Test plan for VQEG GroTruQoE3D1 database

Jing Li
Marcus Barkowsky
University of Nantes, France

Scope and goals

- Establish a ground truth database for Quality of Experience in 3DTV (GroTruQoE3D) measurement methodologies

- Standardization of subjective assessment methodology for different degradations in 3DTV

Outline

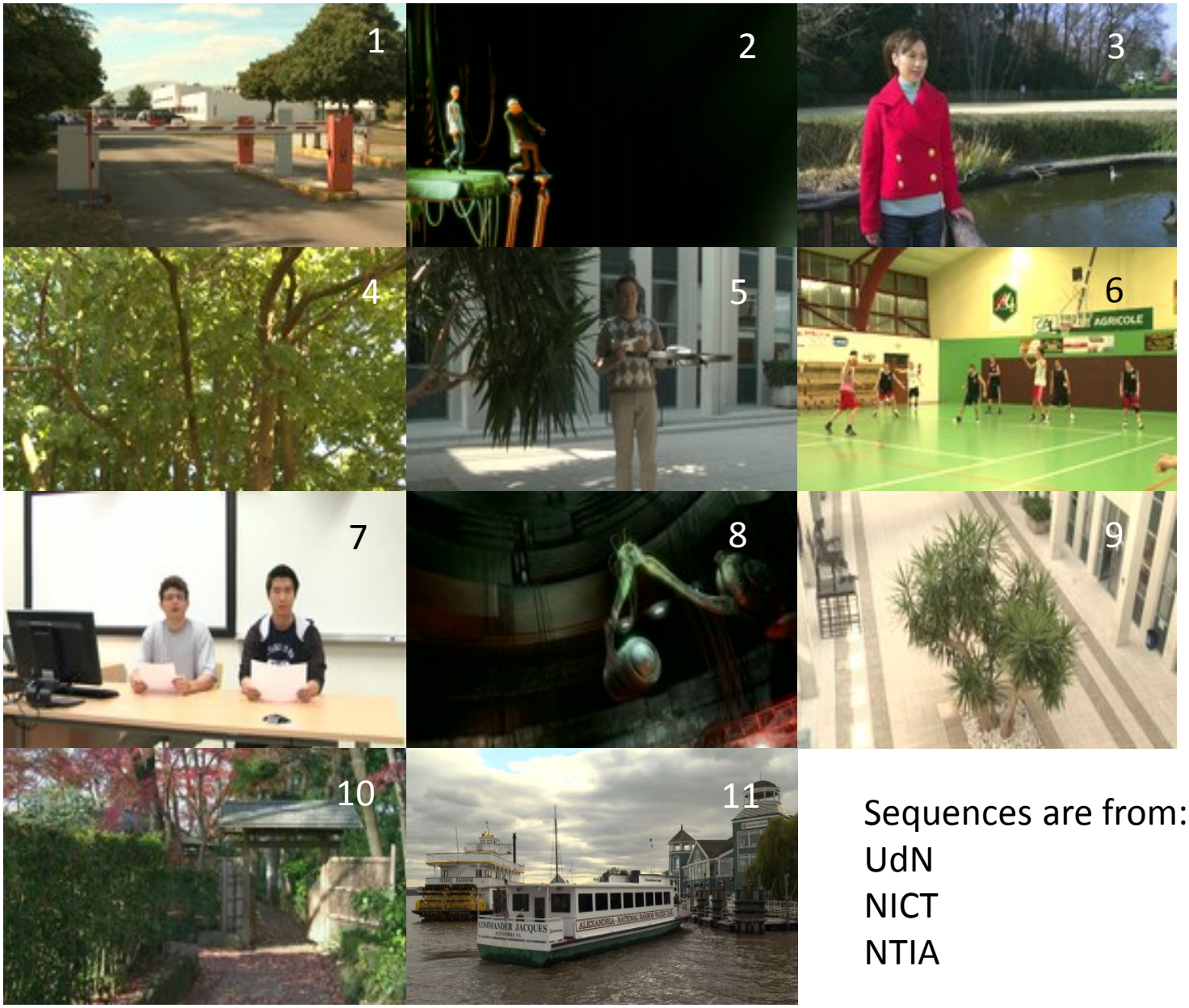
- Introduction of video sequences
- Why Pair Comparison test method is selected
- How to boost pair comparison
- How the experiment being conducted collaboratively
- What is the requirement for test setup
- How to analyze the data

Video sequences-SRC

- 10 (or 11) SRC
- Cover a wide range of different content features
  - coding complexity
  - motion
  - brightness
  - 3D effect
  - maximum disparity range

Comfortable viewing

- YUV422, Full HD, 25fps, 16seconds (except “Umbrella” 13 seconds)



Sequences are from:
UdN
NICT
NTIA

Video sequences-HRC

18 HRC: Selected based on three scales: “image quality”, “visual comfort” and “depth quantity” *

HRC	Remarks	Encoding			Packet loss	Decoding
		Encoder	Bitrate/QP	GOP		
0	(reference 3D)					
1	(reference 2D)					
2	spatial resolutiuon reduction by 4 with lanczos 3 filter					
3	fps reduction by 3					
4	Brightness at 80% (Only one view is changed)					
5	gamma at 0.5 (Only one view is changed)					
6	horizontal disparty offset -30 pixel					
7	horizontal disparty offset 30 pixel					
8	vertical disparity offset -20 pixel					
9	graphical distortion with stirmark (Only one view is changed)					
10	2D to 30 Using geometric deformation					
11		JM 18.2	~/32	IBBP64		JM18.2
12		JM 18.2	~/44	IBBP64		JM18.2
13	asymetric view	JM 18.2	~/32 ~/44	IBBP64		JM18.2
14	Edge enhancement at 40%	JM 18.2	~/32	IBBP64		JM18.2
15	error concealment (Only one view is changed)	JM 18.2	~/32	IBBP64	gilbert strong	JM18.2
16	2D video (left view)	JM 18.2	~/44	IBBP64		JM18.2
17	JPEG2000 Encoding	Jpeg2000 kakadu 6	23Mbps/~			Jpeg2000 kakadu

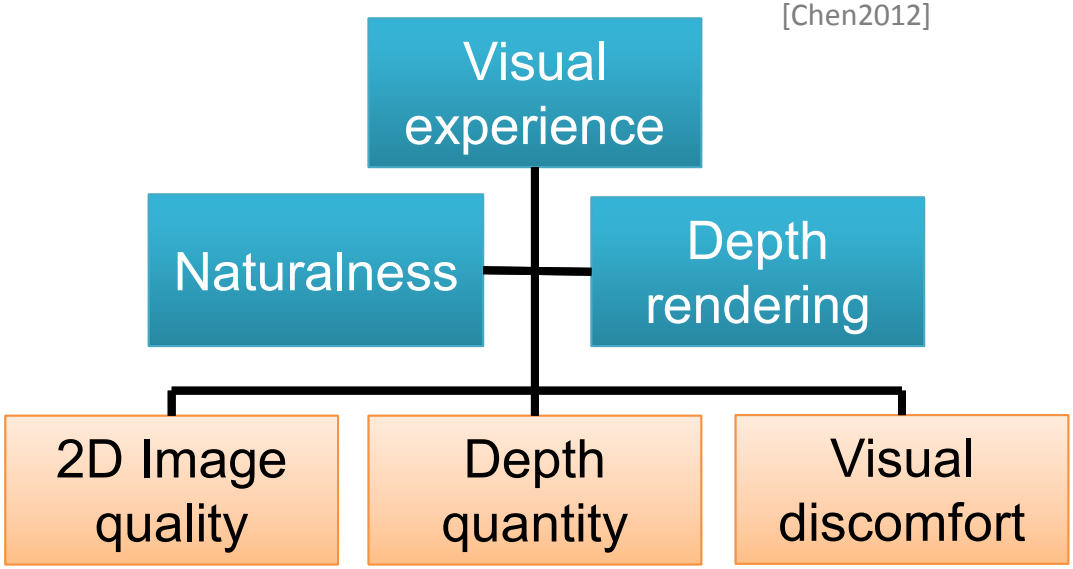
*The selection was performed based on votes from experts from the Mid Sweden University (MIUN)
Details see https://docs.google.com/spreadsheet/ccc?key=0ArzgrjHcemZYdHk3ZGhOZ0w0VzFHblA1M0ROLTJCQ1E#gid=2

Outline

- Introduction of video sequences
- Why Pair Comparison test method is selected
- How to boost pair comparison
- How the experiment being conducted collaboratively
- What is the requirement for test setup
- How to analyze the data

Quality of Experience (QoE) in 3DTV

[Chen2012]



Chen, W., Fournier, J., Barkowsky, M., & Le Callet, P. (2012). Exploration of Quality of Experience of Stereoscopic Images: Binocular Depth. VPQM.

