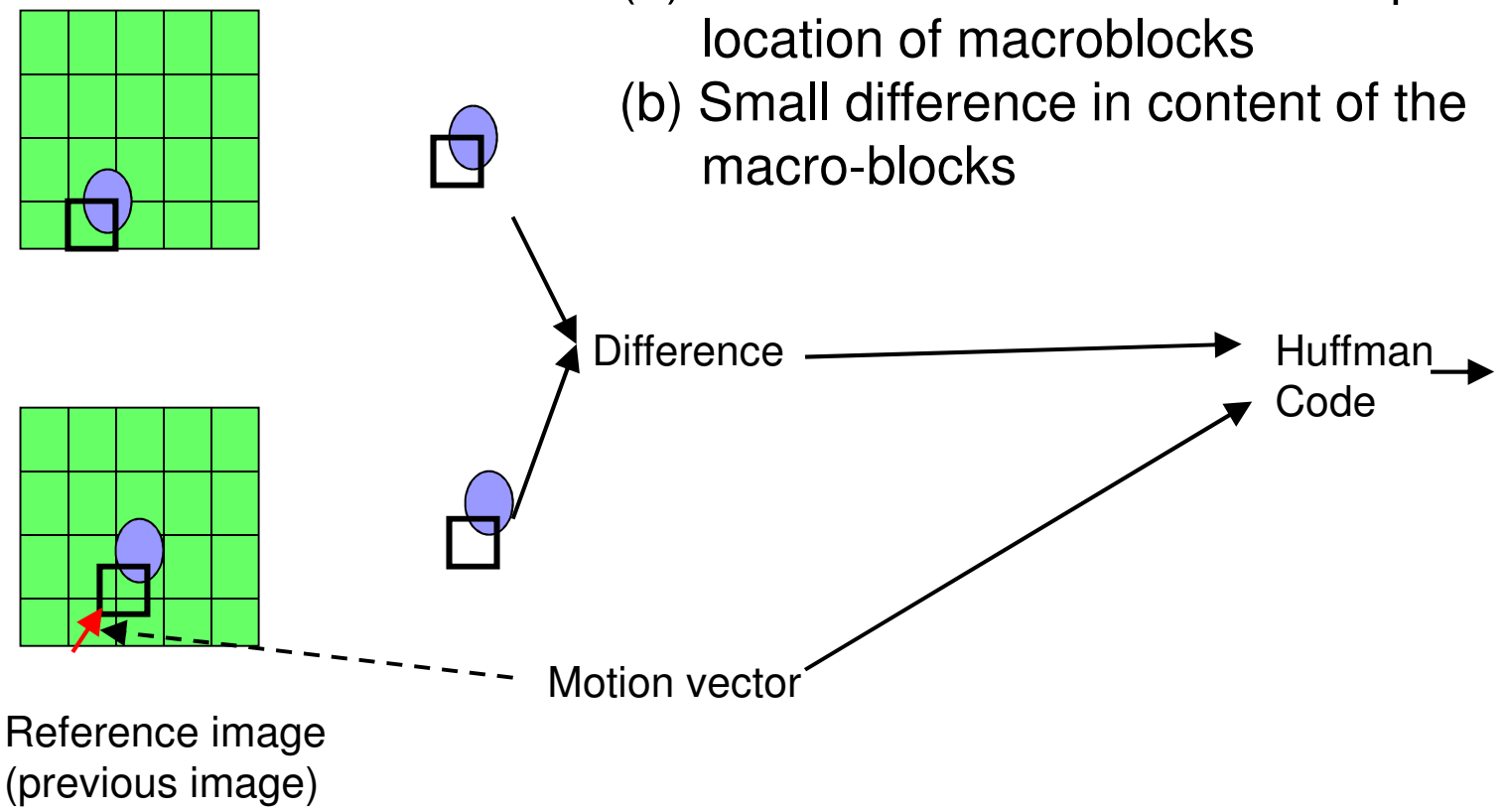


P-frame (Unidirectionally predictive coded frames)

Encode:

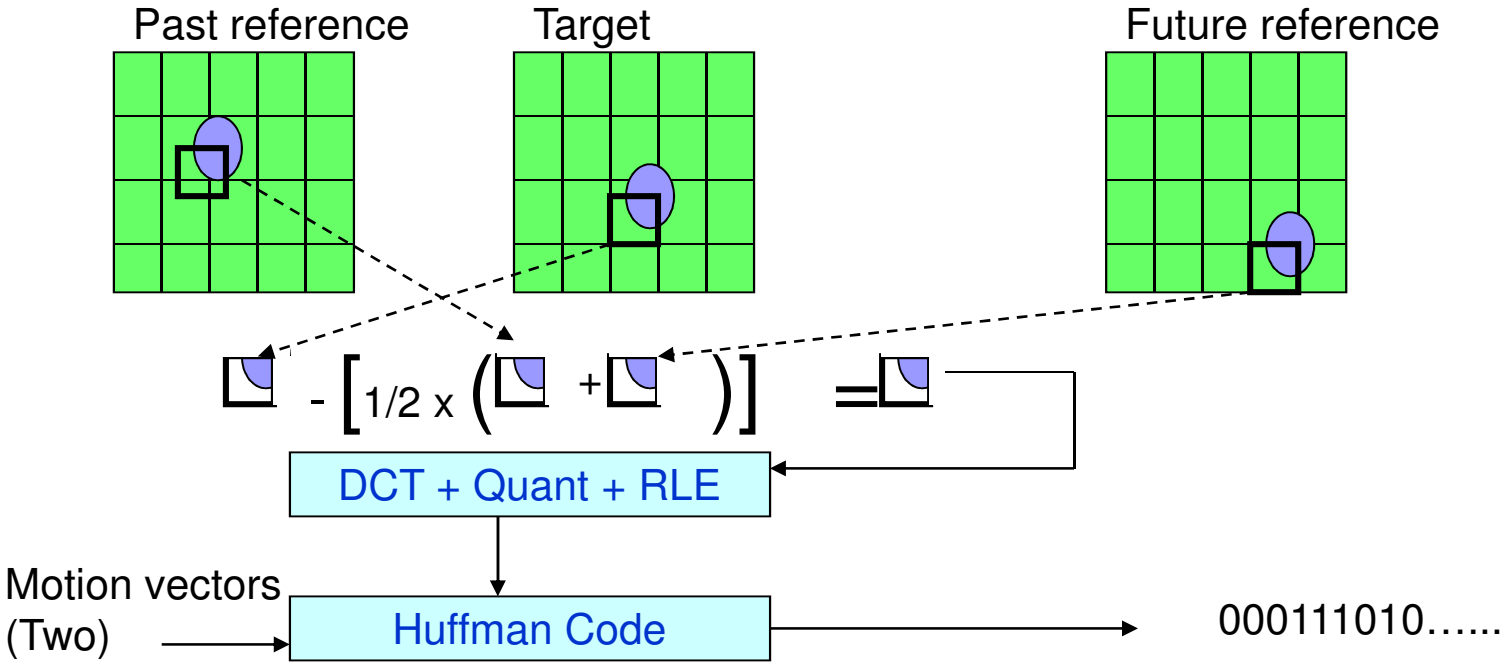
(a) motion vector - difference in spatial location of macroblocks

(b) Small difference in content of the macro-blocks

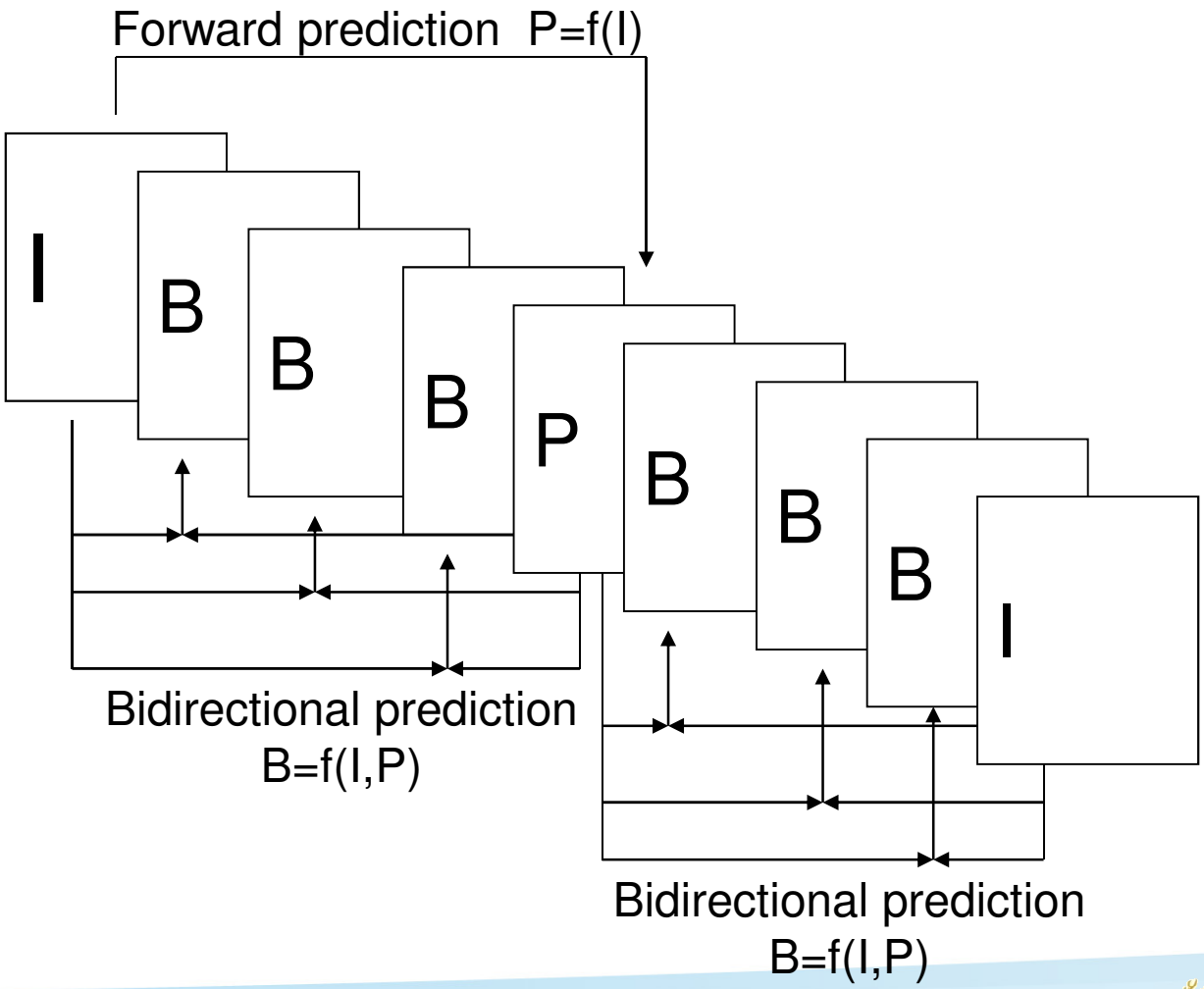


B-frame (Bidirectionally predictive coded frames)

- Two motion vectors are estimated (one to a past frame, one to a future frame).



Motion Compression for Coding MPEG

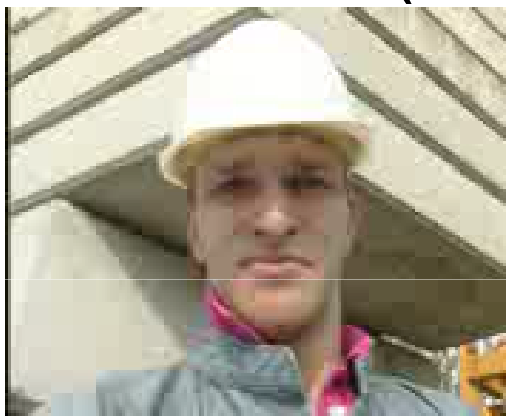


Performance of Comparison of Video Codecs

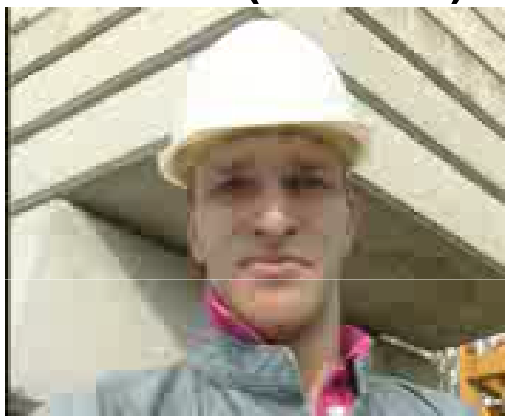
MUSCADE
MUltimedia SCAlable 3D for Europe

Original

H.263 baseline (33 dB)



H.263+ (33.5 dB)



MPEG-4 core (33.5 dB)

H.264 (42 dB)

(QCIF, 30 fps, 100 Kbits/s)

2. Why quality Assessments?

- Signal Processing
- Compression
- Communications
- Display Manufactures
- Service Providers

How to measure the video quality?

- Through **Subjective tests**, which request the user to evaluate & rate the quality of the multimedia content.
- Through **Objective metrics,** which can automatically evaluate the quality of the multimedia content, predicting the subjective judgment.

Pros & Cons of Subjective Tests

Pros

- The human subject is the one who will judge the quality of the multimedia content in the real life.

Cons

- Time consuming
- Difficult to design
- Cannot be performed in real time

Pros & Cons of Objective Metrics

Pros

- Fast & Cheap

Cons

- It is difficult to design metrics which achieve high ***correlation*** with the end user perception!
- Only a few implementations of visual quality metrics are ***publicly available***
- Lack of extensive and reliable ***comparison*** of existing techniques

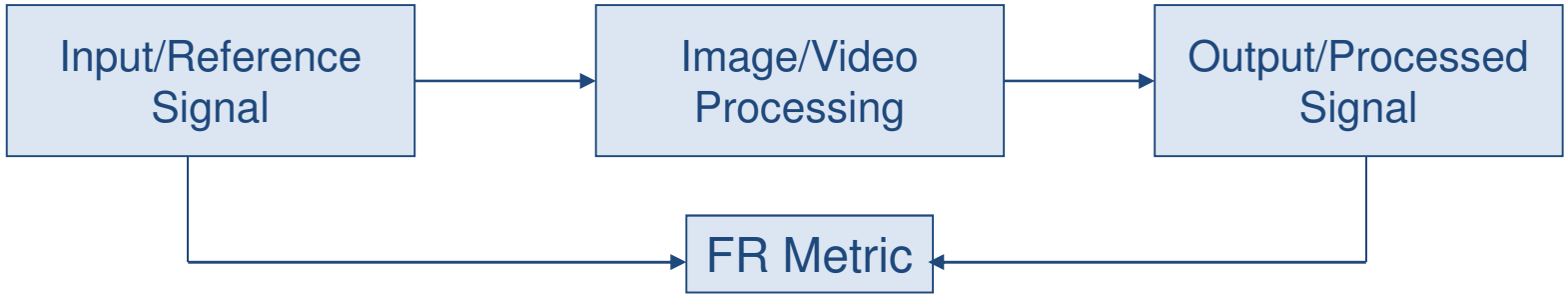
Usage of Objective Quality Metrics

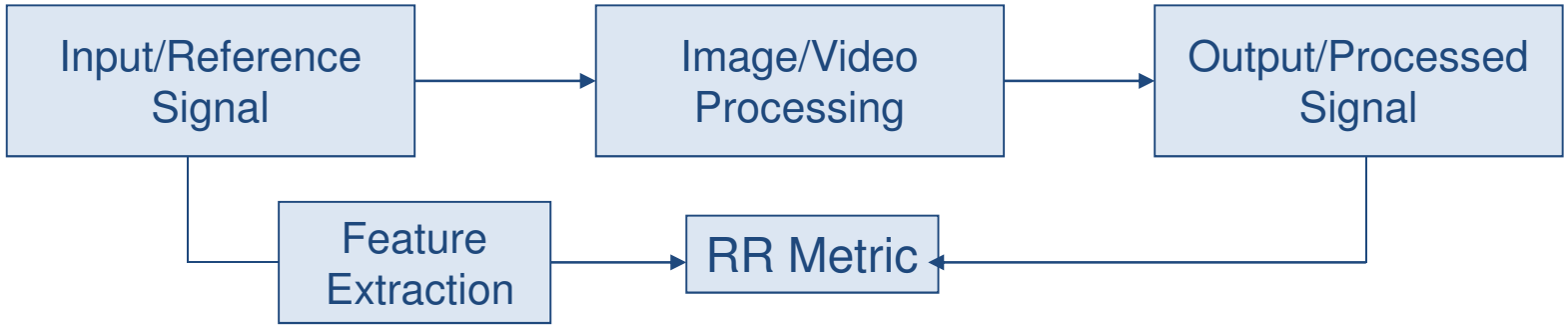
1. Monitor image quality for quality control systems
2. It can be employed to benchmark image and video processing systems and algorithms.
3. It can be embedded into an image and video processing system to optimize the algorithms and the parameter settings

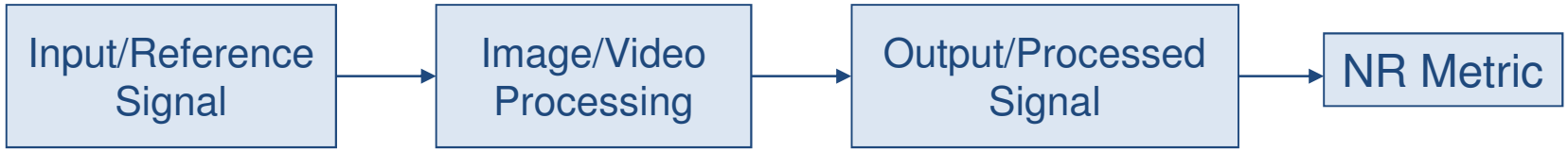
3. Classifications of Quality Metrics

1. Full Reference metrics (FR)
2. Reduced Reference metrics (RR)
3. No Reference metrics (NR)

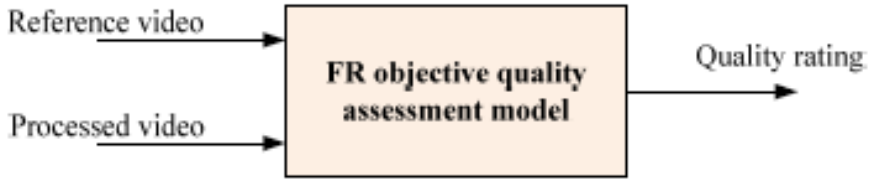
FR, NR and RR Metrics



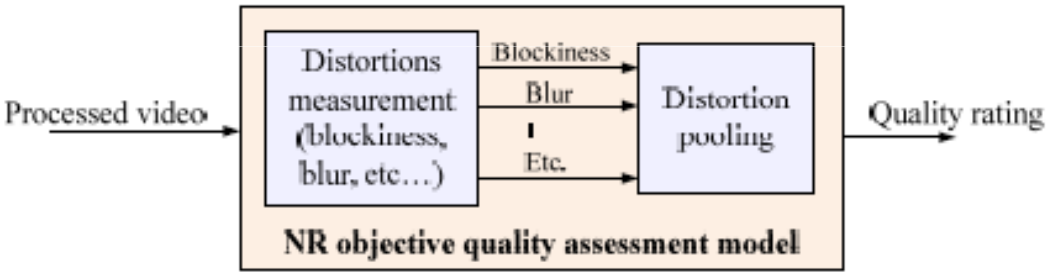




Objective quality assessment techniques



(a) FR objective quality assessment model



(b) NR objective quality assessment model

(c) RR objective quality assessment model

Full Reference Quality Assessment

- Most of the FR quality assessment models share a common error sensitivity based approach

- An image or video signal whose quality is being evaluated can be thought of as a sum of a perfect reference signal and an error signal.