Subjective assessment methodology for QoE in 3DTV

- 2D image quality assessment [P.910][BT.500]
  - **ACR (Absolute Category Rating)**
  - DSCQS (Double-Stimulus Continuous Quality Scale)
  - SSCQE (Single Stimulus Continuous Quality Evaluation)

ACR

5 excellent

4 good

3 fair

2 poor

1 bad

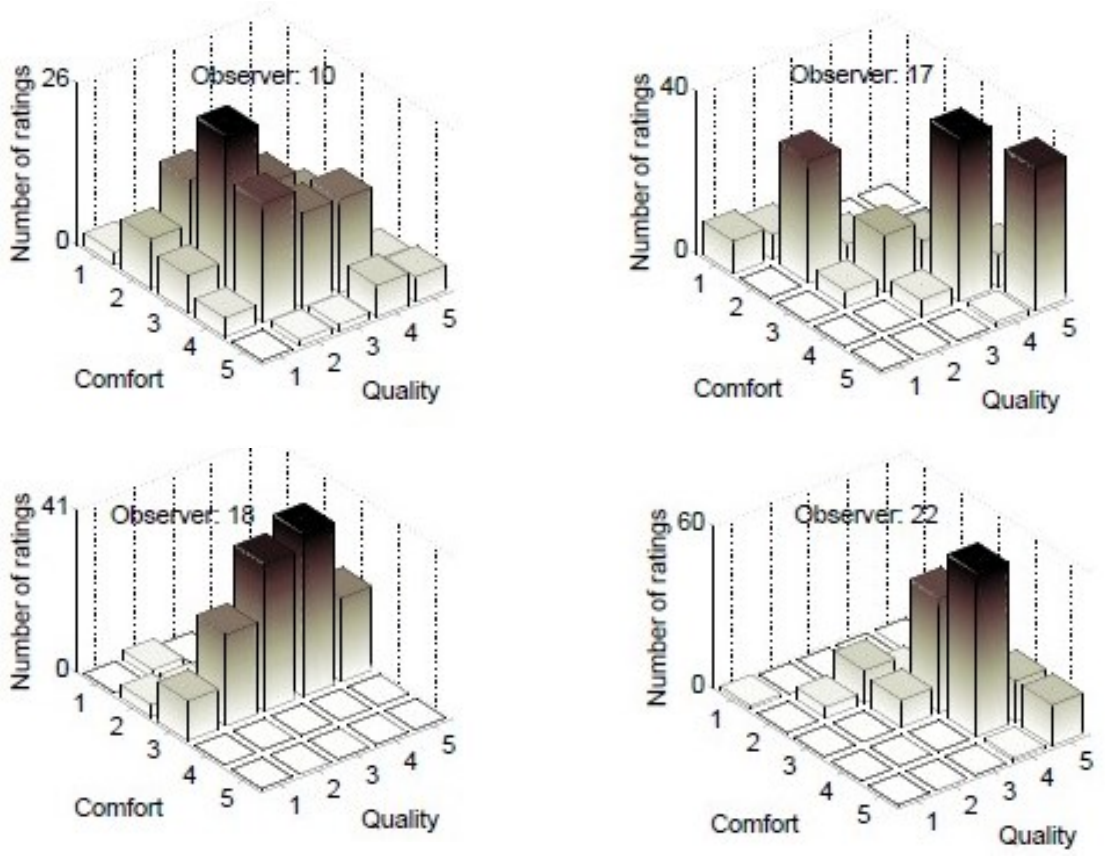
Adapted to 3D

On each dimension of 3D QoE [BT.2021]

- Image quality
- Visual comfort
- Depth quality

Example: scale interpretation & observer variability

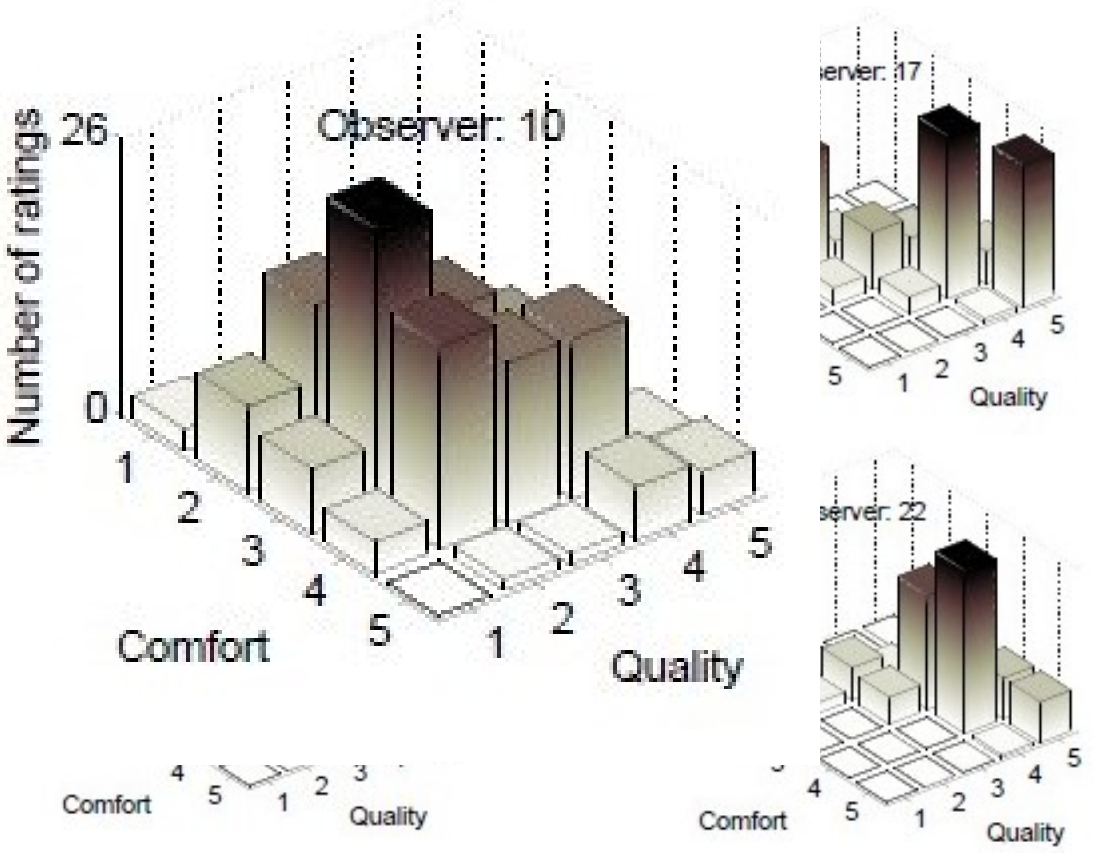
- A Co-joint ACR experiment for visual comfort and image quality in 3DTV [Engelke2011]



Engelke, Ulrich, Yohann Pitrey, and Patrick Le Callet. "Towards an inter-observer analysis framework for multimedia quality assessment." *Third International Workshop on Quality of Multimedia Experience (QoMEX),* IEEE, 2011.

Example: scale interpretation & observer variability

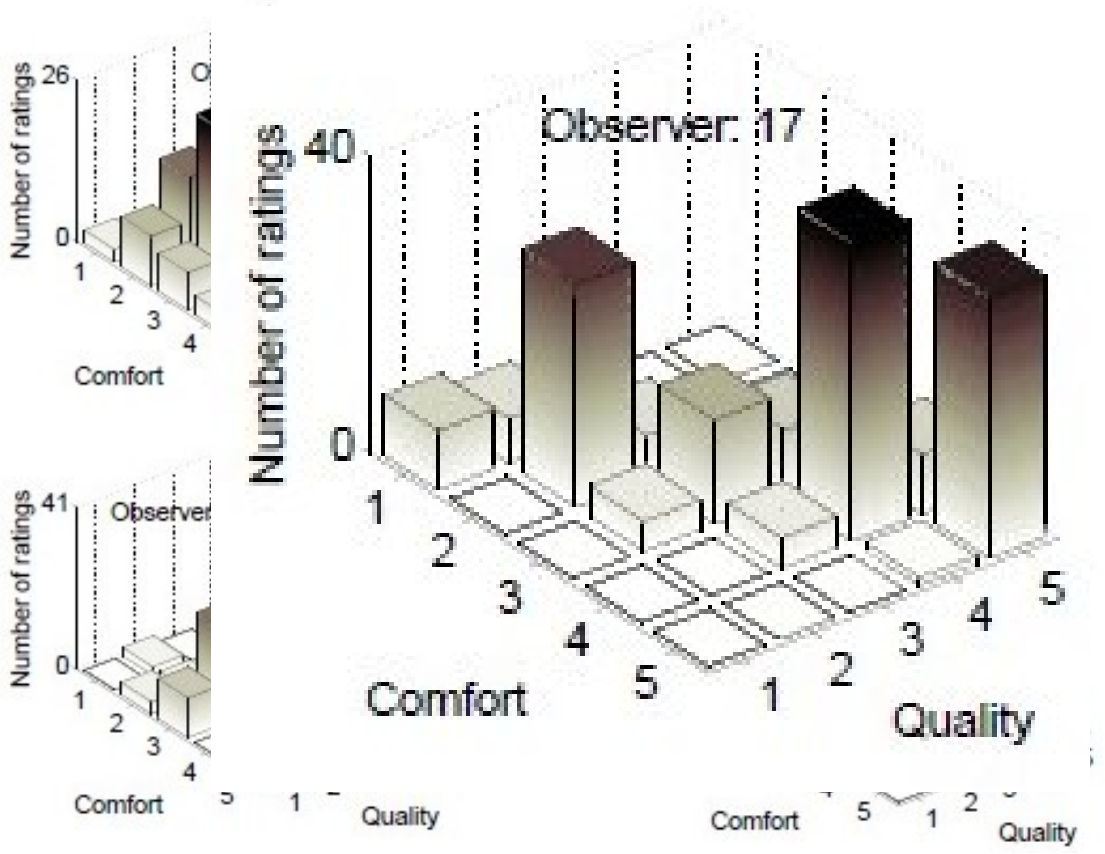
- A Co-joint ACR experiment for visual comfort and image quality in 3DTV [Engelke2011]



Engelke, Ulrich, Yohann Pitrey, and Patrick Le Callet. "Towards an inter-observer analysis framework for multimedia quality assessment." *Third International Workshop on Quality of Multimedia Experience (QoMEX),* IEEE, 2011.

Example: scale interpretation & observer variability

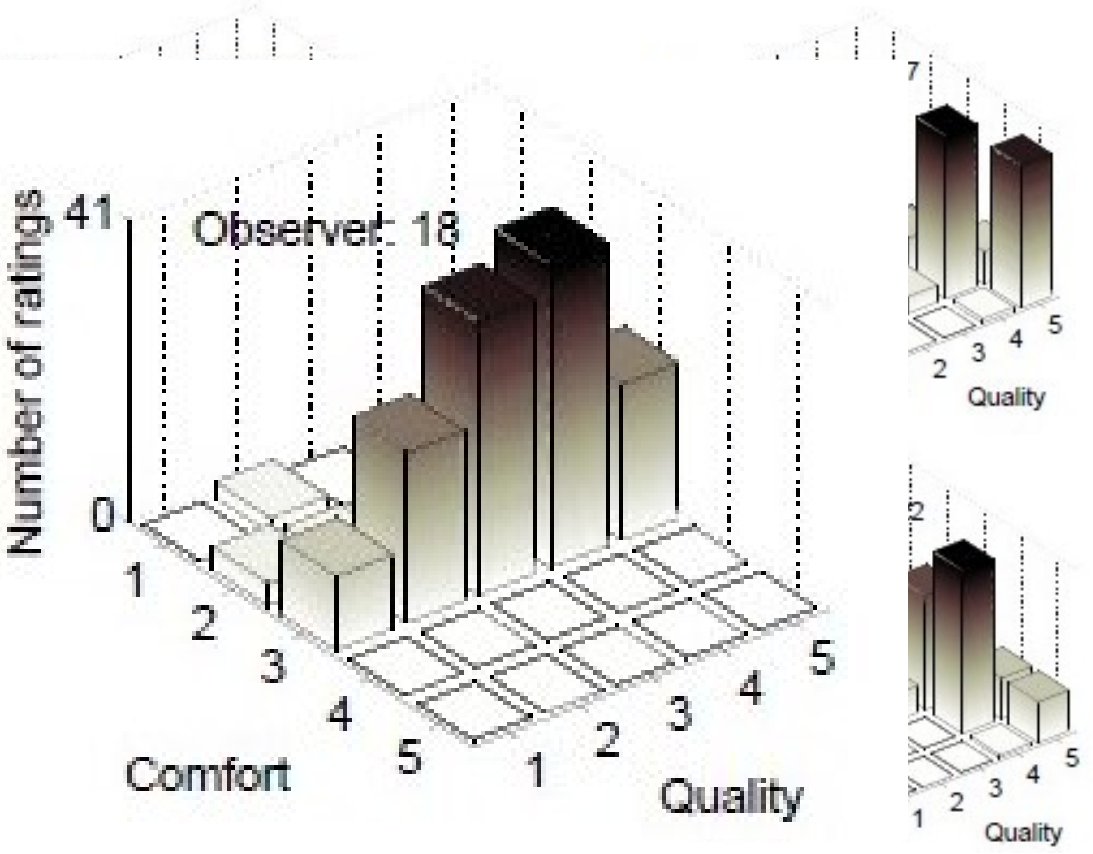
- A Co-joint ACR experiment for visual comfort and image quality in 3DTV [Engelke2011]



Engelke, Ulrich, Yohann Pitrey, and Patrick Le Callet. "Towards an inter-observer analysis framework for multimedia quality assessment." *Third International Workshop on Quality of Multimedia Experience (QoMEX),* IEEE, 2011.

Example: scale interpretation & observer variability

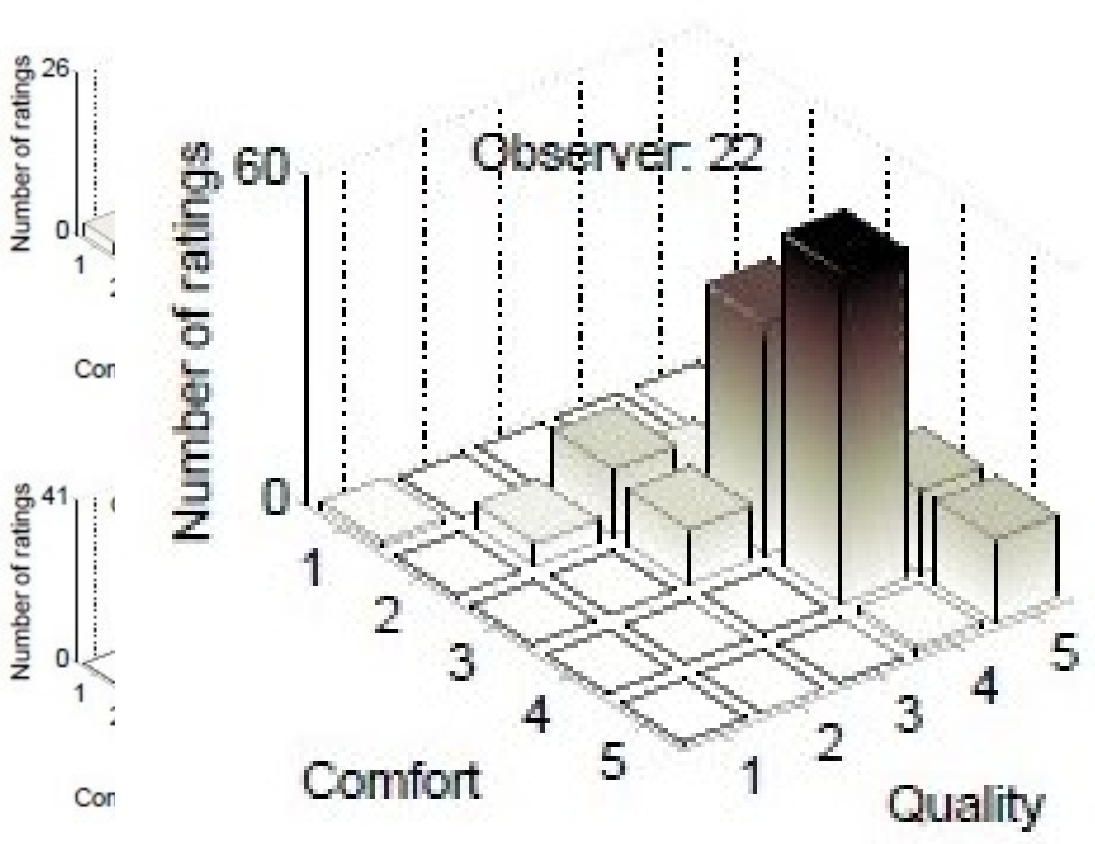
- A Co-joint ACR experiment for visual comfort and image quality in 3DTV [Engelke2011]



Engelke, Ulrich, Yohann Pitrey, and Patrick Le Callet. "Towards an inter-observer analysis framework for multimedia quality assessment." *Third International Workshop on Quality of Multimedia Experience (QoMEX),* IEEE, 2011.