- It is assumed that the loss of quality is directly related to the strength of the error signal.

Full Reference Quality Assessment

- Therefore, a natural way to assess the quality of an image is to quantify the error between the distorted signal and the reference signal.
- Simplest implementation is the MSE

Full Reference Metrics

- Several Metrics have been proposed
- MSE based techniques are most popular

What's wrong with MSE?

- There are a number of reasons why MSE may not correlate well with the human perception of quality
  - Digital pixel values on which the MSE is typically computed, may not exactly represent the light stimulus entering the eye.
  - The sensitivity of the HVS to the errors may be different for different types of errors
  - Two distorted image signals with the same amount of error energy may have very different types of errors

Peak Signal to Noise Ratio (PSNR)

- Based on MSE
- The most popular FR method due to its simplicity

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2$$

$$PSNR = 10\log_{10}\frac{L^2}{MSE}$$

  - L=255
  - N is the number of pixels and x_i and y_i are the original and distorted pixel values

What's wrong with MSE?

a

b

c

d

- Evaluation of "Lena" images with different types of noise
- In all the images b,c and d the MSE is equal to 225
- Note the differences

a) Original 'Lena' Image

b) Impulsive salt-pepper noise contaminated image

c) Additive Gaussian noise contaminated image

d) Multiplicative speckle noise contaminated image

What's wrong with MSE?

a

b

c

d

- Evaluation of "Lena" images with different types of distortion
- In all the images b,c and d the MSE is equal to 225
- Note the differences

a) Mean shifted image

b) Contrast stretched image

c) Blurred image

d) JPEG compressed image (MSE 215)

FR method with Structural Distortion Measurement

- The main function of the human visual system is to extract structural information from the viewing field
- The human visual system is highly adapted for this purpose
- Therefore, a measurement of structural distortion should be a good approximation of perceived image distortion
- The most popular structural distortion based metric is known the Structural Similarity Index (SSIM)

SSIM

- SSIM is a method for measuring the similarity between two images
- Structural Distortion measure
- Considers image degradation as *perceived structural information loss*
- Follows a top down approach
- Publicly available
  - http://www.ece.uwaterloo.ca/~z70wang/research/ssim/

Components of SSIM Index

- Considers 3 aspects of distortion
  - Luminance Change
  - Contrast Change
  - Structural Change

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

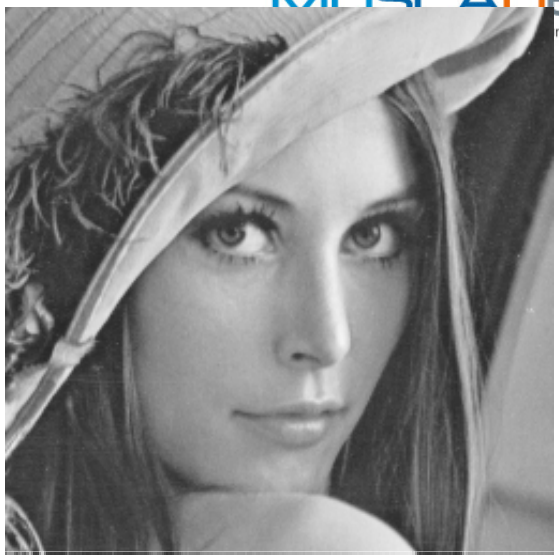
$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

$$SSIM(\boldsymbol{x},\boldsymbol{y}) = l(\boldsymbol{x},\boldsymbol{y}) \cdot c(\boldsymbol{x},\boldsymbol{y}) \cdot s(\boldsymbol{x},\boldsymbol{y})$$

Performance of SSIM

MSE=0, MSSIM=1

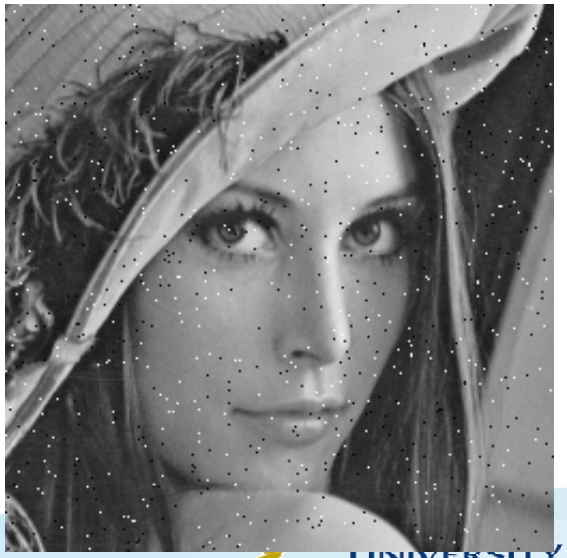
MSE=225, MSSIM=0.949



MSE=225, MSSIM=0.989

MSE=215, MSSIM=0.671

MSE=225, MSSIM=0.688



MSE=225, MSSIM=0.723

Extensions of SSIM (1)

- Color image quality assessment

  [Toet & Lucassen., *Displays,* '03]

- Video quality assessment

  [Wang, *et al*., *Signal Processing: Image Communication,* '04]

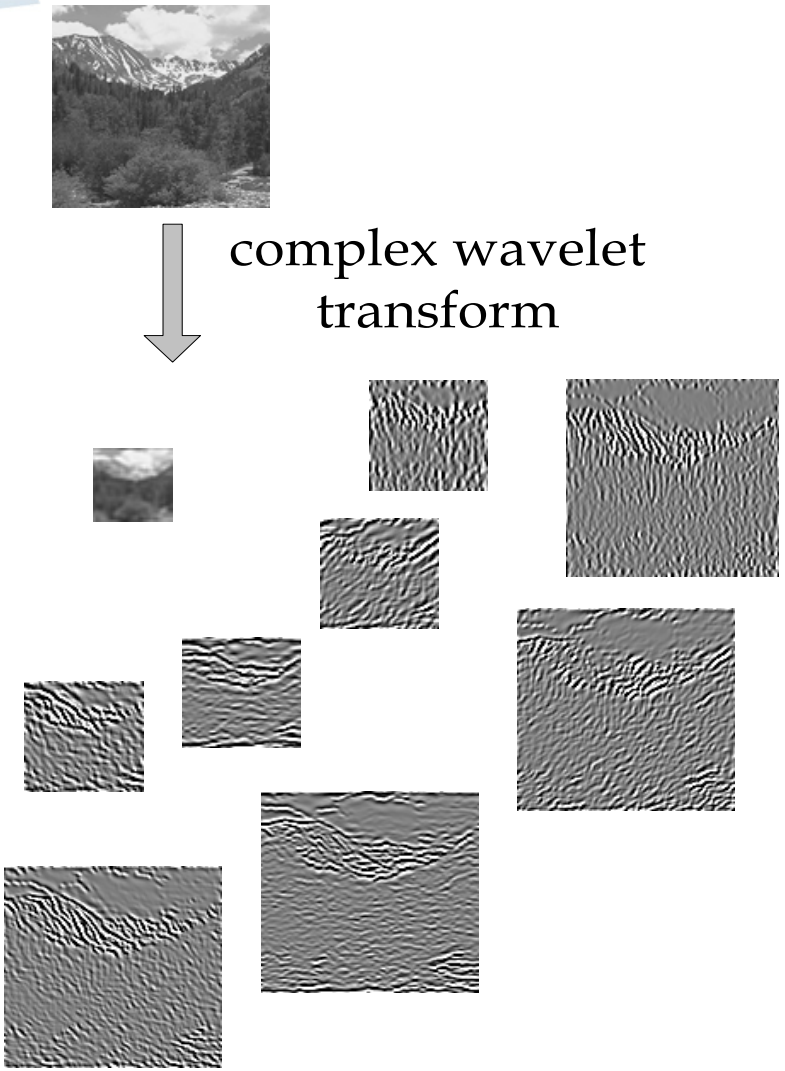
- Multi-scale SSIM

  [Wang, *et al*., Invited Paper, *IEEE Asilomar Conf.* '03]

- Complex wavelet SSIM

  [Wang & Simoncelli, *ICASSP* '05]

Extensions of SSIM (2)



- **Complex wavelet SSIM**
  - Motivation: robust to translation, rotation and scaling

$$\mathrm{SSIM}(\boldsymbol{x},\boldsymbol{y}) = \frac{2\left|\sum c_x \cdot c_y^*\right| + C}{\sum |c_x|^2 + \sum |c_y|^2 + C}$$

c_x, c_y : complex wavelet coefficients in images x and y

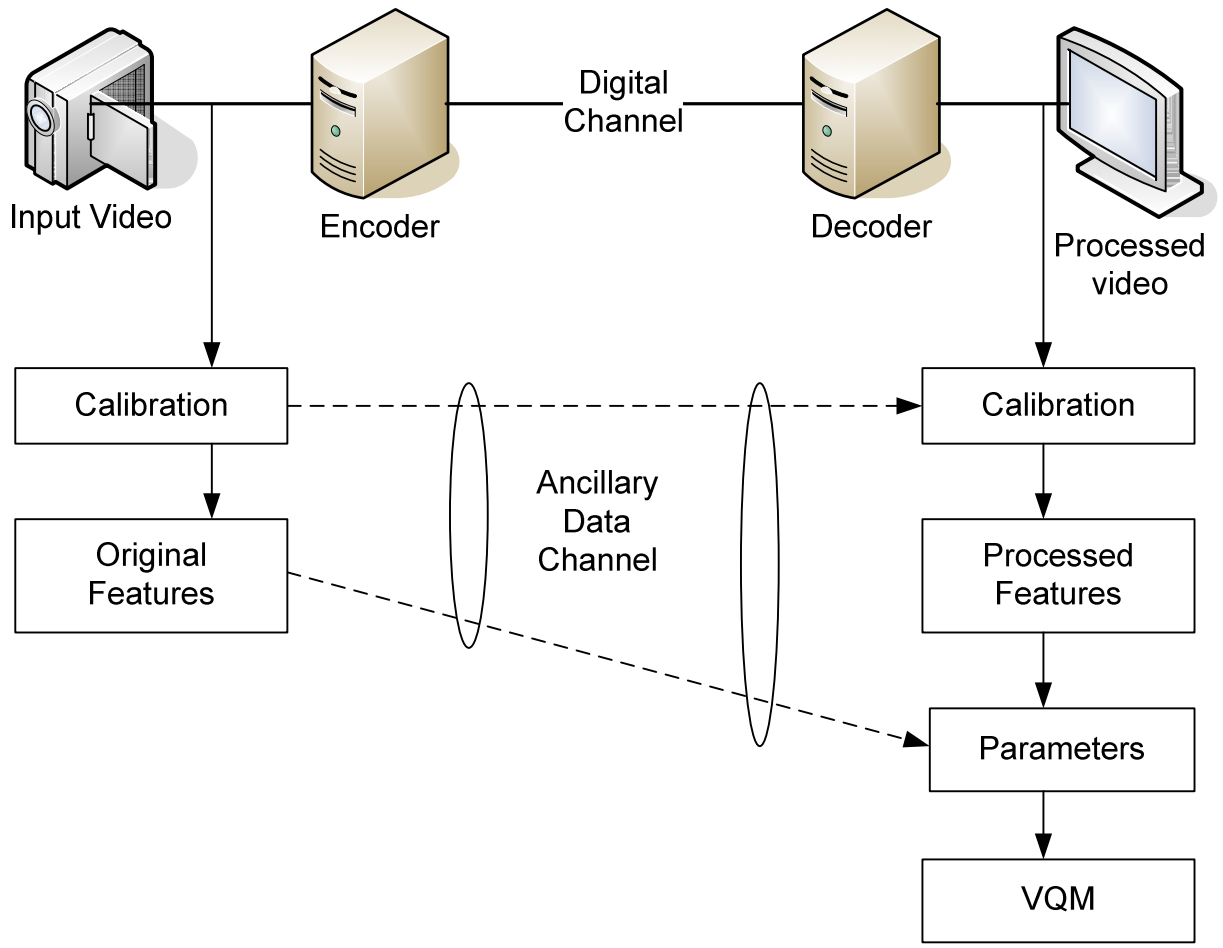
VQM (NTIA General Model)

- This metric for estimating video quality was developed by the National Telecommunications and Information Administration (NTIA)
- Was evaluated by the Video Quality Experts Group (VQEG) in their Phase II Full Reference Television (FR-TV) test
- The NTIA General Model was the only video quality estimator that was in the top performing group for both the 525-line and 625-line video tests

VQM (NTIA General Model)

- The ANSI adopted the NTIA General Model and its associated calibration techniques as a North American Standard in 2003
- The ITU has also included the NTIA General Model as a normative method in two Draft Recommendations.
- “A new standardization method for objectively measuring video quality”, IEEE Transactions on Broadcasting. [Pinson & Wolf, ‘04]
- Was successfully tested for HDTV scenarios [Wolf & Pinson, VPQM ’07]

VQM Model


Input Video
Encoder
Digital Channel
Decoder
Processed video
Calibration
Original Features
Ancillary Data Channel
Calibration
Processed Features
Parameters
VQM

VQM Model (Cont…)

- The General Model utilizes reduced-reference technology
- provides estimates of the overall impressions of video quality
- Reduced-reference measurement systems utilize low-bandwidth features that are extracted from the source and destination video streams
- The ancillary data channel is needed to transmit the extracted features (bandwidth of 9.3% of the uncompressed video sequence)

VQM Model (Cont…)

- The calibration of the original and processed video streams includes
  - spatial alignment
  - valid region estimation
  - gain & level offset calculation
  - temporal alignment
- VQM calculation involves
  - extracting perception-based features
  - computing video quality parameters
  - combining parameters to construct the General Model

VQM Model (Cont…)

- The General Model contains seven independent parameters
  - si_loss
    - detects a decrease or loss of spatial information (e.g., blurring)
  - hv_loss
    - detects a shift of edges from horizontal & vertical orientation to diagonal orientation
  - hv_gain
    - detects a shift of edges from diagonal to horizontal & vertical
  - chroma_spread
    - detects changes in the spread of the distribution of two-dimensional color samples

VQM Model (Cont…)

- The General Model contains seven independent parameters
  - si_gain
    - measures improvements to quality that result from edge sharpening or enhancements
  - ct_ati_gain
    - computed as the product of a contrast feature, measuring the amount of spatial detail, and a temporal information feature
  - chroma_extreme
    - detects severe localized color impairments, such as those produced by digital transmission errors

Comparison of Objective Evaluation Methods

Method	Mathematical Complexity	Calculation Capacity Needed	Chrominance Consideration	Correlation with Subjective Measures
PSNR	Simple	small	separate	poor
SSIM	Complex	medium-large	not included	good
VQM	Very Complex	large	included	fairly good

No reference Methods

- Estimating end-user's perception of a video stream without using an original stream as a reference

No reference / blind

- Complications
  - Unquantifiable factors when reference is not available include but not limited to:
    - Aesthetics
    - Cognitive relevance
    - Learning
    - Visual context
- Philosophy
  - All images/videos are perfect unless distorted during:
    - Acquisition
    - Processing
    - Reproduction

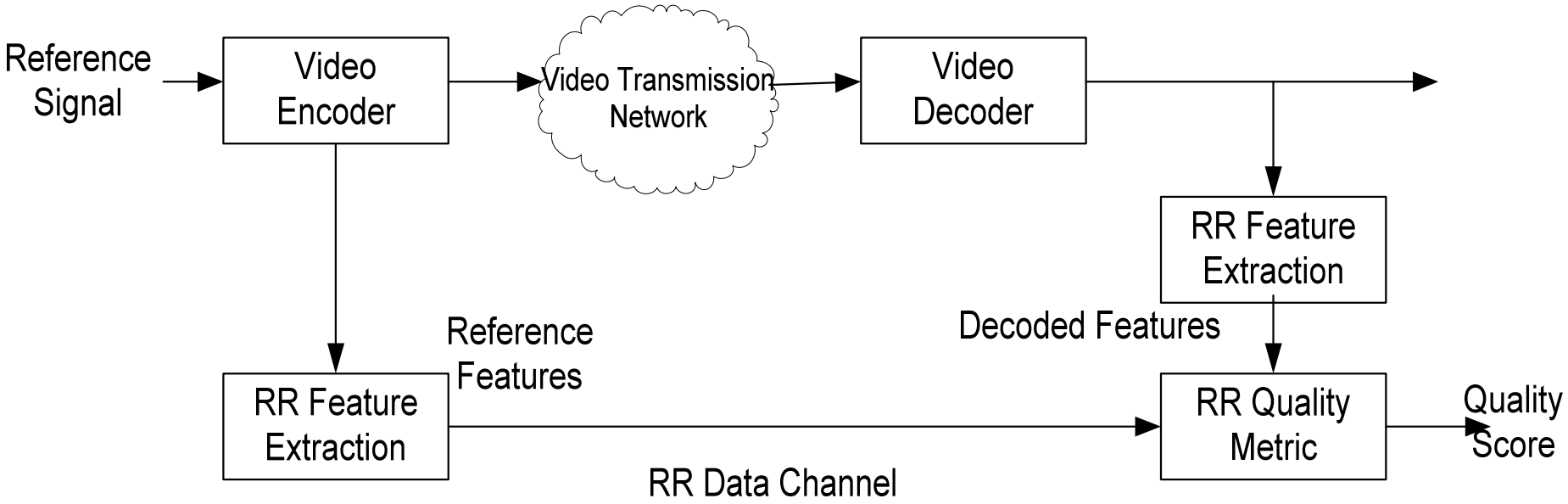
No reference

- Determining the possible distortion introduced during these stages
- Reference is “perfect” natural images/videos
  - Measured with respect to a model best suited to a given distortion type or application
    - E.g., natural images/videos do not contain blocking artifacts
- To improve prediction
  - Some HVS aspects are also modeled
    - Texture and luminance masking