Example: scale interpretation & observer variability

- A Co-joint ACR experiment for visual comfort and image quality in 3DTV [Engelke2011]



Engelke, Ulrich, Yohann Pitrey, and Patrick Le Callet. "Towards an inter-observer analysis framework for multimedia quality assessment." *Third International Workshop on Quality of Multimedia Experience (QoMEX),* IEEE, 2011.

- Subjects are not always capable of expressing their perceptions or impression by means of an exact numerical value.

ACR

Observers

- not used to 3D

→ difficult to link the perception with experience

- Language differences

  excellent, bon, … in french

Judgment might be unreliable

A good alternative…

ACR

5	excellent
4	good
3	fair
2	poor
1	bad

alternative

Pair Comparison

A B

- Which one do you prefer? → QoE or PoE

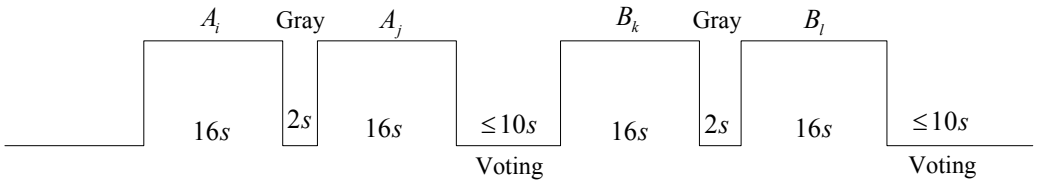
PoE: Preference of Experience

- Avoid the scale interpretation issues
- Easy to understand for subjects
- Easy to implement for subjects

Pair Comparison

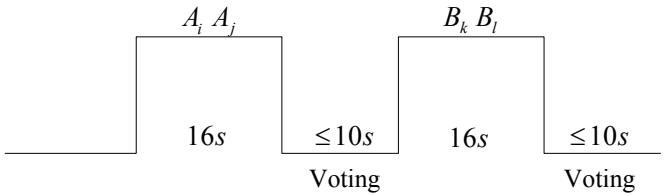
- Compare different degradations of same content

- Time-sequential



A_i A_j : Stimulus with content A under test condition i and j, respectively.
B_k B_l : Stimulus with content B under test condition k and l, respectively.

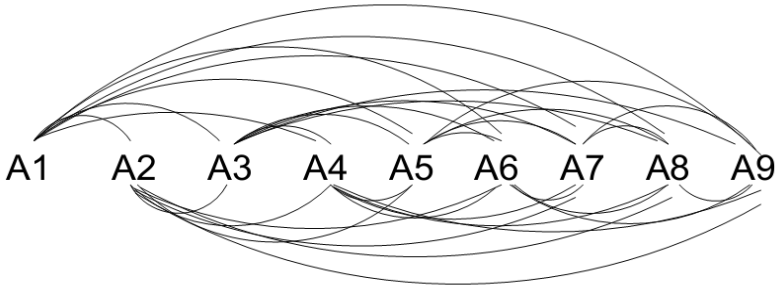
- Time-parallel

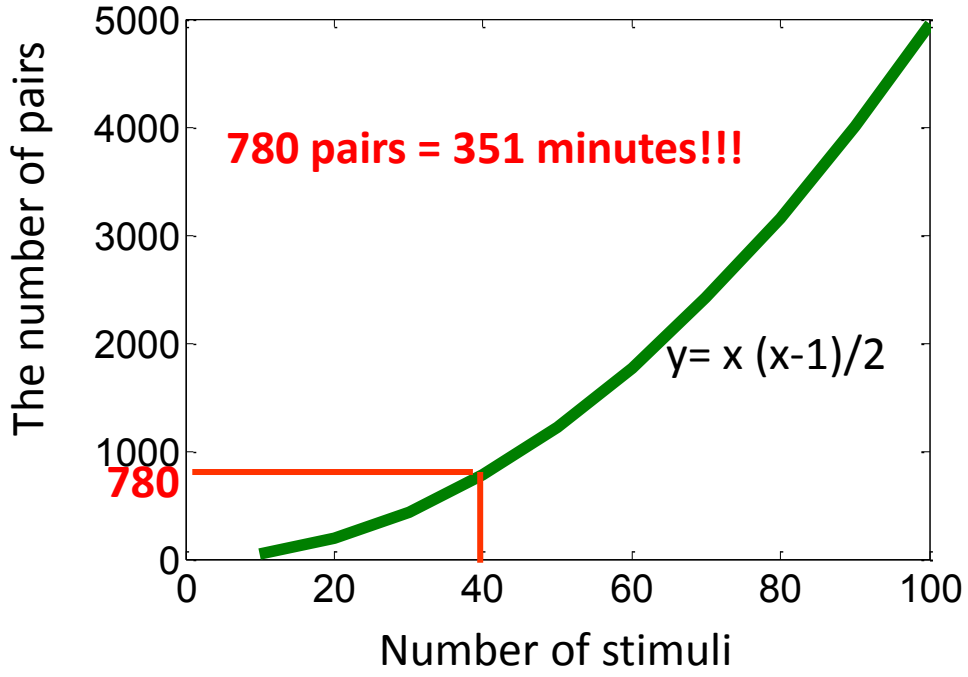


A_i A_j : Stimulus with content A under test condition i and j, respectively.
B_k B_l : Stimulus with content B under test condition k and l, respectively.

Limitation of Pair Comparison in real application

All pairs have to be compared…





For a ACR test:
40 stimuli *15s
= 10 minutes !

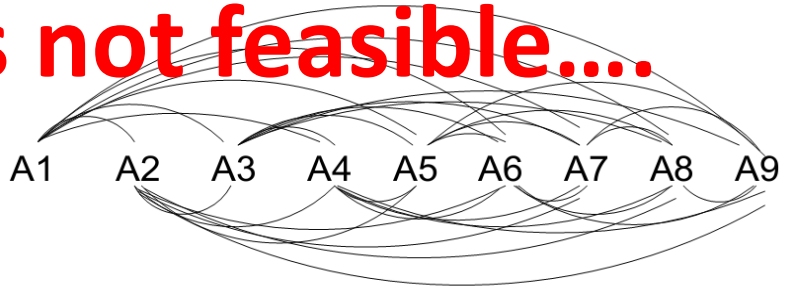
Using Time-sequential presentation, for example:

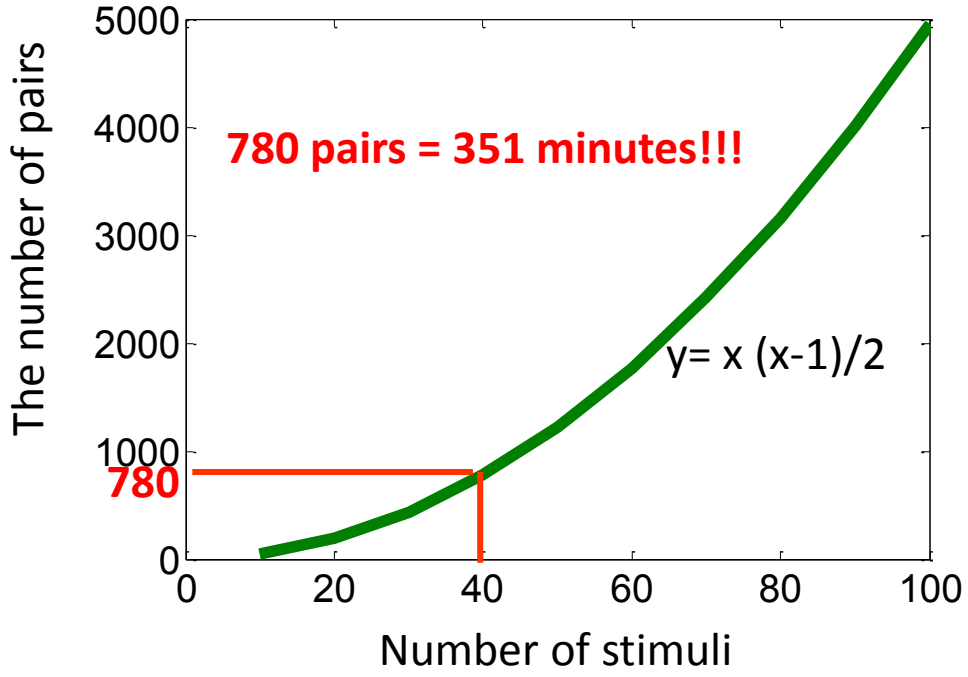
For each comparison: A1 (10s)+gray(2s)+A2 (10s)+voting (5s) = 27s

Limitation of Pair Comparison in real application

Pair Comparison is not feasible….

All pairs have to be compared…





For a ACR test:
40 stimuli *15s
= 10 minutes !

Using Time-sequential presentation, for example:

For each comparison: A1 (10s)+gray(2s)+A2 (10s)+voting (5s) = 27s

Outline

- Introduction of video sequences
- Why Pair Comparison test method is selected
- How to boost pair comparison
- How the experiment being conducted collaboratively
- What is the requirement for test setup
- How to analyze the data

To reduce the number of comparisons…

✓ Select a subset of the whole pairs for comparison

- **Efficient:**

  →the selected pairs should provide more information on the final scale values than other pairs.

- **Balanced:**

  The occurrence frequency of each stimulus is equal.

  → to avoid any bias effects from presentation frequency of a particular stimulus

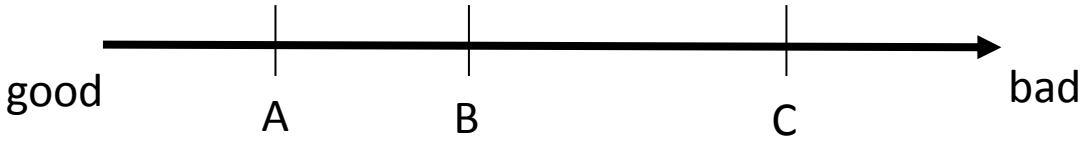
- **Robust:**

  The selection of the pairs would be more robust to observation errors that often happen in a subjective test [Li2012].

Jing Li, Marcus Barkowsky, Patrick Le Callet, "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment", ICIP, 2012.

Analysis

If we can only compare two pairs to determine the distances between A,B,C, which two pairs should we select?



Example:

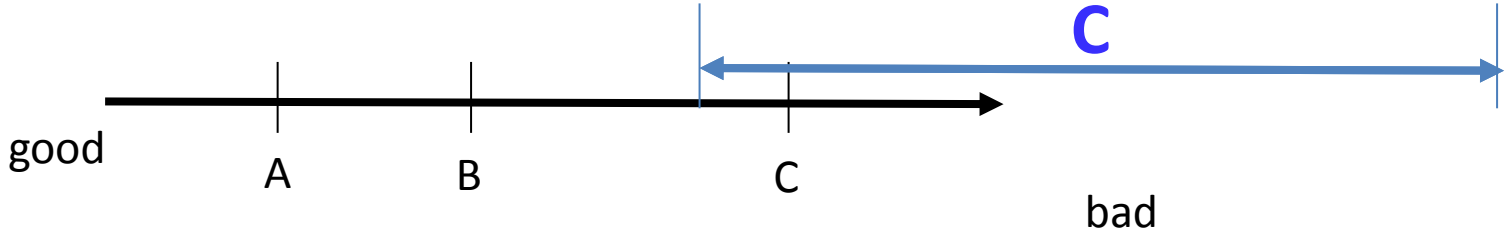
Choose AB → closer, 15 out of 20 observers select A

Choose AC → too far away

→ With a small number of observations, almost all observers choose A

→ $P_{AC} = 1$

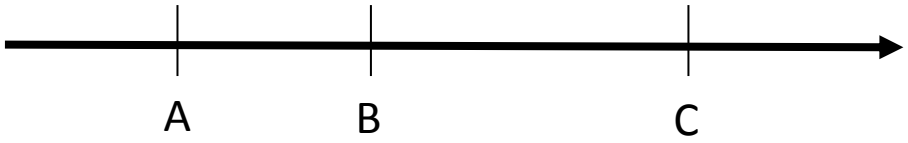
→ cannot define the distance between A and C



Jing Li, Marcus Barkowsky, Patrick Le Callet, (2012). "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment", ICIP.

Analysis

If we can only compare two pairs to determine the distances between A,B,C, which two pairs should we select?



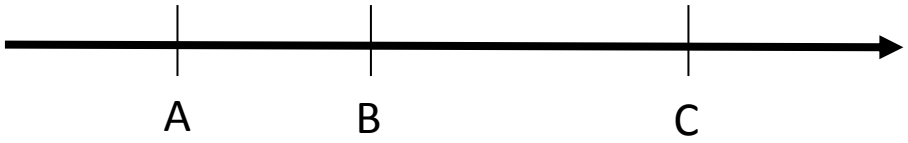
- Statistical analysis showed that comparison on closer pairs would
  - generate more precise results than distinct pairs;
  - be more robust to observation errors than distinct pairs [LilCIP2012]

➢ To design an efficient and robust design, comparisons should be concentrated on closer pairs! → AB, BC rather than OTHERS

Jing Li, Marcus Barkowsky, Patrick Le Callet, (2012). "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment", ICIP.

Analysis

If we can only compare two pairs to determine the distances between A,B,C, which two pairs should we select?



- Statistical analysis showed that comparison on closer pairs would
  - generate more precise results than distinct pairs;
  - be more robust to observation errors than distinct pairs [LilCIP2012]

➢ To design an efficient and robust design, comparisons should be concentrated on closer pairs! → AB, BC rather than OTHERS

How about balance?

Jing Li, Marcus Barkowsky, Patrick Le Callet, (2012). "Analysis and improvement of a paired comparison method in the application of 3DTV subjective experiment", ICIP.

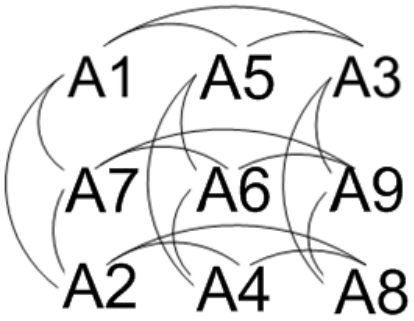
Optimized Rectangular Design

Supposing the rank ordering of the video sequences is available:

A1 A5 A3 A9 A8 A4 A2 A7 A6

Arrangement of matrix

A1	A5	A3
A7	A6	A9
A2	A4	A8



1. Closer stimuli are arranged close to each other → efficient & robust
2. Only compare stimuli which are in the same column or row
   → reduce the number of comparison
   → occurrence of each stimulus is balanced (e.g., 4 times for each stimulus)

Jing Li, Marcus Barkowsky, Patrick Le Callet, "Boosting Paired Comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs", SPIE SD&A, 2013.

Outline

- Introduction of video sequences
- Why Pair Comparison test method is selected
- How to boost pair comparison
- How the experiment being conducted collaboratively
- What is the requirement for test setup
- How to analyze the data

Situation

- 10 SRCs
- 18 HRCs →using 3×6 ORD method→63 pairs
- Comparison only conducted within SRC→ 63*10=630 pairs

For one pair:

- Time-sequential: 16(A)+2(gray)+16(B)+5(vote)=39 seconds
- Time-parallel: 16 (AB)+5(vote)=21 seconds

For 630 pairs

- Time-sequential: 630*39 = 409.5 minutes per observer
- Time-parallel: 630*21 = 220.5 minutes per observer

We need collaboration!!