Reduced Reference Methods

- Reference videos require tremendous amounts of storage space
- Reduced-reference (RR) quality assessment does not assume the complete availability of the reference signal
- only partial reference information (features that are extracted) is needed through an ancillary data channel

Deployment of a Reduced-reference Video Quality Assessment Metric



Subjective measures

- Mean opinion scores (MOS)
  - Endorsed and standardized by the ITU
  - Many obj. measures provide MOS
  - Requires mapping to labels (excellent, … bad)
  - shortcomings labels are inconsistent across languages, intervals not of equal size
- Binary measures: watchability, acceptability
  - Easy, low mental load on participants
  - Provides cut off points on “fit for purpose”
  - can be used for utility curves for service providers

Subjective Evaluation

- DSIS: Double Stimulus Impairment Scale
- DSCQS: Double Stimulus Continuous Quality Scale
- SSCQE: Single Stimulus Continuous Quality Evaluation
- SDSCE: Simultaneous Double Stimulus for Continuous Evaluation
- SCACJ: Stimulus Comparison Adjectival Categorical Judgement

Double Stimulus Impairment Scale

- Videos are shown consecutively in pairs
  - First one is the reference
  - Second one is impaired
- Expert must give opinion after playback using the opinion scale(5 the best and 1 the worst)
- Recency effect – Most recent clip has more effect on decision

Double Stimulus Continuous Quality Scale

- Videos are played in pairs
- Both videos are shown simultaneously
- Each pair is repeated a given number of times
- One of the videos is the reference and the other is the distorted one.
- Expert is not aware of the classification.
- Most commonly used method.
  - Especially when qualities are similar

Single Stimulus Continuous Quality Evaluation

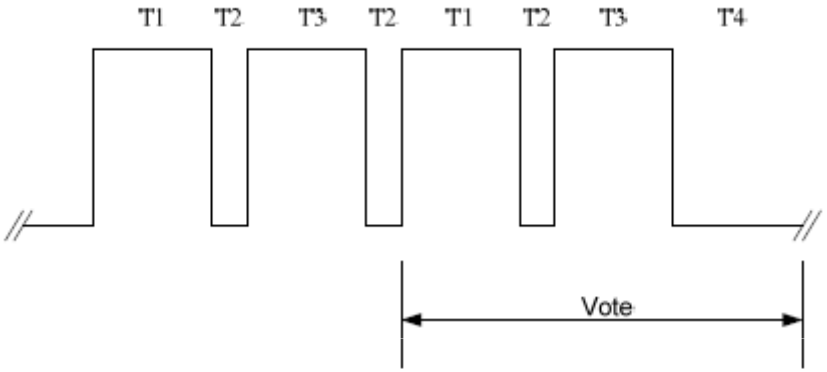
- Longer program(20 – 30 mins)
- Reference is not presented
- Continuous rating.
- Ratings are sampled throughout
- Rate changes between frames can be measured

- Memory effect
- Distractions with grading

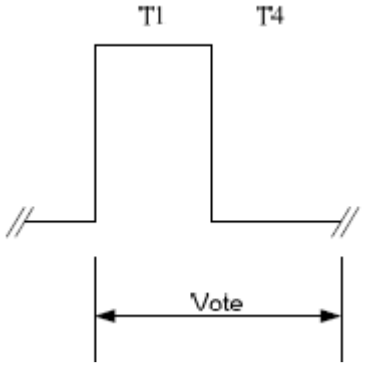
Simultaneous Double Stimulus for Continuous Evaluation

- Similar to the SSCQE, but videos shown simultaneously

Presentation Structure of Material



a. DSCQS method



b. SSCQS method

Phases of presentation:

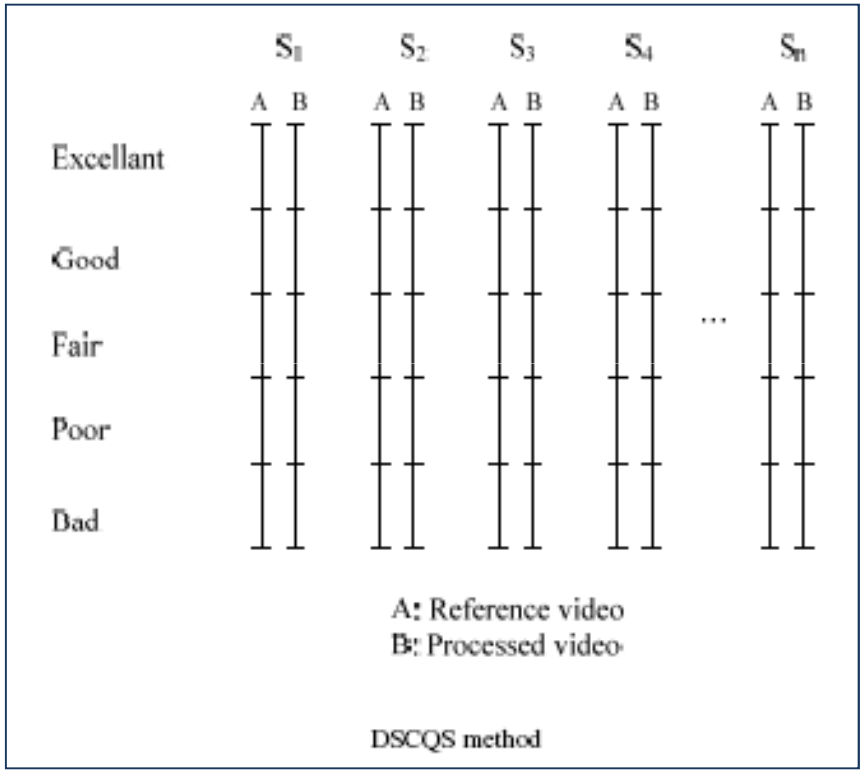
T1 = 10 seconds: Test sequence A

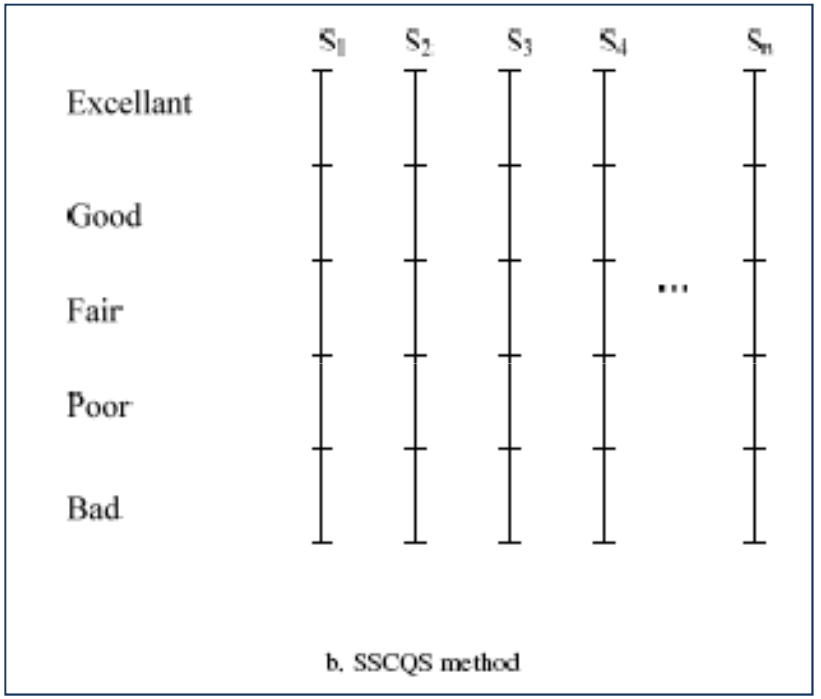
T2 = 3 seconds: Mid-gray B

T3 = 10 seconds: Test sequence B

T4 = 5-11 seconds: Mid-grey

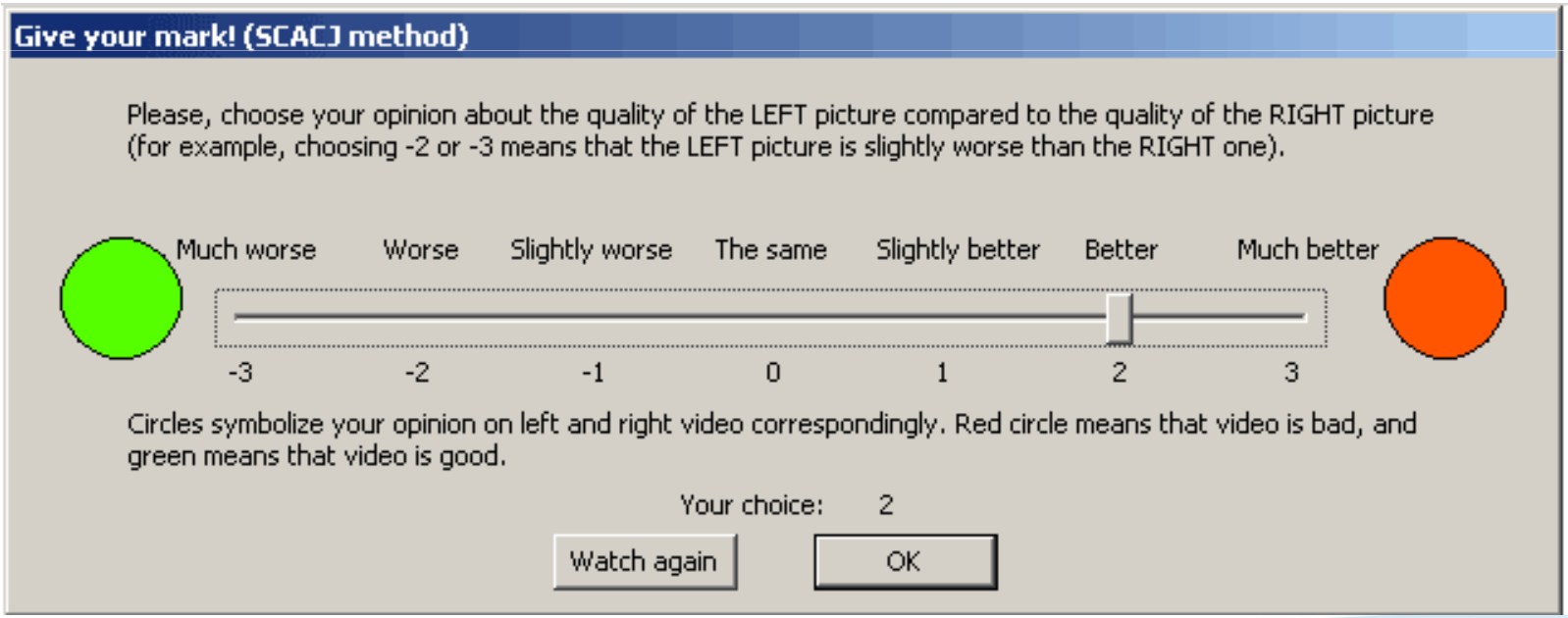
Quality rating scale for subjective assessment





Stimulus Comparison Adjectival Categorical Judgement

- Two sequences played simultaneously
- Expert has to give opinion after that on this scale

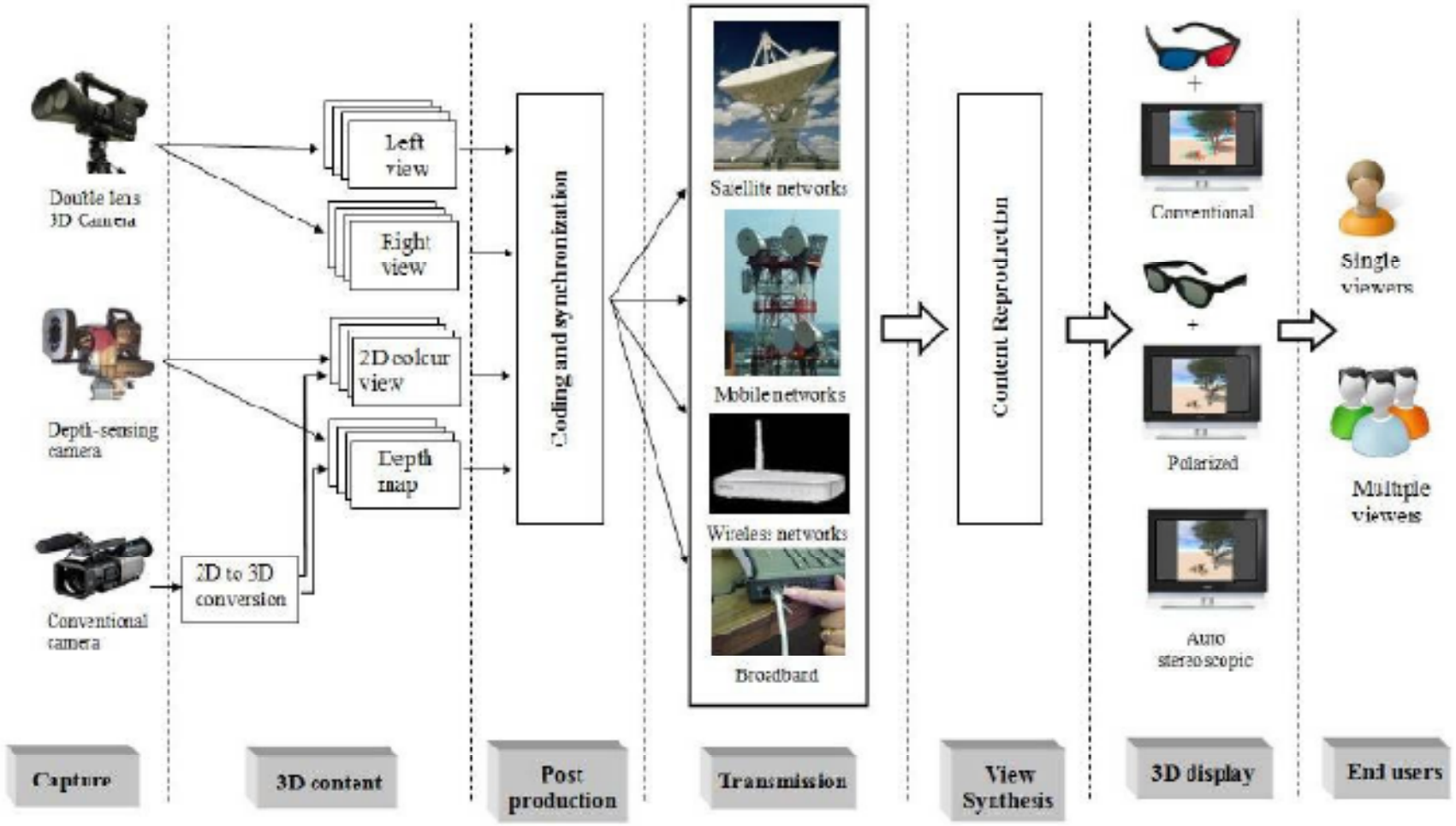


Parameter	DSIS	DSCQS	SSCQE	SDSCE
Explicit reference	Yes	No	No	Yes
Hidden reference	No	Yes	No	No
High anchor	No	Yes	No	No
Low anchor	No	Yes	No	No
Scale	Bad to excellent	Bad to excellent	Bad to excellent	Bad to excellent
Sequence length	10s	10s	5 min	10s
Picture format	All	All	All	All
Two simultaneous stimuli	No	No	No	Yes
Presentation of test material	I: Once II: Twice in succession	Twice in succession	Once	Once
Voting	Only test sequence	Test sequence and reference	Test sequences	Difference between the test sequence and the reference simultaneously shown
Possibility to change the vote before proceeding	No	No	No	No
Continuous quality evaluation	No	No	Yes (moving slider in a continuous way)	Yes (moving slider in a continuous way)
Minimum accepted votes	15	15	15	15
Assessors per display	One or more	One or more	One or more	One or more
Display	Mainly TV	Mainly TV	Mainly TV	Mainly TV

5.Introduction to 3D Video

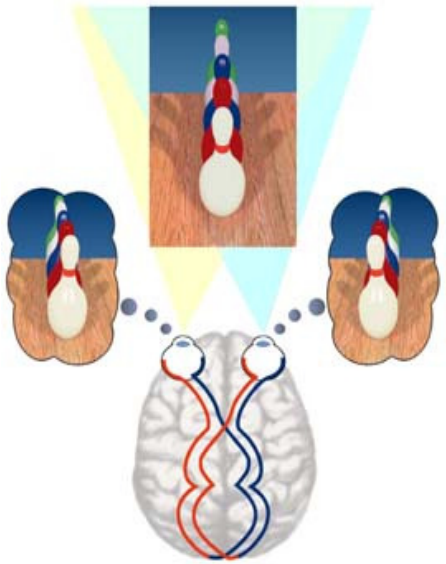
- 3D Video has received great interest in recent years.
  - achieve a more involving and immersive representation of visual information
  - provide more natural methods of communication