Problem & Solution

- Different labs
  - different display technology
  - different observers
  - different screen size
  - different presentation method
  - …
- Before collecting the data from different labs

→ We need a common set to evaluate the possibility of collection

→ A precise test plan is provided

try to minimize the unnecessary effects

Design of Common Set

1. The number of pairs of the common set should be large enough for analysis

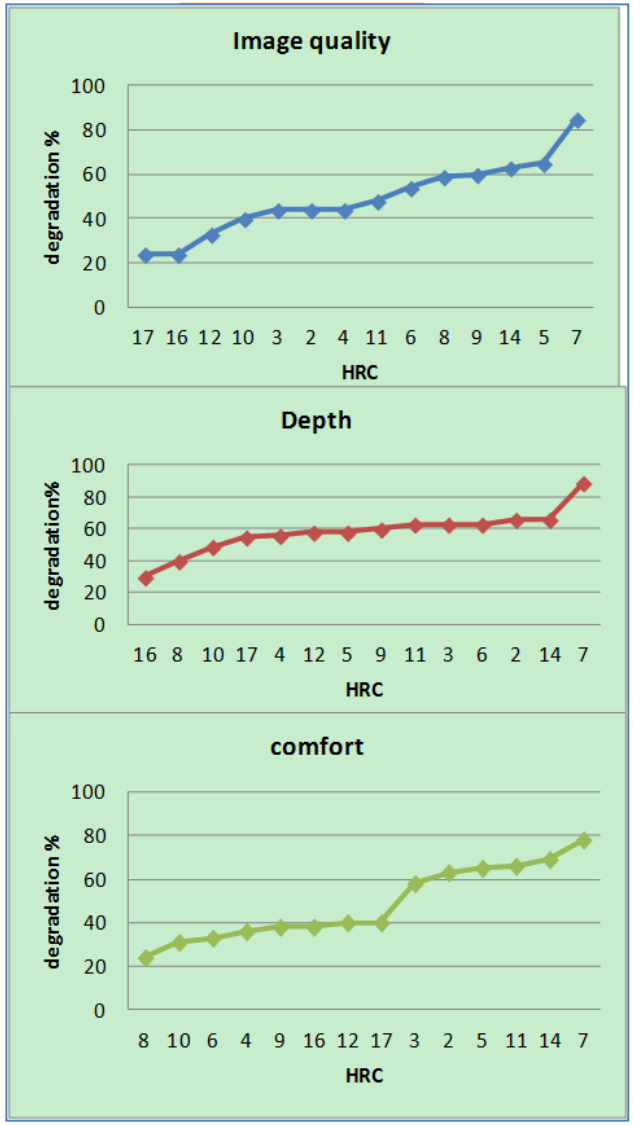
2. The pairs should be included in the whole test pairs by ORD matrix

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	0

NO

3. As the total number of comparisons for each observer is very limited (within 50mins), the common set should be not too large…

Design of Common Set



For all HRCs

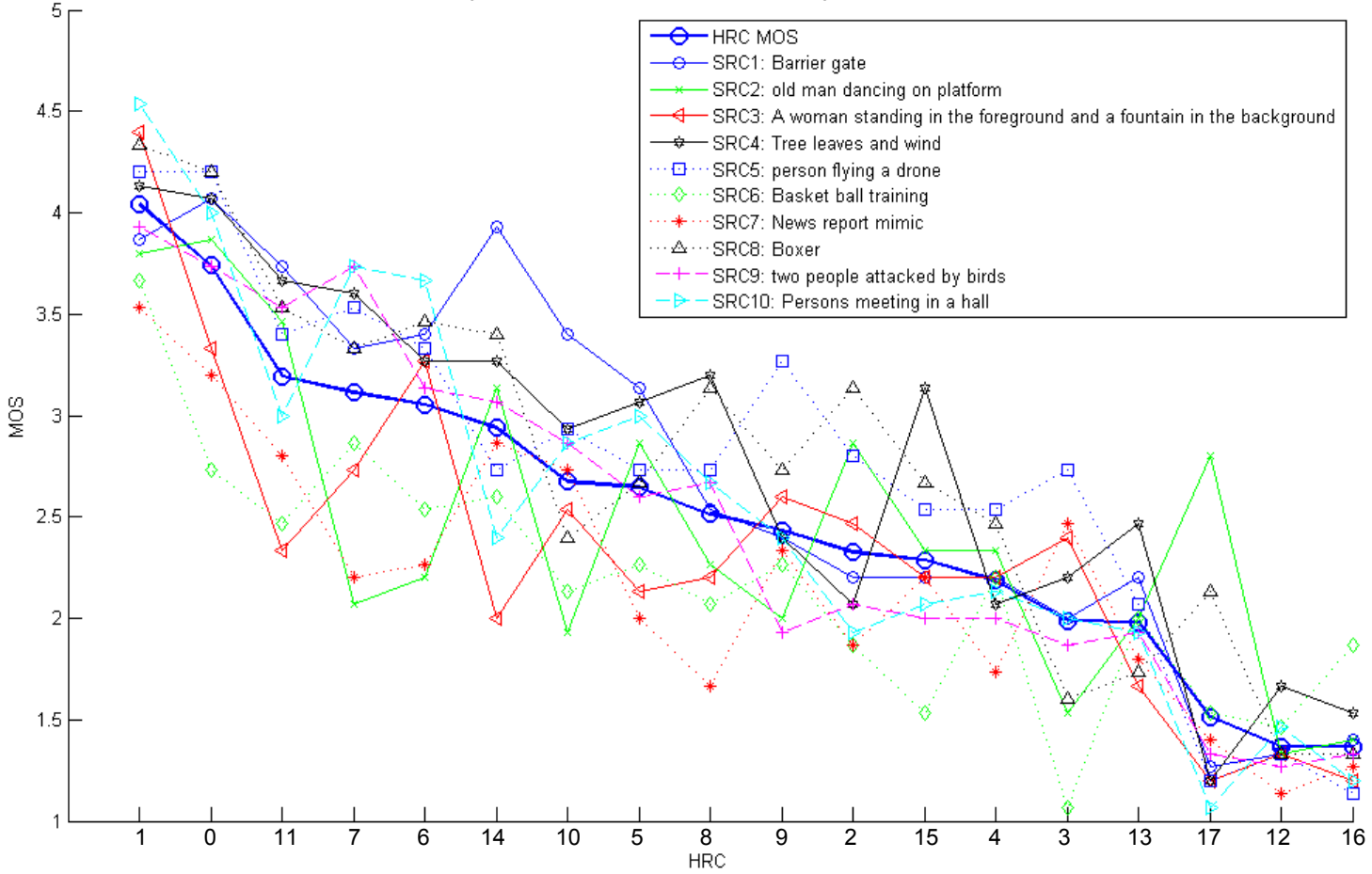
Degradations have been evaluated by 4 experts from the Mid Sweden University (MIUN) on the three scale
“picture quality”,
“depth quantity”,
“visual discomfort”

- Common set should cover the whole range of each of the three dimensions

Design of Common Set

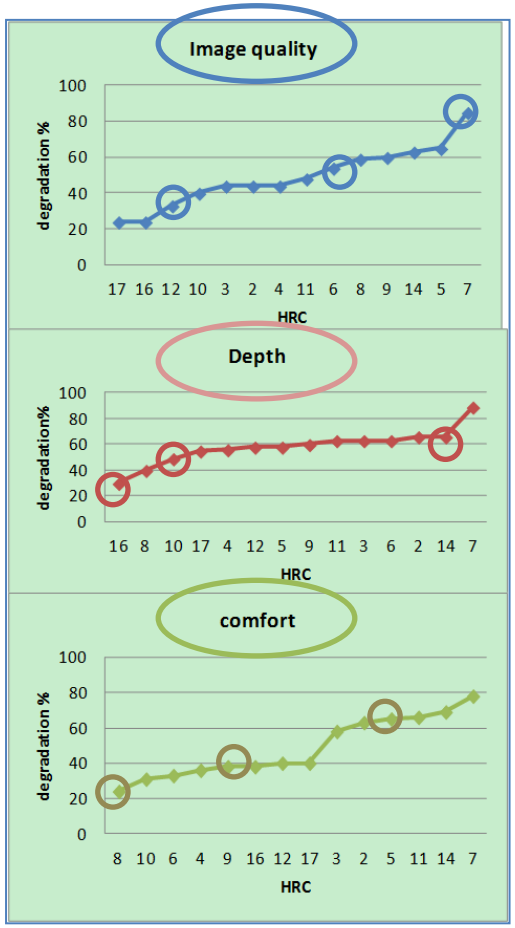
Common set should be part of the whole test pairs

Our ACR test results (15 naive observers)