End-to-End 3D Video Chain



Stereoscopic Video

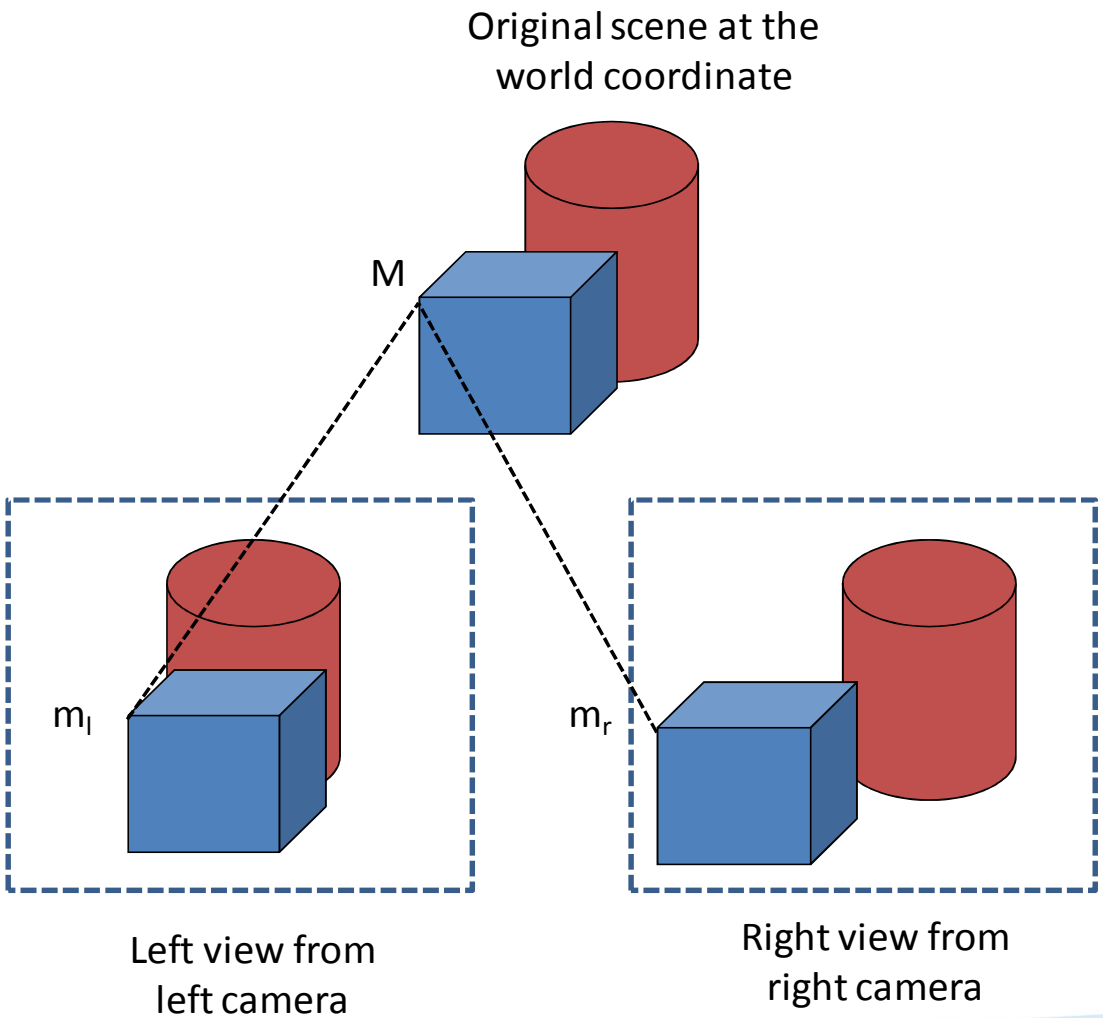
The left and right views are fused in the visual cortex of the brain to perceive the depth of a scene



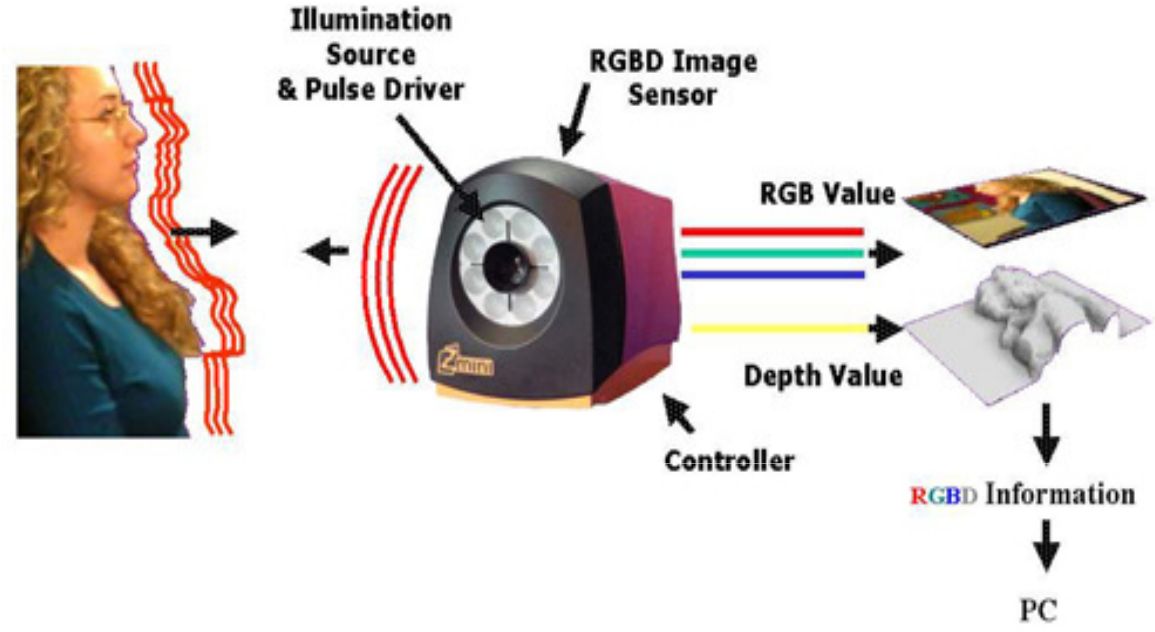
- Advantages of stereo video over other representations of 3D video
  - Simple representation
  - Cost effective display systems
  - Easy to adapt for existing audio-visual communication technologies

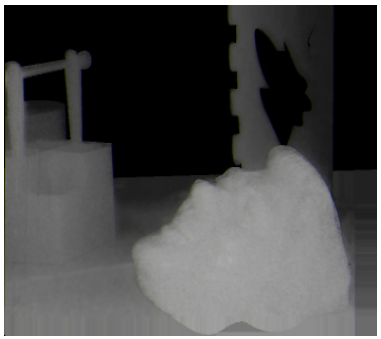
Left-Right-View 3D Representation

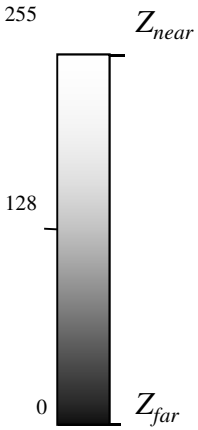
– Left and Right Cameras



Colour-Plus-Depth 3D Representation







- In case of 8-bits YUV format, the depth value is stored in Y-component with value between 0-255.

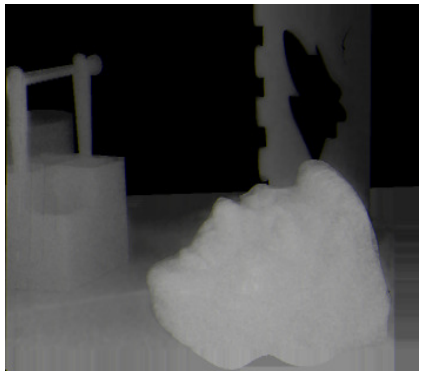
- The nearest distance is represented by white colour (255) and the furthest distance is represented by black colour (0).

Depth Image-Based Rendering

Colour

+

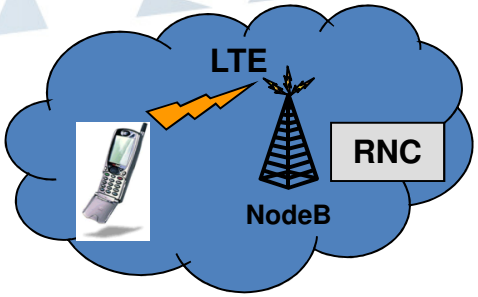
Depth

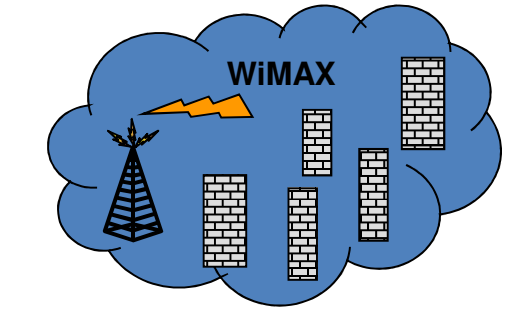


Left View

Right View

Motivations





3DTV

Mobile 3DTV

Large Terminal
Group of co-located users with a large fixed terminal

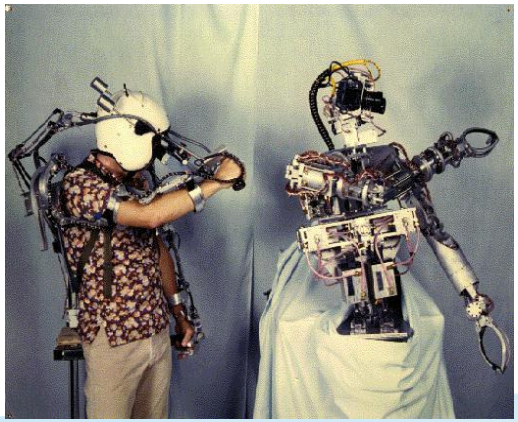
Virtual Collaboration System

Mobile terminal
Single remote user with a mobile terminal

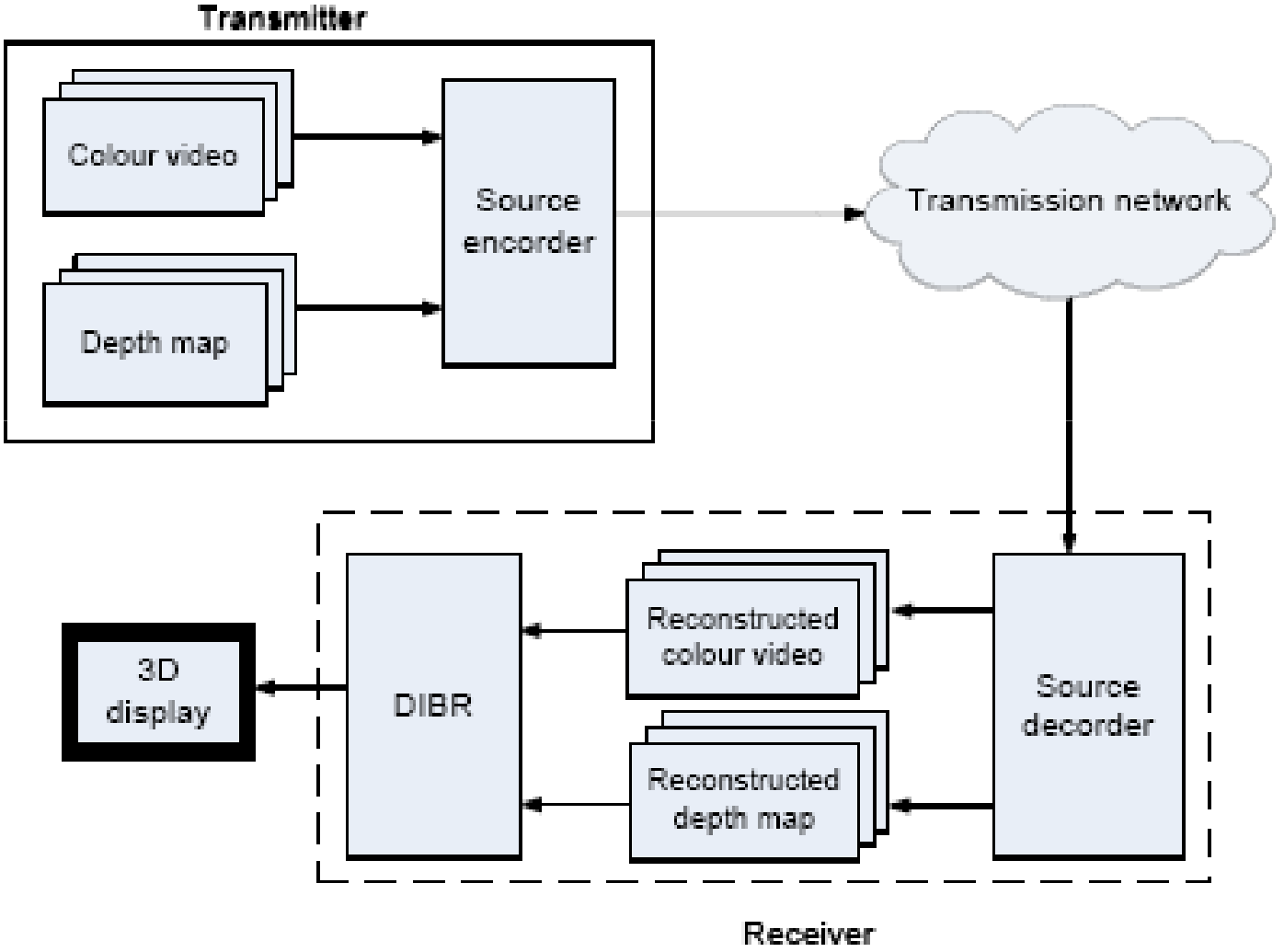
Small Terminal
Single remote user with a small fixed terminal

The potential applications are:

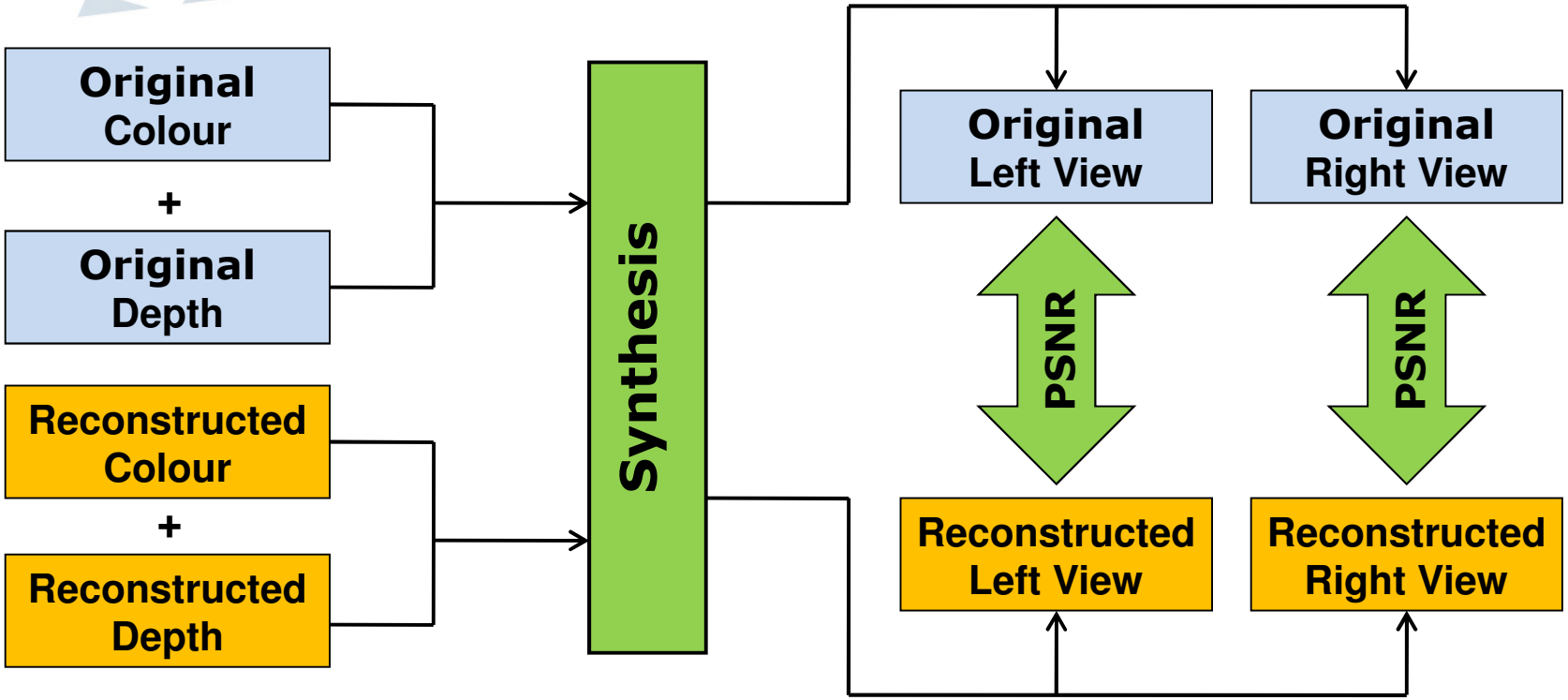
- Tele-conferencing
- Tele-robotics
- Tele-operation
- 3D-television (3D-TV)



Colour-plus-depth 3D video communication system



6. 3D-Objective Evaluations



- 3D quality is objectively evaluated in terms of average PSNR between the reconstructed left-right views and the original left-right views.