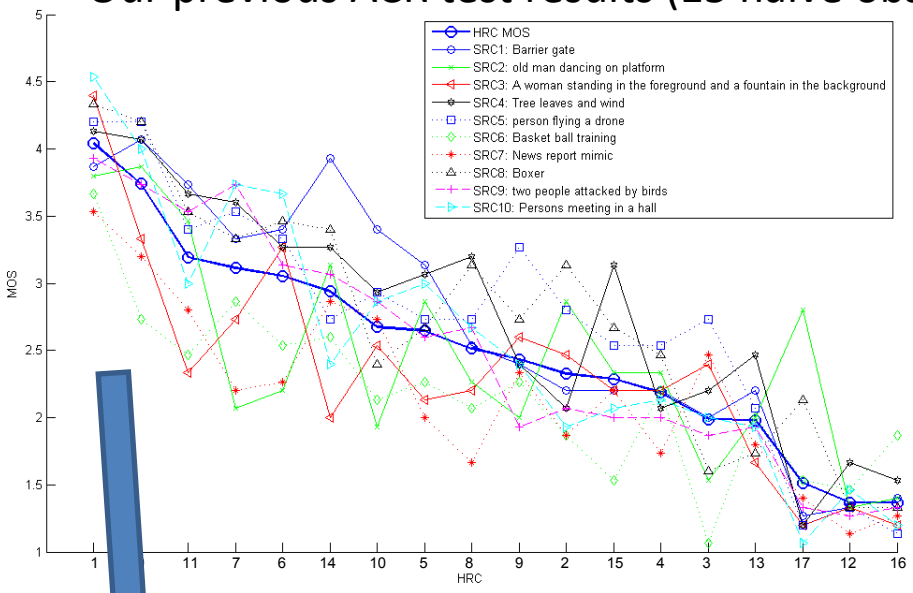


Closer pairs should be arranged in the same column or row of the ORD matrix

Design of Common Set



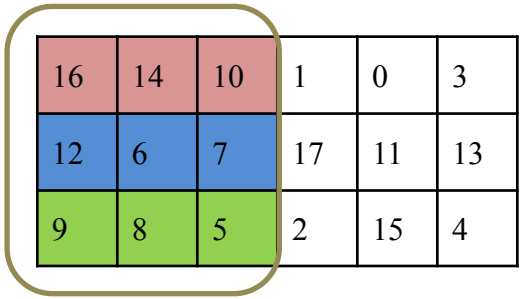
Our previous ACR test results (15 naive observers)



16	14	10	1	0	3
12	6	7	17	11	13
9	8	5	2	15	4

- Common set includes representative of three dimensions
- Common set should be a part of the 3 ×6 matrix for ORD method
- Closer HRC pairs are arranged in the same column or row

Design of Common Set



16	14	10	1	0	3
12	6	7	17	11	13
9	8	5	2	15	4

- Common set includes 18 HRC pairs
- Two SRCs are used to avoid too many repeated video contents

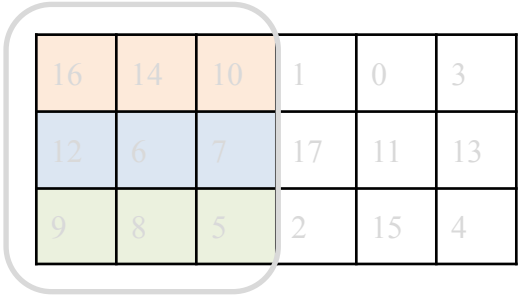
→ 18*2 = 36 common set pairs for each lab

For time-sequential lab (39 seconds/pair):
The test should be within 50 minutes for one observer → 50*60/39 = 77 pairs

36 common set pair with 2 video contents
The rest 77-36 = 41 pairs are shared by the rest 8 contents

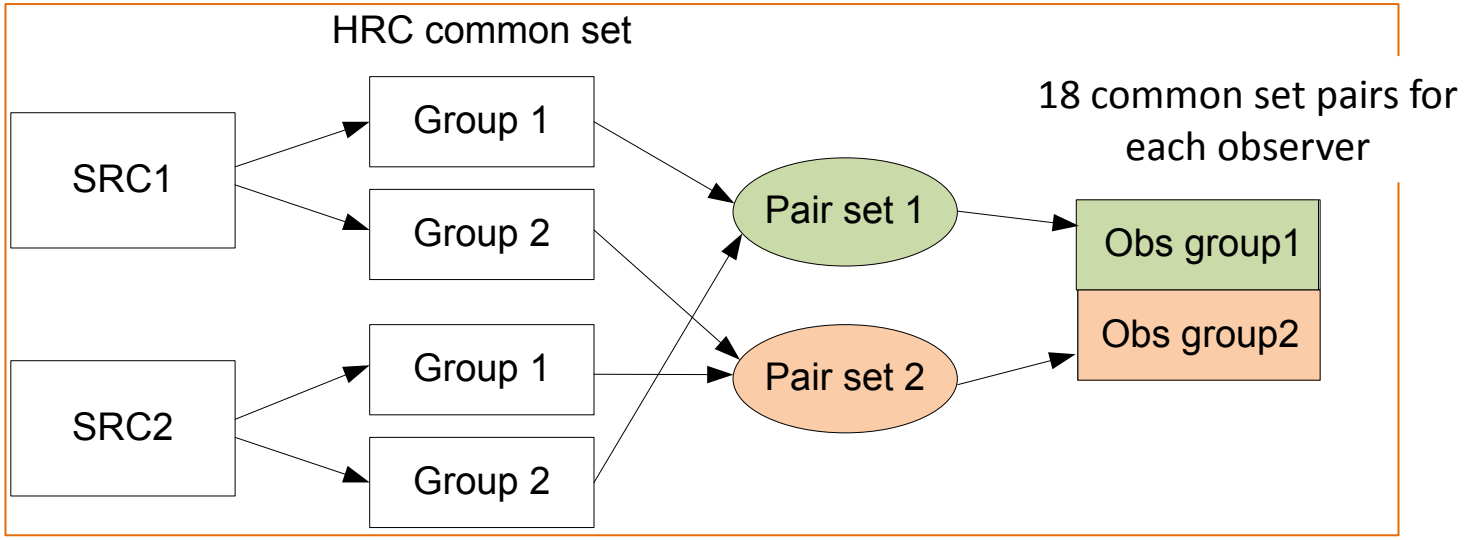
NO! Too many common set pairs!!!

Design of Common Set



- Common set includes 18 pairs
- Two SRCs are used to avoid too many repeated video contents

→ 18*2 = 36 common set pairs for each lab

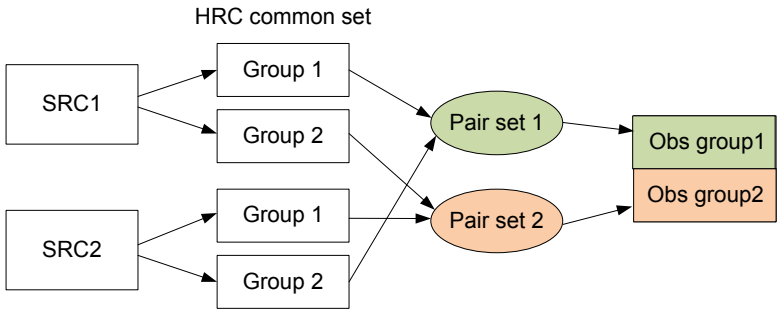


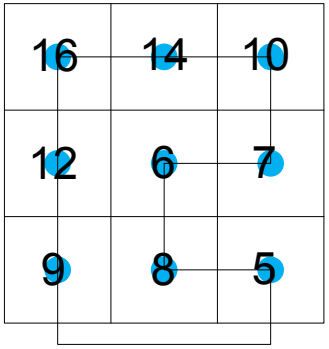
Design of Common Set

16	14	10	1	0	3
12	6	7	17	11	13
9	8	5	2	15	4

- Common set includes 18 pairs
- Two SRCs are used to avoid too many repeated video contents

→ 18*2 = 36 common set pairs for each lab





HRC Group1: (16,14), (14,10), (10,7), (7,6), (6,8), (8,5), (5,9), (9,12), (12,16).

HRC Group2: (16,10), (12,6), (12,7), (9,8), (16,9), (14,6), (14,8), (10,5), (7,5).