Visual experience assessments

- Experimentation setup for
  - 42” Philips WOWvx multi-view auto-stereoscopic display
  - 28 non-expert volunteers, who were aged between 20 and 40 years
    - Colour vision was verified
    - Eye acuity was tested
  - MVD1 test sequences
  - Double Stimulus Continuous Quality Scale (DSCQS) method specified in ITU-R BT.500 standard

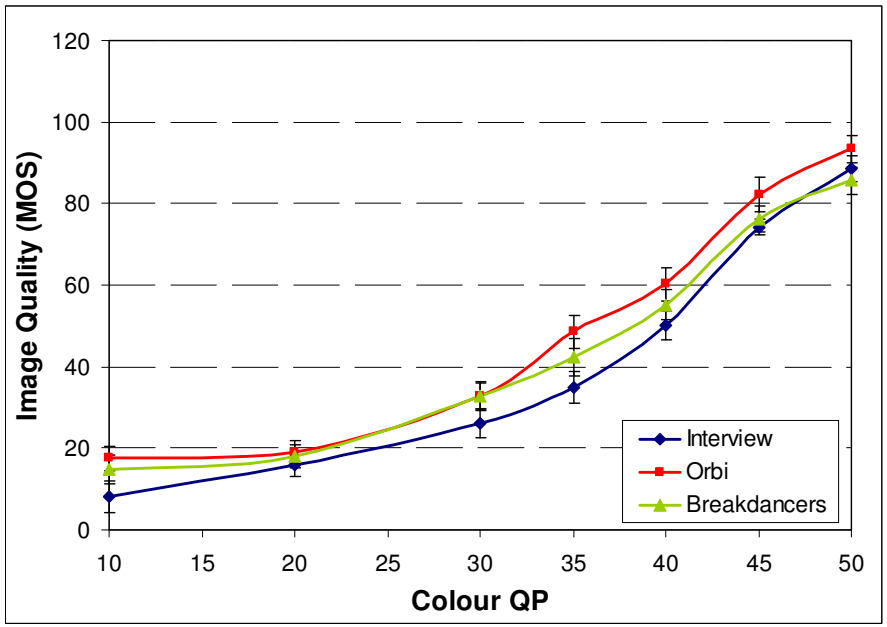
The logistic function

- Relationship between the objective quality assessment models and the subjective quality ratings given by the observers (i.e., MOS) were approximated the symmetrical logistic function

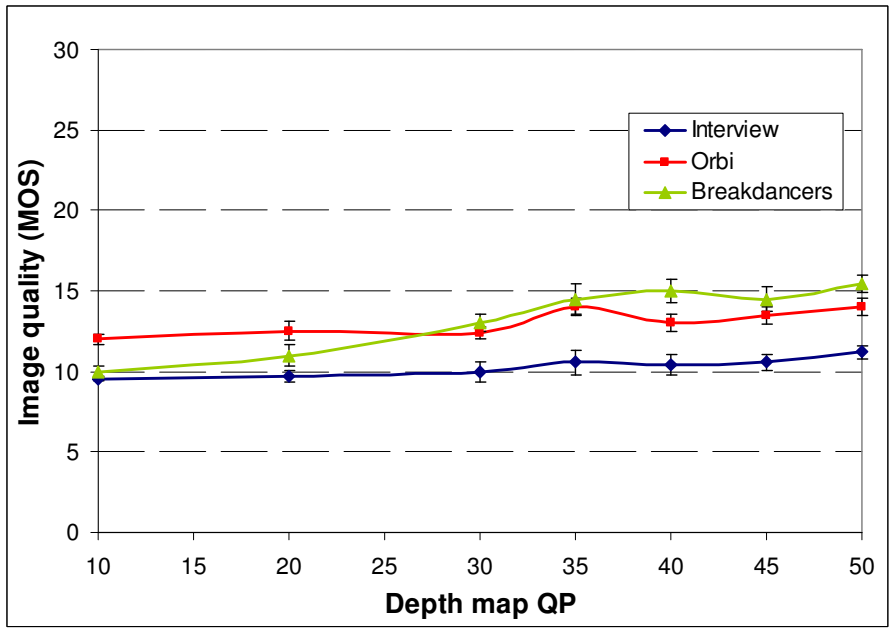
$$p = \frac{1}{1 + e^{(D - D_M) \cdot G}}$$

- Performance comparison metrics
  - Correlation Coefficient (CC)
  - Root Mean Squared Error (RMSE)
  - Sum of Squares due to Error (SSE)

Effect of compression on the image quality



Perceived image quality (MOS) vs video quantisation parameter



Perceived image quality (MOS) vs depth map quantisation parameter

Conclusion: Image quality mostly depends on the quality of the video component of MVD1 contents

Assessing available objective image quality metrics

- Assessing the available objective image quality metrics on MVD1 contents
  - Metrics to be assessed:

$$Obj = \{PSNR, SSIM, VQM\}$$

  - Measure 1 - Average of the objective quality of rendered left and right views:

$$Measure1 = \frac{1}{2}\left(Obj_L + Obj_R\right)$$

  - Measure 2 - Weighted average based on the number of left and right eye dominant assessers:

$$Measure2 = \frac{1}{N_{tot}}\left(N_L \cdot Obj_L + N_R \cdot Obj_R\right)$$

Assessing available objective image quality metrics

– Measure 3 - Weighted average based on the eye acuity of assessers:

$$Measure3 = \frac{1}{N_{tot}} \left(\frac{\sum_{i=1}^{n} Acuity_{L,i}}{Acuity_{Max}} \cdot Obj_L + \frac{\sum_{i=1}^{n} Acuity_{R,i}}{Acuity_{Max}} \cdot Obj_R \right)$$

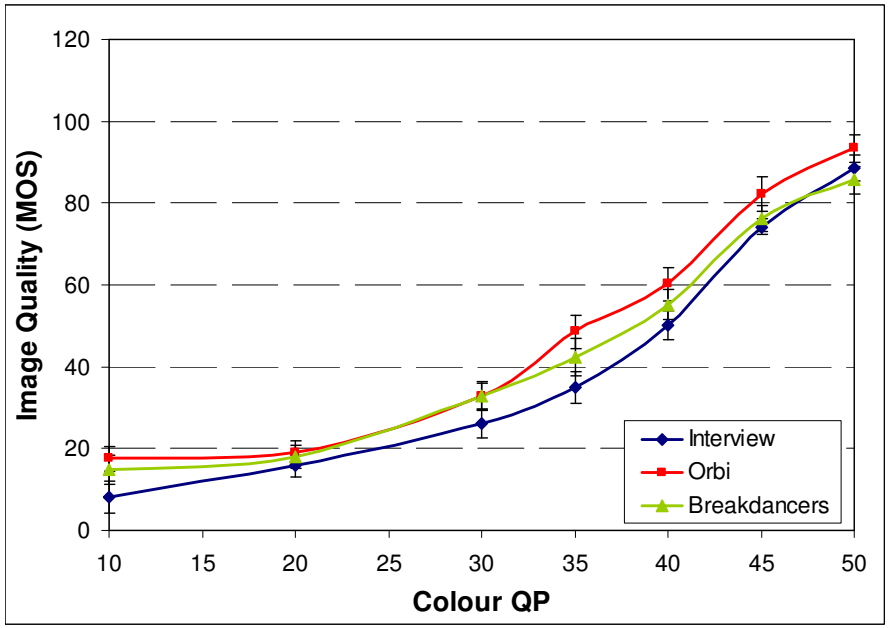
– Measure 4 - Quality of the colour image:

$$Measure4 = Obj_{colour}$$

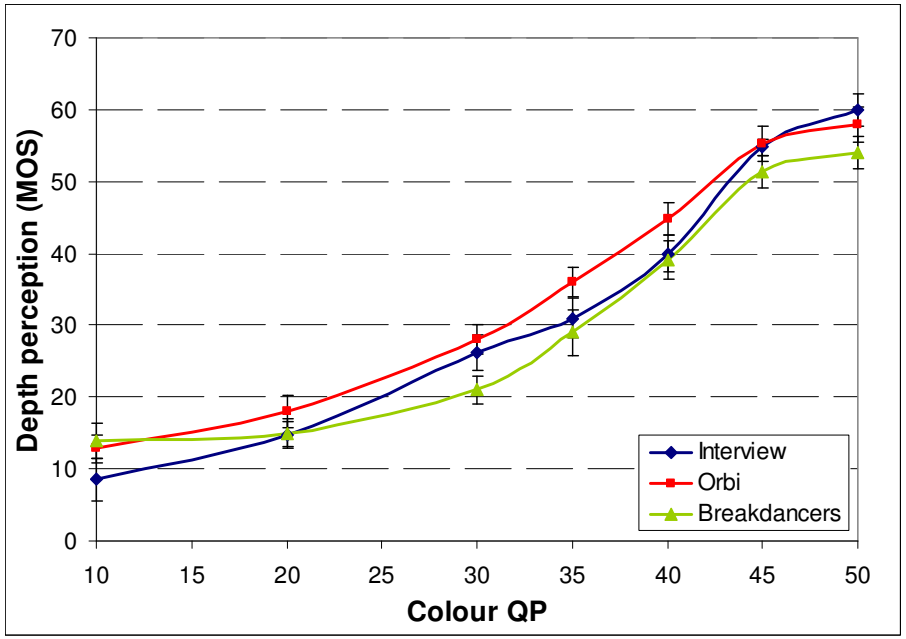
Performance analysis

	Objective Quality measures	Correlation between the measure and the MOS		
		CC	RMSE	SSE
PSNR	*Measure1*	0.8808	0.0696	0.06226
	Measure2	0.8811	0.06957	0.06223
	Measure3	0.8885	0.06881	0.06148
	Measure4	0.8917	0.06763	0.06041
SSIM	*Measure1*	0.8424	0.1798	0.09934
	Measure2	0.8484	0.1708	0.09026
	Measure3	0.8692	0.1527	0.07934
	Measure4	0.936	0.04241	0.0439
VQM	*Measure1*	0.9433	0.03681	0.04174
	Measure2	0.9189	0.06201	0.05568
	Measure3	0.9321	0.04483	0.04514
	Measure4	**0.96**	**0.03506**	**0.03992**

Effect of compression on the depth perception



Depth perception (MOS) vs video quantisation parameter



Depth perception quality (MOS) vs depth map quantisation parameter

Conclusion: Depth perception is sensitive to the quality of both video and depth map component of MVD1 contents

Modelling of depth perception

- Human brain independently perceive the degradation in monocular and binocular cues.

$$Depth\_perception = D_M^{\alpha} \cdot D_B^{\beta}$$

D_M : *Contribution of monocular cues*

D_B : *Contribution of binocular cues*

α *and* β: *+ve constants*

Contribution from the Colour Texture Video

- PSNR, SSIM and VQM are considered
- Subjective experiment
  - Uniform depth map was used
    - Users effectively see 2D video
  - Quality of the colour video is varied
  - Depth perception attribute was rated

Subjective Vs. Objective

- The relationship between the subjective quality (MOS) and objective measures is approximated by the symmetrical logistic function
- The quantitative measures for each prediction model
  - Correlation Coefficient (CC)
  - Root Mean Squared Error (RMSE)
  - Sum of Squared Error (SSE)

Performance analysis

Objective Quality Model	Depth perception (monocular)		
	CC	RMSE	SSE
PSNR	0.8226	0.1659	0.0934
SSIM	0.8431	0.1027	0.0735
VQM	0.9129	0.1007	0.0596

CC=1, RMSE=0 and SSE=0 perfect correlation
CC=0, RMSE=1 and SSE=1 worst correlation

Contribution from the depth map

- Depth map is not a natural image
  - Existing 2D quality metrics are less suitable
- Disparity Distortion Model (DDM) is proposed
  - HVS identifies depth by proper visual recognition of depth-planes of binocular vision
  - Factors that serve in visualization,
    - Relative distance to different depth planes
    - Consistency of the content in depth planes
    - Structural information [4]

[4]- Wandell, B.A., Foundations of vision, Sinauer Associates, Inc, Sunderland, Massachusetts, USA, 1995.

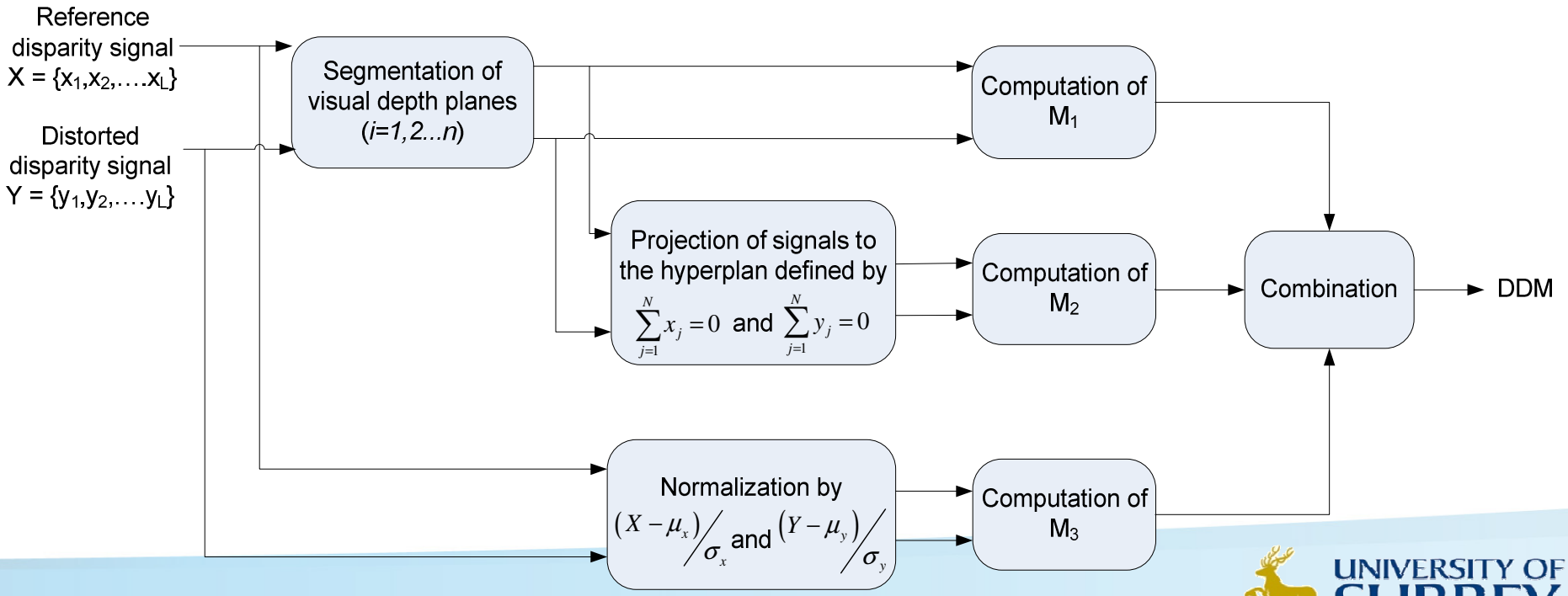
7. Disparity Distortion Model (DDM)

$$DDM = f\left(M_1(x,y), M_2(x,y), M_3(x,y)\right)$$

M_1 - distortion of the relative distance in depth axis among depth planes

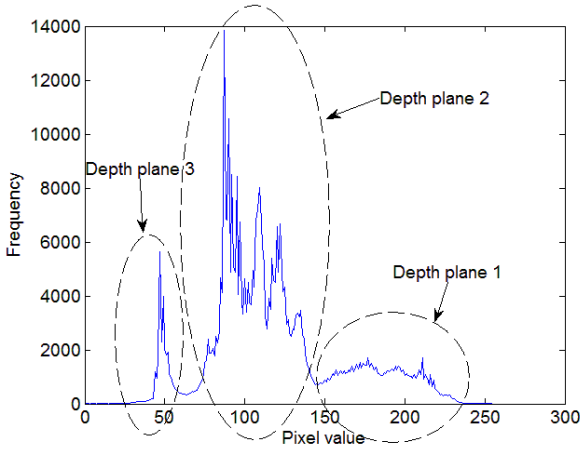
M_2 - distortion of the consistency in perceived depth within the depth planes

M_3 - structural comparison

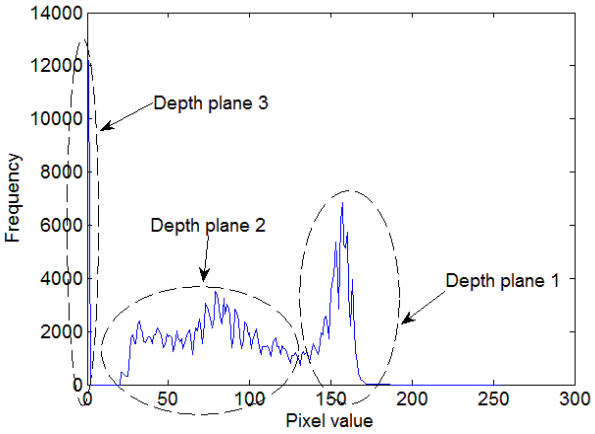


DDM: Visual DPs (1)

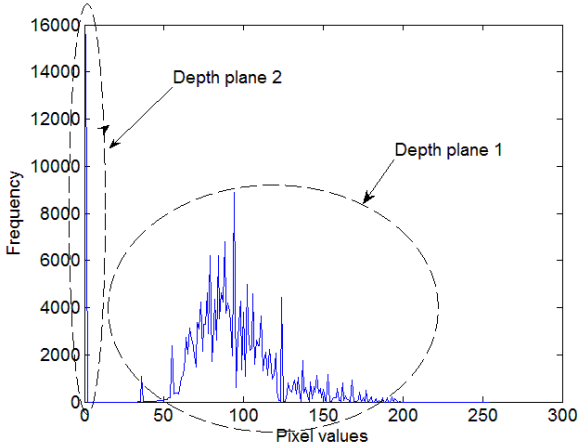
- Visually recognized depth planes (DPs)
  - Histogram of the disparity signal is examined and visually recognized depth planes are identified



"*Breakdancers*" sequence



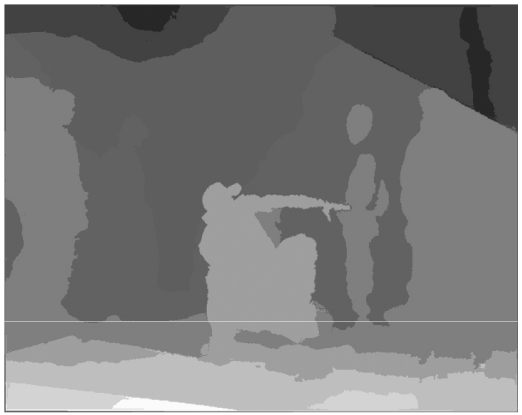
"*Orbi*" sequence



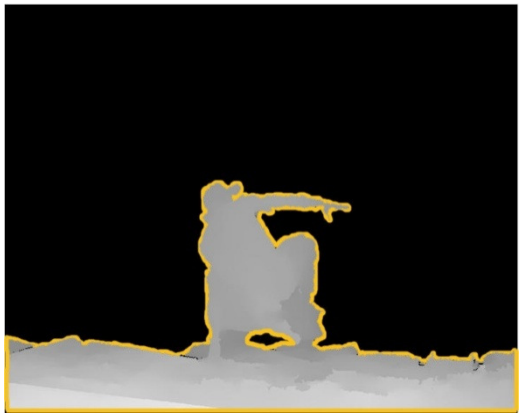
"*Interview*" sequence

DDM: Visual DPs (2)

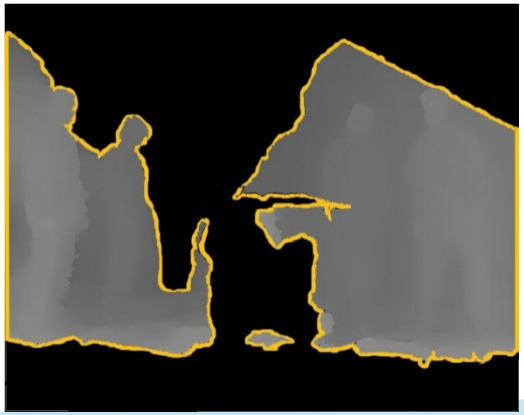
- Depth map segmented according to Visual DPs



Original depth image



Segmented depth plane 1



Segmented depth plane 2

Segmented depth plane 3

DDM: M1

- M1: Distortion of the relative distance (DRD) in depth axis among depth planes

Relationship between the depth value *Z* and gray scale bit value *m*

$$z = \frac{m}{255} \cdot \left(k_{near} W + k_{far} W\right) - k_{far} W \quad \text{--- (a)}$$

The distortion of the relative distance between adjacent depth planes *i* and *i+1*

$$z_x^i - z_x^{i+1} = \frac{\left(k_{near} W + k_{far} W\right)}{255} \cdot \left(m_x^i - m_x^{i+1}\right) \text{--- (b)}$$

$$z_y^i - z_y^{i+1} = \frac{\left(k_{near} W + k_{far} W\right)}{255} \cdot \left(m_y^i - m_y^{i+1}\right) \text{--- (c)}$$

i – depth plane index
n – number of DPS
m_x m_y- mean intensity of DPs

$$DRD^{i,i+1} = \frac{\left[k_{near} W + k_{far} W\right]}{255} \cdot \left|\left(m_x^i - m_x^{i+1}\right) - \left(m_y^i - m_y^{i+1}\right)\right| \quad \text{--- (d)}$$

$$M_1 = \sum_{i=1}^{n} DRD^{i,i+1} \quad \text{--- (e)}$$

DDM: M2 and M3

- **M2: Distortion of the consistency in the perceived depth within the depth planes**

Error signal, $e^i = x^i - y^i$ --- (a)

For depth plane *i*, $\sigma_e^i = \left[\frac{1}{N} \sum_{j=1}^{N} \left(e_j^i - \mu_e^i \right) \right]^{\frac{1}{2}}$ --- (b)

N – number of pixels in DP
j – pixel index
n – number of DPS

For all depth planes, $M_2 = \sum_{i=1}^{n} \sigma_e^i$ --- (c)

- **M3: Structural comparison (SC) of the depth maps**
  - Evaluated over 16 x 16 macroblocks of the entire image

$$SC_{[6]} = \frac{\sigma_{xy} + k_1}{\sigma_x \cdot \sigma_y + k_1} \text{ --- (d)} \qquad M_3 = \frac{1}{m} \sum_{j=1}^{m} SC_j \text{ --- (e)}$$

σ_{xy} – covariance of x and y
m – number of macroblocks

Depth perception model

- Mean disparity distortion measure (MDDM)

$$MDDM(X,Y)=\frac{1}{M}\sum_{j=1}^{M}\left[\frac{M_3(x_j,y_j)}{M_1(x_j,y_j)\cdot M_2(x_j,y_j)+k_2}\right] \quad \text{--- (a)}$$

M – Number of frames

- Proposed depth perception model

$$Depth\_perception=(1-VQM)^{\alpha}\cdot MDDM^{\beta} \quad \text{--- (b)}$$

$$\alpha=1.5$$

$$\beta=1$$

Performance analysis

- Test 4 : Quality of both colour texture video and depth map are varied
  - Subjects rated depth perception

Objective Quality Model	Depth perception		
	CC	RMSE	SSE
Average PSNR of the Rendered Left and Right videos	0.7788	0.0737	0.0579
Average SSIM of the Rendered Left and Right videos	0.8065	0.0674	0.0547
Average VQM of the Rendered Left and Right videos	0.7753	0.0739	0.0603
Proposed Depth Perception Model	0.8716	0.0325	0.0379

CC=1, RMSE=0 and SSE=0 perfect correlation
CC=0, RMSE=1 and SSE=1 worst correlation

Overall visual QoE

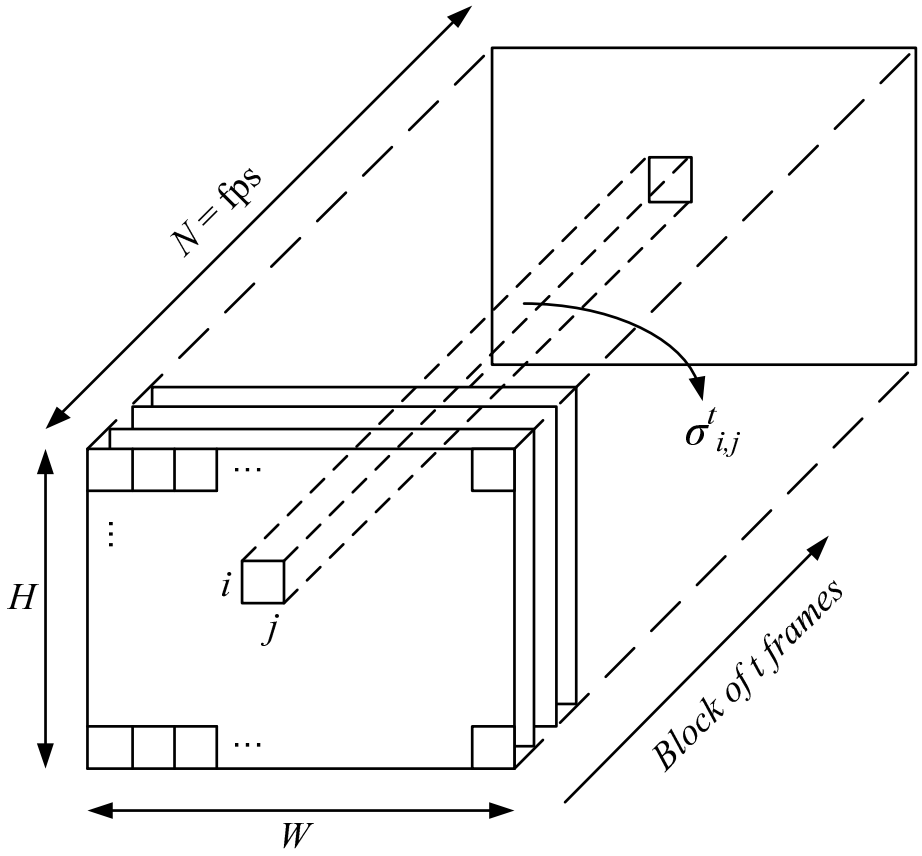
$$QoE_{video} = f(IQ, DP, cxt)$$

$$QoE_{video} = f_1(content) \cdot IQ + f_2(content) \cdot DP$$

$$f_1(content) + f_2(content) = 1$$

- z-direction (depth direction) motion activity (ZMA) of 3D video is used to model the weighting functions f_1 and f_2

Overall visual QoE



- Z-direction (depth direction) motion activity (ZMA)
  - Depth map component of the 3D video is segmented into blocks of *N* frames, where *N* is the frame rate,
  - The standard deviation of each pixel position is evaluated for each block
  - ZMA is then normalized (nZMA)

$$\sigma^t_{Y_{i,j}} = \left[\frac{1}{N} \sum_{k=1}^{N} \left(Y^k_{i,j} - \mu^t_{Y_{i,j}} \right) \right]^{\frac{1}{2}}$$

$$ZMA = \frac{1}{M} \sum_{t=1}^{M} \left[\sum_{j=1}^{H} \sum_{i=1}^{W} \sigma^t_{Y_{i,j}} \right]$$

$$nZMA = \frac{ZMA}{(W \cdot H) \times (2^8 - 1)}$$

Overall visual QoE

- Based on the subjective experiments to assess the overall 3D visual perception

$$f_1(nZMA) = 1 - 0.997 \cdot nZMA^{0.2393}$$

$$f_2(nZMA) = 0.997 \cdot nZMA^{0.2393}$$

- Overall QoE model:

$$QoE_{video} = \left(1 - 0.997 \cdot nZMA^{0.2393}\right) \cdot \left(1 - VQM_{colour}\right) + \left(0.997 \cdot nZMA^{0.2393}\right) \cdot \left[\left(1 - VQM_{colour}\right) \cdot MDDM_{depth\_map}\right]$$

Performance analysis of the overall visual QoE model

QoE model	Overall 3D video quality		
	CC	RMSE	SSE
Average PSNR of the Rendered Left and Right views	0.7061	0.1363	0.1091
Average SSIM of the Rendered Left and Right views	0.7387	0.0949	0.0887
Average VQM of the Rendered Left and Right views	0.8092	0.0570	0.0501
Proposed Compound 3D Quality Model	0.8461	0.0337	0.0319

QUALITY MODEL FOR STEREO VIDEO CONTENTS

Subjective experimentation setup

- Objective
  - To subjectively assess the impact of distortion in each view on the perceived image quality, depth perception and the overall 3D visual experience
- MVD2 video sequences are used for this experiment
  - The video is encoded using JMVC codec
    - Left view is coded independently
    - Right view is coded with respect to the left view (or depth map)
  - Four different QP settings are used: 20, 25, 30, and 40
- The sequences were displayed on JVC passive stereoscopic display

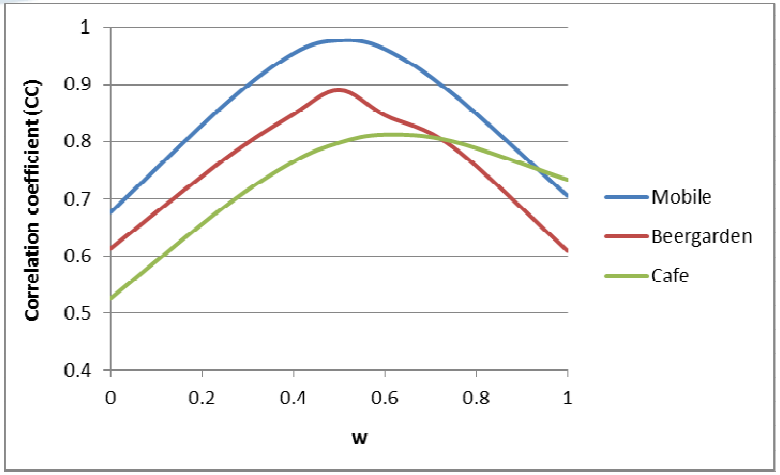
Experimental results

- Correlation between subjective and objective results
  - Distortion metric:

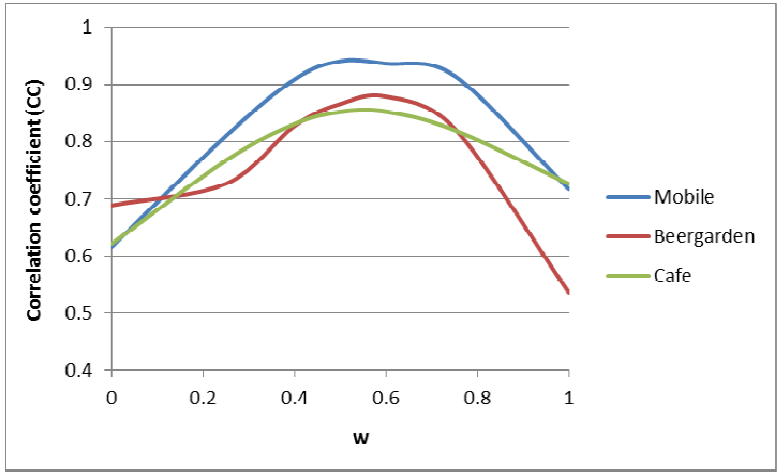
$$D = (1 - w).VQM_{left} + w.VQM_{right}$$

VQM_{left} – VQM of the left view
VQM_{left} – VQM of the right view
w – weighing factor

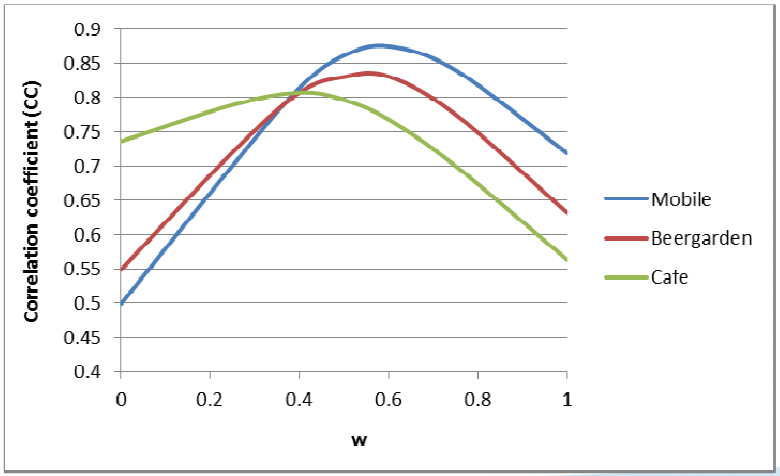
Experimental results



Correlation coefficient vs. *w* for image quality prediction

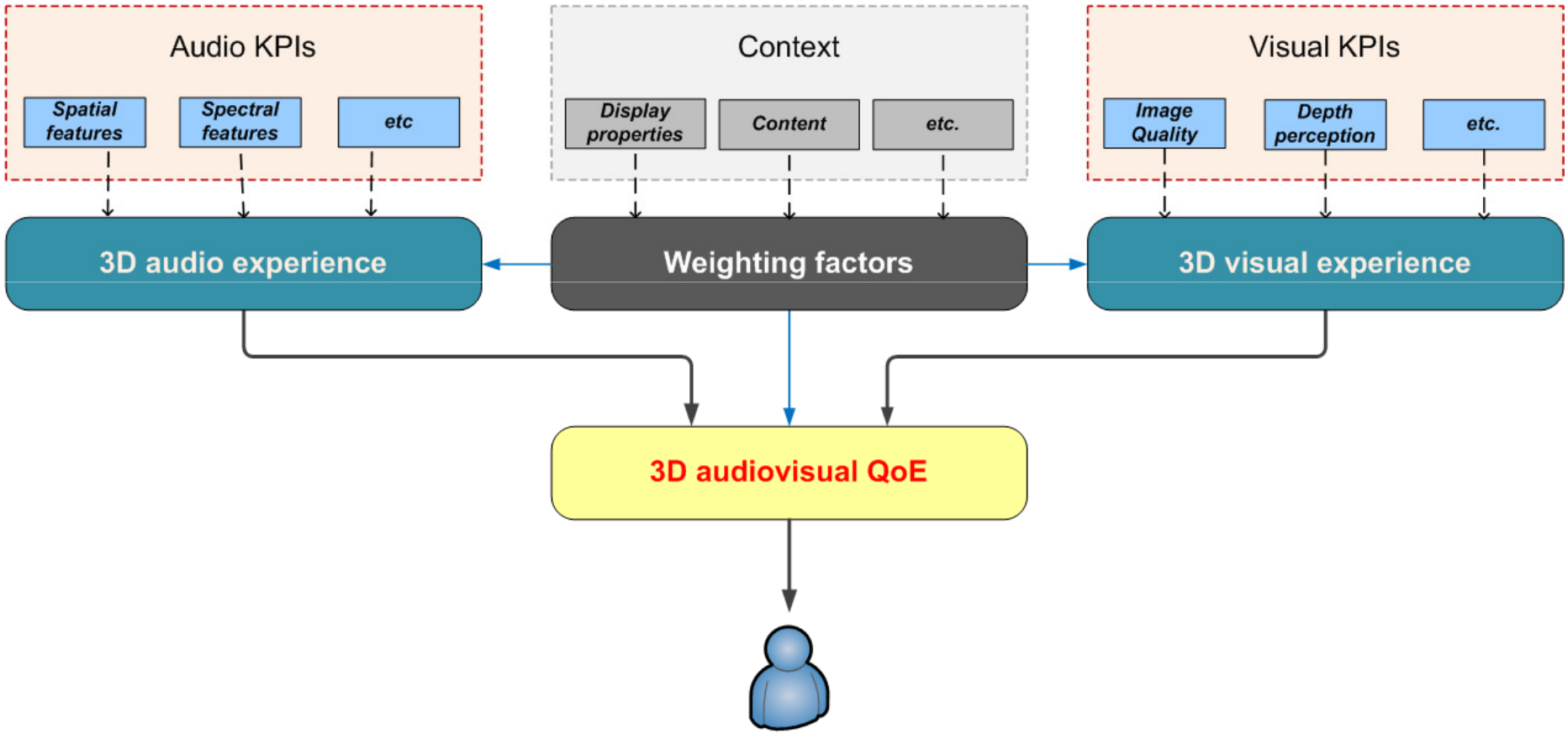


Correlation coefficient vs. *w* for depth perception prediction

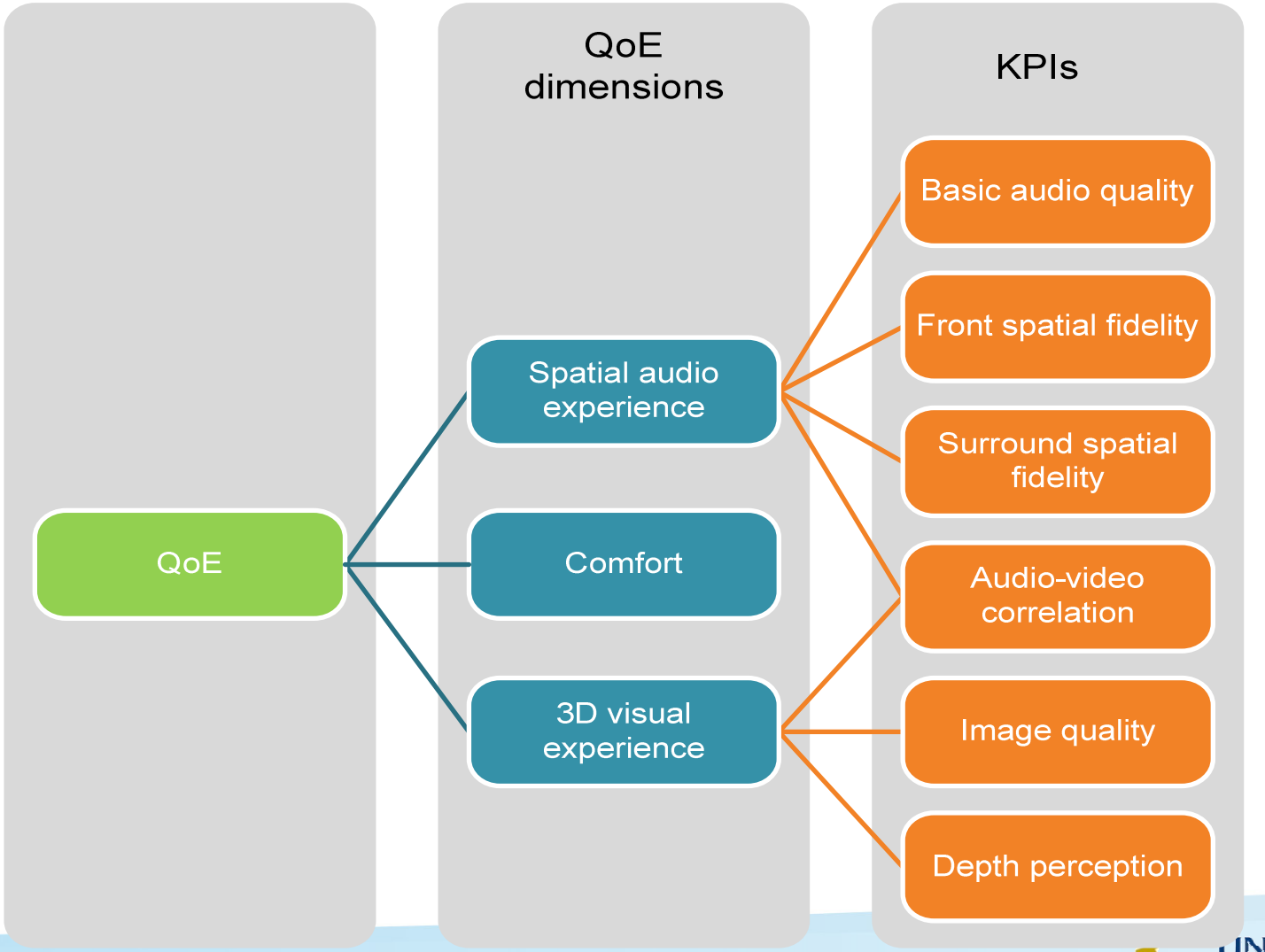


Correlation coefficient vs. *w* for overall 3D visual experience prediction

8. Quality of Experience



QoE dimensions and KPIs



Spatial audio experience

Factors:

- Original material quality
  - Basic audio quality
  - Front spatial fidelity
  - Surround spatial fidelity
  - ✓Some models exist
- Reproduction system fidelity in reproducing front and surround spatial material
  - Systems: Binaural, stereo, 5.1 home theatre, ambisonics, and WFS.
  - ✓Some comparison results exist
- Interaction of audio with video (concerns front spatial fidelity)
  - Temporal and spatial synchronicity
  - ✓Testing needed
- Listening environment and listener
  - Acoustical characteristics, ambient noise, etc.
  - Listener demographics (Age, gender, cultural background, occupation, etc.), previous experience, mood, etc.
  - ✓ Testing needed
- Production type (News, movies, sports, advertisements, etc.)
  - ✓ Testing needed

✓**Further testing to combine these factors**

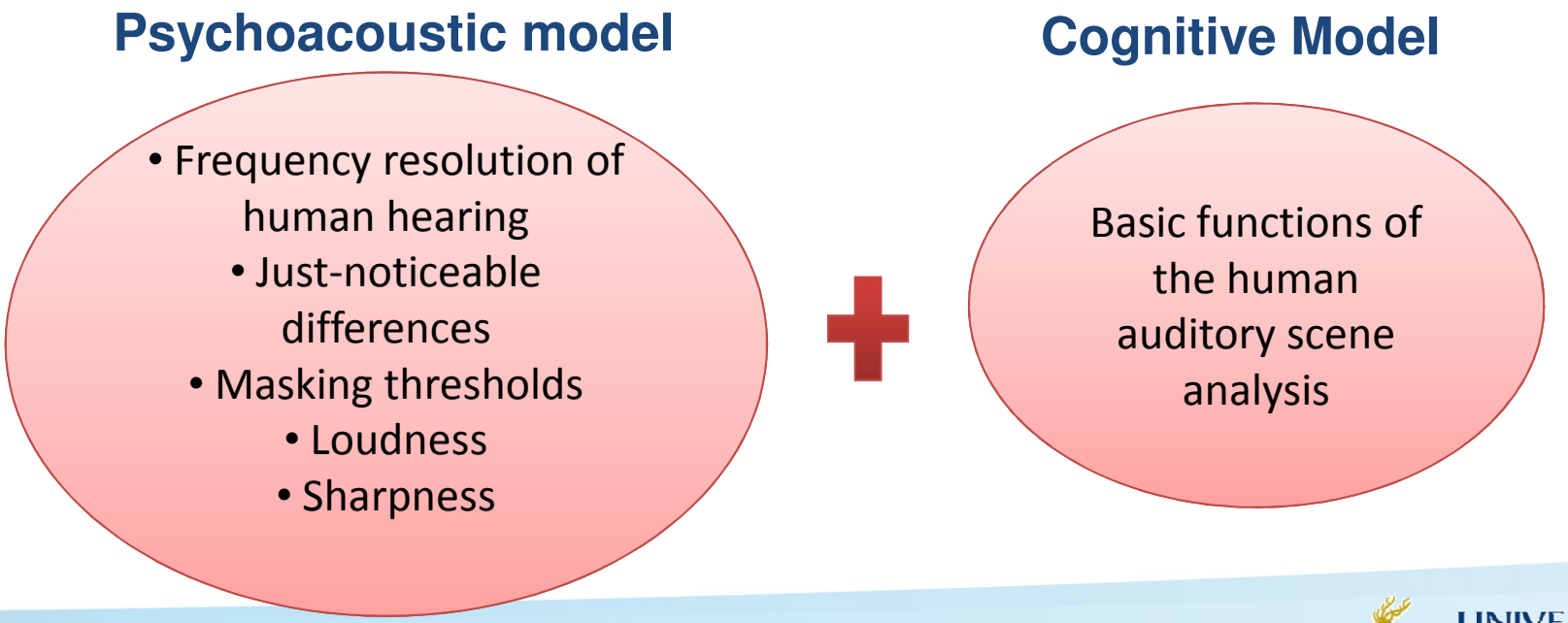
Spatial audio experience

Spatial Audio KPIs

- Basic Audio Quality
- Front spatial fidelity
- Surround spatial fidelity
- Correlation of audio and video

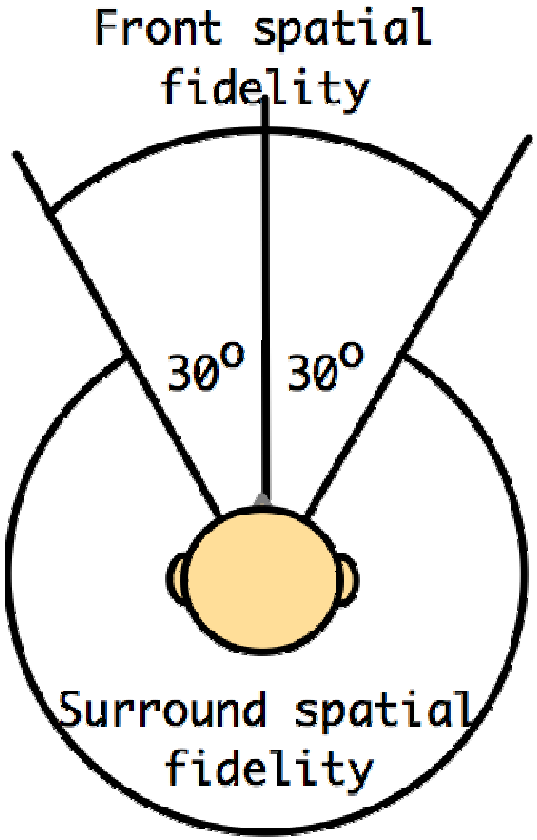
Spatial audio experience

- Basic Audio Quality (BAQ)
  - Difference between the original (reference) material and the processed material
  - Objective model: Rec. ITU-R BS.1387-1 method for objective measurements of perceived audio quality (known as PEAQ)



Spatial audio experience

- Front and Surround Spatial Fidelity



- Sounds in the front:
  - Localisation has better accuracy around 4°-10°
  - Important for visual cues (spatial synchronicity)
- Surrounding sounds:
  - Important for envelopment & immersiveness
- Spectral features
  - Centroid of spectral coherence (related to 'brightness')
  - Spectral rolloff
- Spatial features
  - Interaural cross correlation (IACC) at 0 degree (related to source width)
  - Maximum values of IACC at various degrees (related to source localisation)
  - Back-to-front energy ratio (related to 'spatial impression')
  - Lateral energy (related to 'envelopment')

Spatial audio experience

- Correlation of audio and video
  - Temporal synchronicity
    - Not related to 3D
    - Subjective test results exist that measure level of annoyance
    - Positive skew (audio ahead of video) and negative skew (video ahead of audio) – Not symmetric!
  - Spatial synchronicity
    - Related to 3D
    - Correlation of audio and video cues for positions of the media objects
      - Azimuth
      - Elevation
      - Depth
    - Requires testing as no models exist

Relevant subjective audio quality testing recommendations

- ITU Rec. BS.1284: General methods for the subjective assessment of sound quality
- ITU Rec. BS.1534: Method for the subjective assessment of intermediate quality levels of coding systems (known as MUSHRA)
- ITU Rec. BS.1116: Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems
- ITU Rec. BS.1286: Methods for the subjective assessment of audio systems with accompanying picture
- Effect of user interaction with the scene: Procedures for the evaluation of spatial audio coding systems (SAOC), issued by the MPEG.

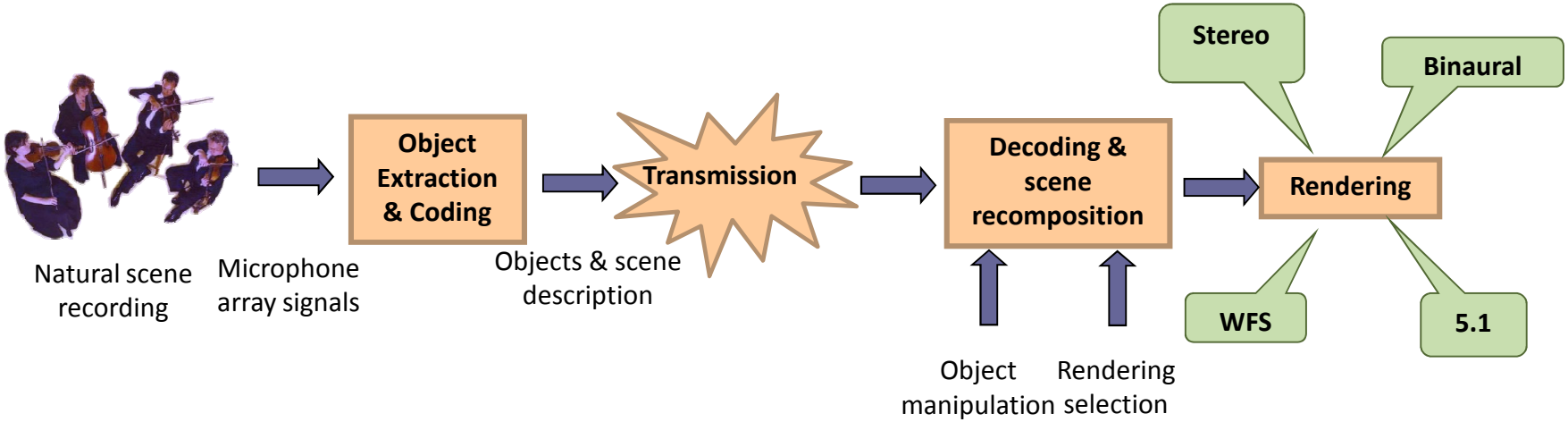
Mapping and validation

- Combining different attributes (Mapping submodel outputs to a single number)
  - Regression analysis or
  - An artificial neural network trained on the test results.

- Further subjective tests for validation

- Criterion for the audio QoE model accuracy (i.e., measures of success)
  - Correlation between subjective and objective results
  - Absolute error score (AES) as defined in ITU-R BS.1387-1 (takes into account the confidence intervals of the average values of the subjective tests.)

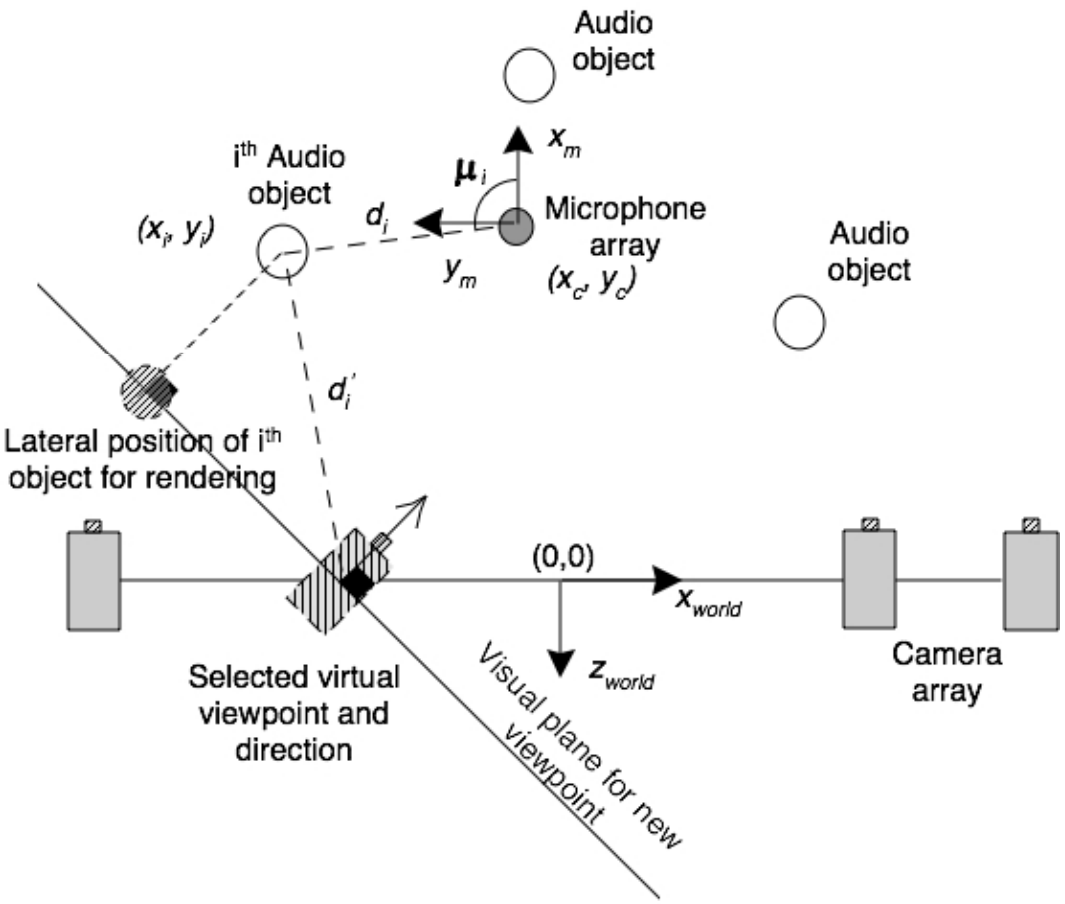
SAOC with Object Extraction



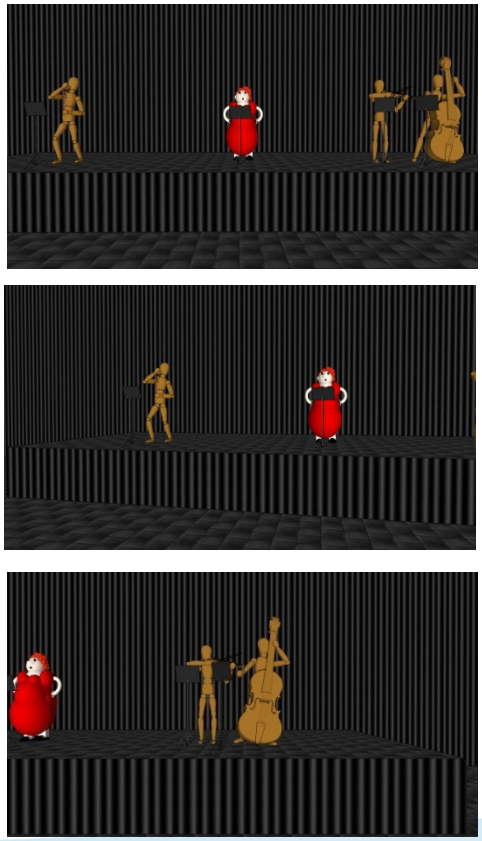
User Interaction/Manipulation

- ✓ Listen to each player separately
- ✓ Make an instrument quieter or louder
- ✓ Switch between different rendering systems
- ✓ Correct loudspeaker positions
- ✓ Compensate room acoustics

Spatial Synchronisation of Audio and Video Objects



Scene from different viewpoints



Processing for Spatial Synchronisation

Capturing side:

- Determine scene geometry (Metadata created during capturing)
- Decompose scene into its constituent audio objects (during capturing)
- Calibrate camera: calculate camera intrinsic matrix (Metadata created before actual capturing or during post-production)
- Encode audio objects with scene description and metadata, then transmit

Receiver side:

- Select rendering system, positions of loudspeakers, listening position, screen size and listening condition (Only once)
- Select viewpoint
- Calculate new audio object distances and correct sound level using inverse law
- Calculate the audio projection matrix (converts the coordinates of audio objects with respect to the microphone array into screen coordinates
  - Uses camera intrinsic matrix
  - Video projection matrix
  - Scaling factor that converts pixels into distances on the screen

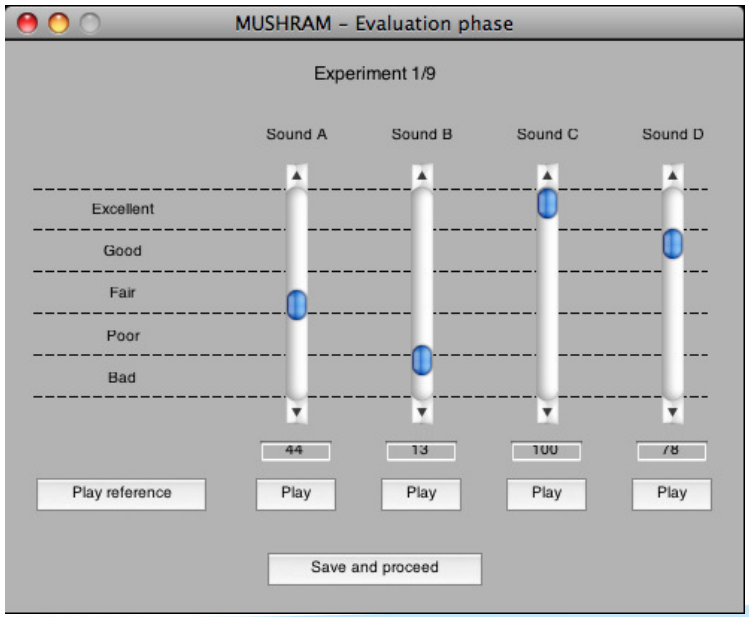
$$\mathbf{H_a} = \rho \mathbf{A}\ [\mathbf{R_v} | \mathbf{T_v} + \mathbf{T_a}]$$

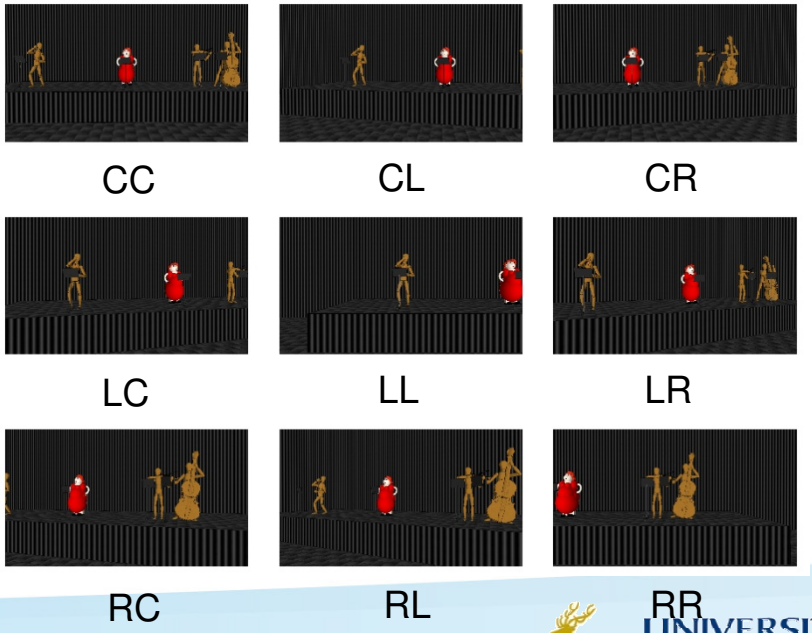
- Calculate rendering positions of audio objects that match the visual positions
- Process channels and play

Subjective Test

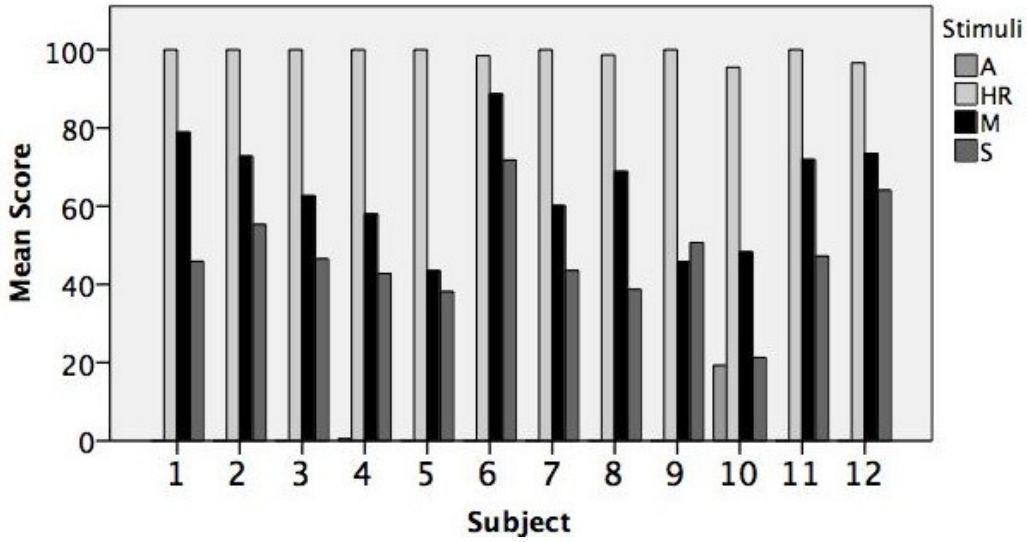
Aims:

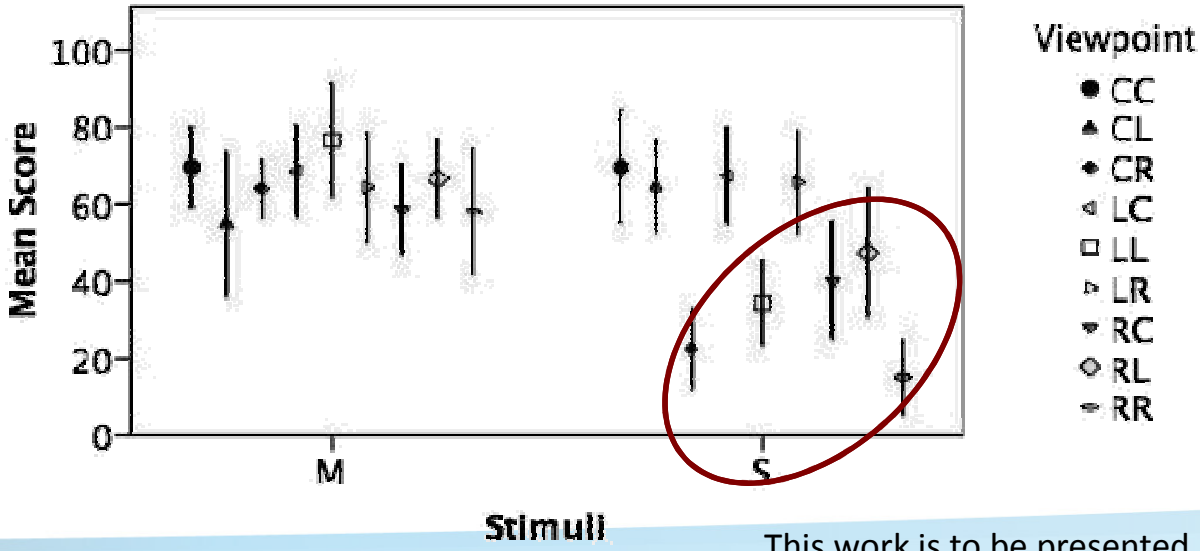
- Determine the perceptual effect of audiovisual object position mismatches
- Test the performance of the audio object extraction and rendering algorithm
  - MUSHRA test (ITU-R BS.1534) with accompanying images (ITU-R BS.1286).
  - 12 participants
  - 9 viewpoints
  - 4 different aural stimuli for each viewpoint: Reference (R), Anchor (A), Method (M), Standard Stereo (S)
  - An aria containing soprano, clarinet, contrabass and violin
  - Attribute for comparison: Correlation of source positions derived from visual and audible cues (No depth this time!)





Test Results

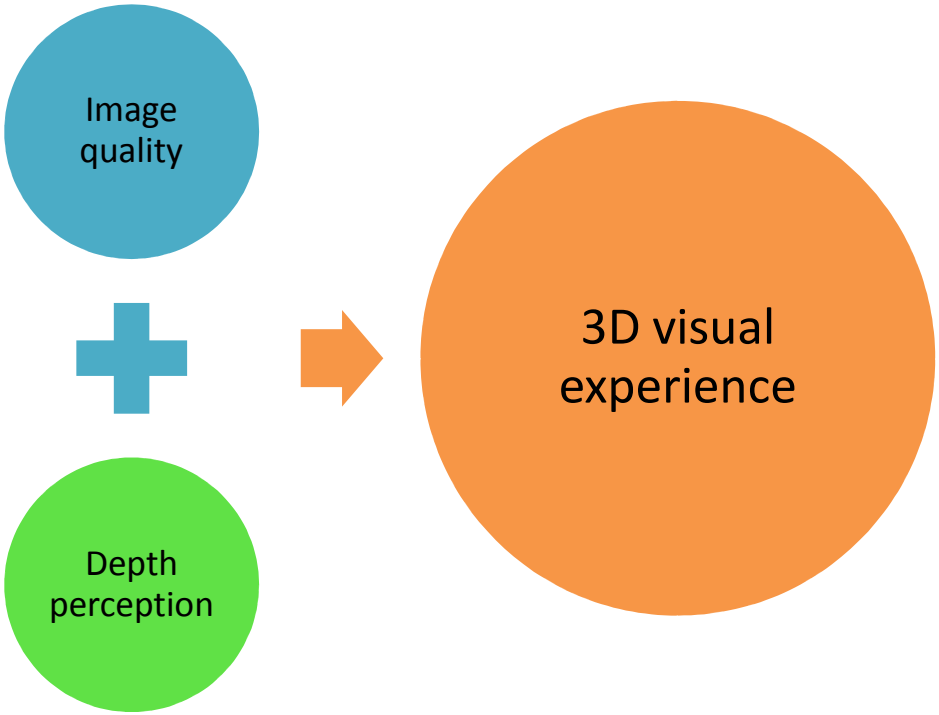




- Mismatches between audio and video objects are highly perceptible.
- The sounds prepared with the proposed method (M), performs much better than standard stereo recordings (S).
- The discrepancies are more pronounced for the scenes (CR, LL, RC, RL, RR) where significant changes occur in the video object positions as compared to the centre viewpoint (CC).

3D visual experience

- KPIs
  - Image quality
    - Quality of the 2D view
  - Depth perception
    - Quality of the perceived depth

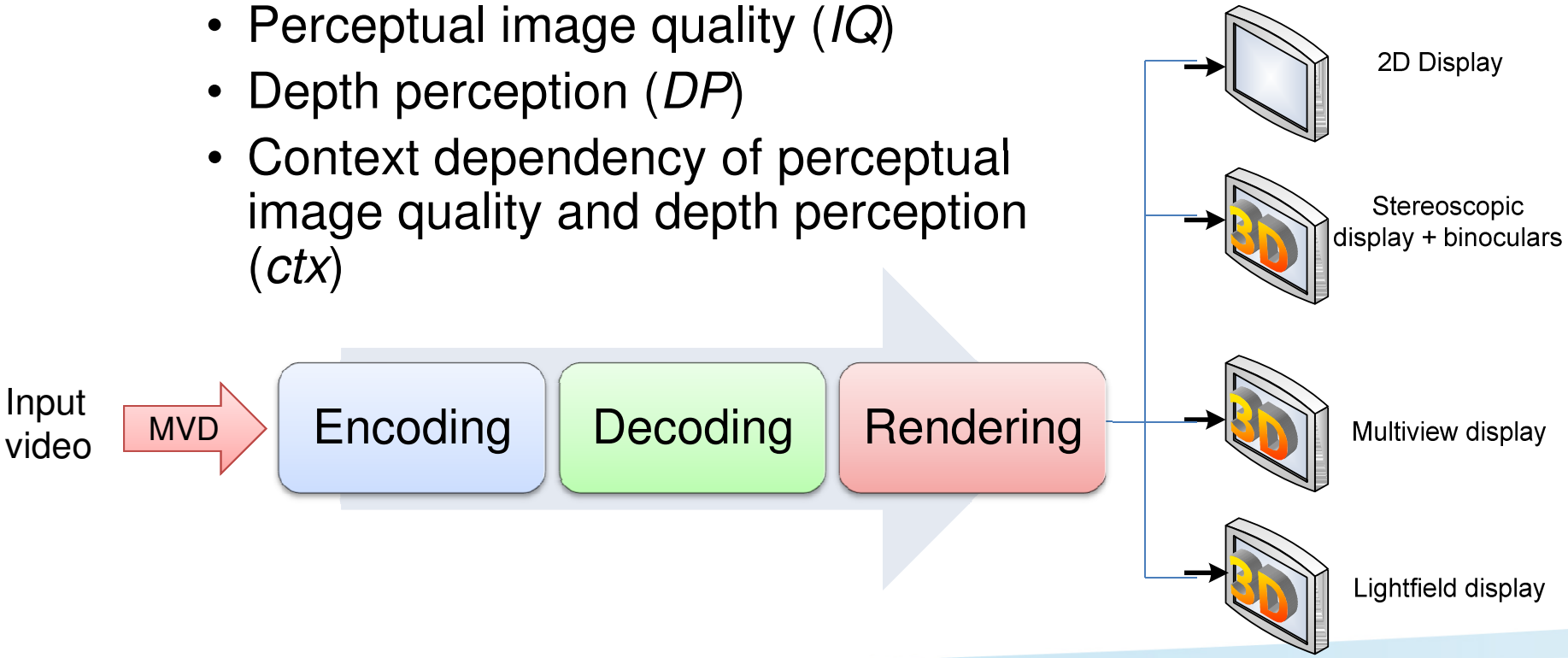


- The model

$$QoE_{Video} = f(IQ, DP, ctx)$$

3D visual experience

- Methodology
  - Subjective experiments will be performed to assess
    - Perceptual image quality (*IQ*)
    - Depth perception (*DP*)
    - Context dependency of perceptual image quality and depth perception (*ctx*)

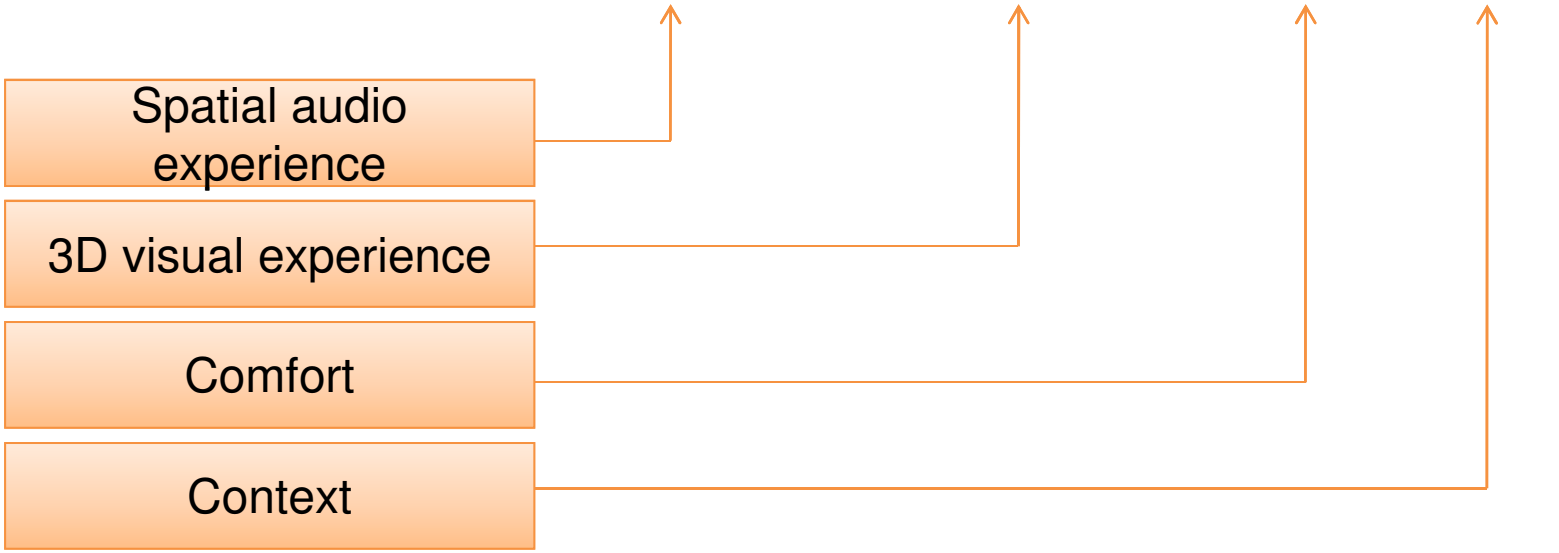


Comfort

- Audio and visual experience is comfort dependant
- Comfort is independent of both the audio and visual quality
- Comfort associated with different audio and video devices will be incorporated in the overall QoE model
- Comfort factor will be assessed through a questioners filled by volunteers attended for subjective experiments

Proposed overall audiovisual QoE model

$$QoE = f\left(QoE_{Audio}, QoE_{Video}, cmf, cxt\right)$$



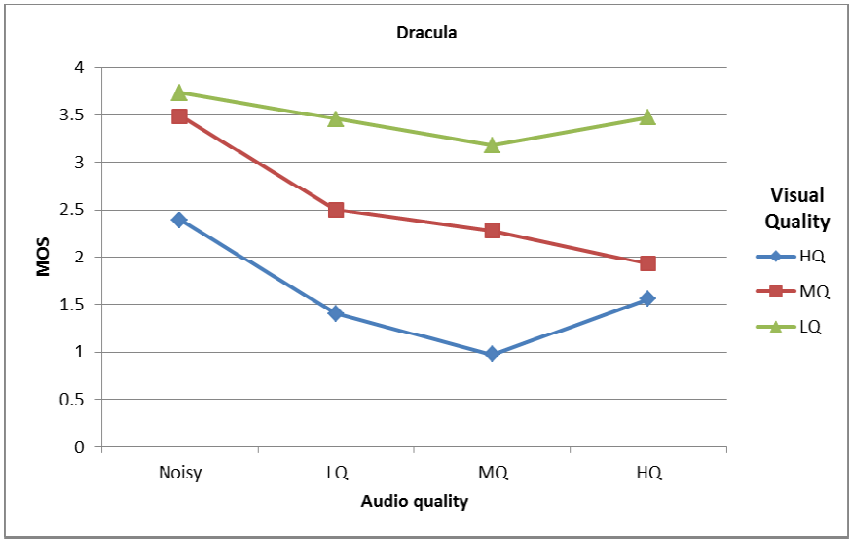
QOE MODEL FOR AUDIOVISUAL CONTENTS

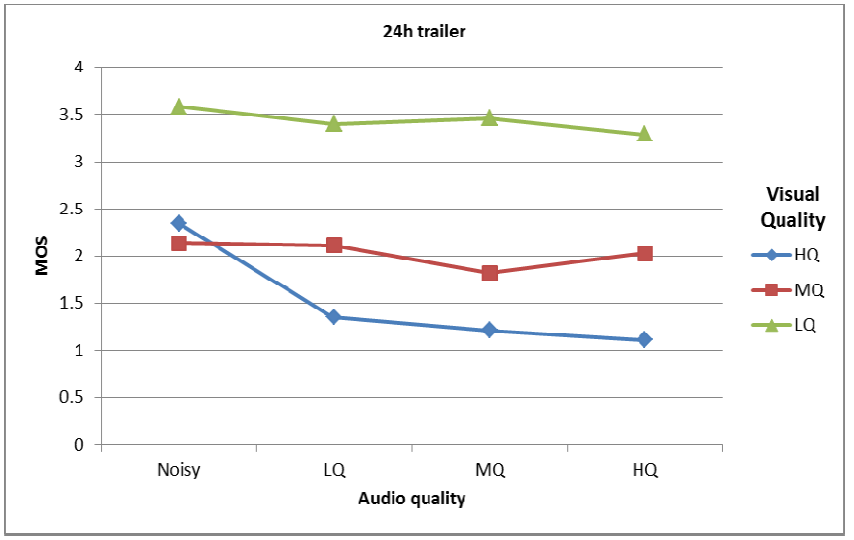
Subjective experimentation setup

- Objective
  - To assess the relative importance of the audio and visual components on overall 3D audiovisual experience
- Three stereo video sequences with stereo audio are used for this experiment
  - Video is encoded using JMVC codec
  - Audio is encoded with AAC codec
- The sequences were displayed on JVC passive stereoscopic display

Experimental results

- Dracula test sequence
  - A stereoscopic animated movie clip





- Quality of visual content has more impact on audiovisual experience
- Future work
  - Determining the weighting factors for audio and visual QoE for predicting overall audiovisual QoE

9. Summary

- Quality measurements techniques
  - FR, RR, NR
- 2D Video quality measurements
  - PSNR, SSIM, VQM
- 3D Video quality measurements
- Depth Perception Modelling
- Quality of Experience

Acknowledgements

- MUSCADE (www.muscade.eu)
- Varuna De Silva, Hemantha Kodikara Arachchi(I-Lab University of Surrey)

MUSCADE
MUltimedia SCAlable 3D for Europe

THANK YOU!