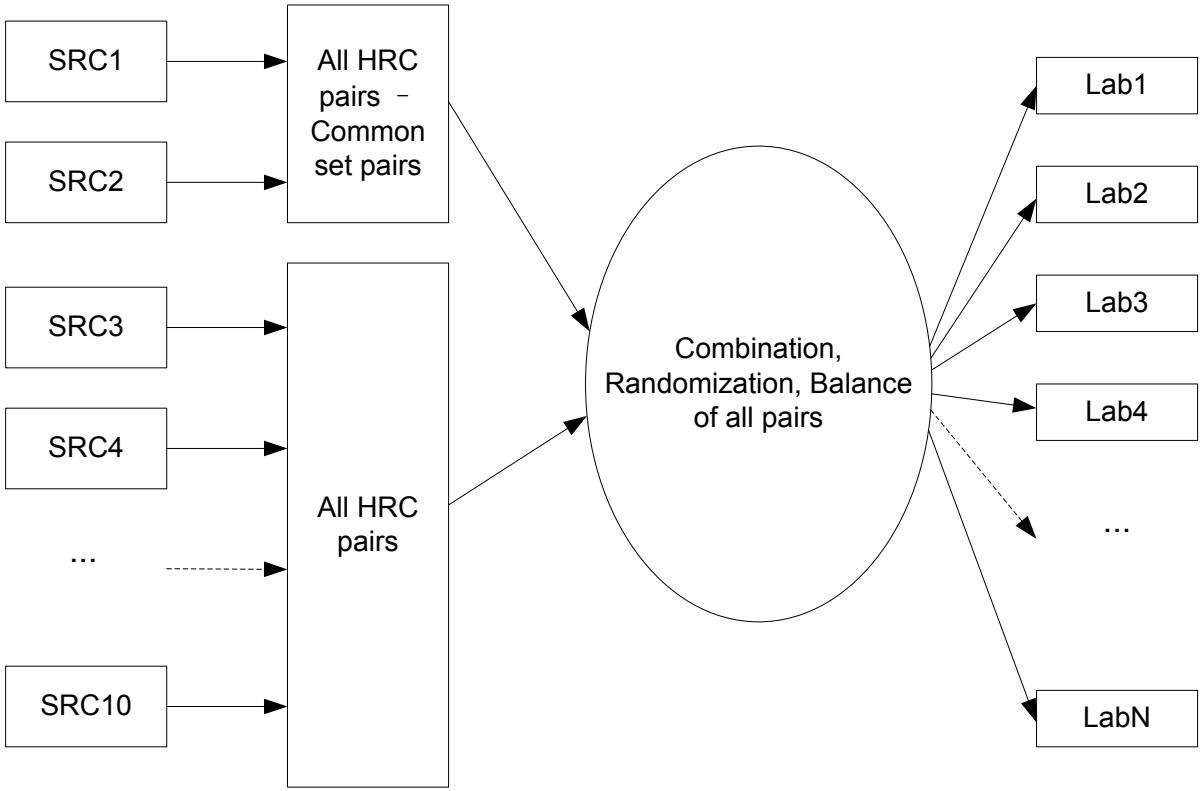
All HRCs have the same occurrency → balance

Assignment of all the other pairs

16	14	10	1	0	3
12	6	7	17	11	13
9	8	5	2	15	4

ORD

63 pairs for each SRC



Assignment of all HRC pairs

SRC	Time-parallel method				Time-sequential method			
	Lab1	Lab2	Lab3	Lab4	Lab5	Lab6	Lab7	Lab8
1	8+9(9)	8+9(9)	8+9(9)	8+9(9)	4+9(9)	3+9(9)	3+9(9)	3+9(9)
2	8+9(9)	8+9(9)	8+9(9)	8+9(9)	3+9(9)	4+9(9)	3+9(9)	3+9(9)
3	11	11	11	11	5	5	5	4
4	11	11	11	11	5	5	4	5
5	11	11	11	11	5	4	5	5
6	11	11	11	11	4	5	5	5
7	11	11	11	11	5	5	5	4
8	11	11	11	11	5	5	4	5
9	11	11	11	11	5	4	5	5
10	11	11	11	11	4	5	5	5
No. of Pairs/obs	122	122	122	122	63	63	62	62

Playlist for each lab

For each lab, the playlist has the following constraints:

- The consecutive contents are not same.
- The presentation order AB and BA are balanced for all observers.

Each lab will be provided a playlist by IRCCyN with the following format:

Order	Observer 1			Observer 2			…
	SRC	HRC-left	HRC-right	SRC	HRC-left	HRC-right	
1	3	12	6	4	6	11	
2	1	7	17	7	12	7	
3	6	14	1	8	15	11	
…							

▪ The pair presentation order MUST be strictly follow the playlist provided for each observer in each lab

Subjective data format

- 3 individual spreadsheets for
  - " experiment description”
  - “ observer information”
  - “ result data”

Subjective data format

- 3 individual spreadsheets for
  - **" experiment description”**
  - "observer information”
  - "result data”

•Lab name
•Subjective test description (e.g., VQEG GroTruQoE3D)
•Test environment (living room or standard lab)
•Test method (e.g., Time-Parallel pair comparison)
•Display model
•Display size
•Display resolution
•Display technology (shutter glasses or polarize)
•Display calibration tool
•Viewing distance
•Number of observers
•Source content description (the ID should be consistent for all labs)
•HRC description (the ID should be consistent for all labs)

Subjective data format

- 3 individual spreadsheets for
  - " experiment description”
  - **"observer information”**
  - "result data”

Observer ID
Visual Acuity Left
Visual Acuity Right
Color Vision
Depth Acuity
Experience 3D
Age
Nationality
Gender
Experience of Quality Test
Eye correction

Subjective data format

- 3 individual spreadsheets for
  - " experiment description”
  - "observer information”
  - **"result data” for each observer**

Order	SRC	HRC-left	HRC-right	Video file name	Voting duration	Voting result
1	1	0	2	src1_hrc0_v01.avi src1_hrc2_v01.avi	4.1s	L
2	3	1	4	Src3_hrc1_v01.avi src3_hrc4_v01.avi	2.2s	L
3	7	2	3	Src7_hrc2_v01.avi src7_hrc3_v01.avi	3.4s	R
…						

Outline

- Introduction of video sequences
- Why Pair Comparison test method is selected
- How to boost pair comparison
- How the experiment being conducted collaboratively
- What is the requirement for test setup
- How to analyze the data

Test setup requirement

- Test environment: BT500
- 3D display: must be calibrated
- Viewing distance: 3H for shutter glasses, 4.5H for polarized display
- Number of observers: 40
- Pre-test vision check
- Training
- Prior and post- QUESTIONNAIRES
- … see test plan

Test setup requirement

For Time-parallel presentation:

- Two 3D displays
- They must be the same model
- They must be synchronized when displaying

For all:

- Observer’s ID, voting results, stimulus pair information must be recorded synchronously.

Outline

- Introduction of video sequences
- Why Pair Comparison test method is selected
- How to boost pair comparison
- How the experiment being conducted collaboratively
- What is the requirement for test setup
- How to analyze the data

Data analysis

- Common set

  → cross-lab results verification

Data analysis

- Common set

  → cross-lab results verification

➢ Barnard's exact test

Pair AB	Lab 1	Lab 2	Total
Choose A	m1	m2	M=m1+m2
Choose B	N1-m1	N2-m2	N-m
Total Number	N1	N2	N

Test: if m1/N1 is significantly different from m2/N2

For example, 5/20 is NOT significantly different from 8/20 (*p*-value = 0.26 >0.05)

Data analysis

- Common set

  → cross-lab results verification

➢ Barnard's exact test

total number of sig. diff. pairs

HRC pair	Lab 1	Lab2	Sig.diff
5 vs 7	5/20	6/20	NO
5 vs 8	8/20	7/20	NO
6 vs 7	10/20	10/20	NO
6 vs 8	7/20	9/20	NO
7 vs 10	3/20	9/20	YES
7 vs 12	14/20	13/20	NO
8 vs 9	11/20	10/20	NO
…			

Data analysis

- Common set

  → cross-lab results verification

➢ Barnard's exact test

  <u>18 common set pairs per SRC</u>

- If the total number of sig. diff. pair (per SRC)between lab1 and lab2 < =2

  → results of lab1 and lab2 are NOT sig. diff. (error level =0.05)

  → can be combined…
- If the total number of sig. diff pair (per SRC) between lab1 and lab2 <=3

  → results of lab1 and lab2 are NOT sig. diff. (error level =0.1)
- Otherwise, further analysis is necessary.

More details see <test plan>

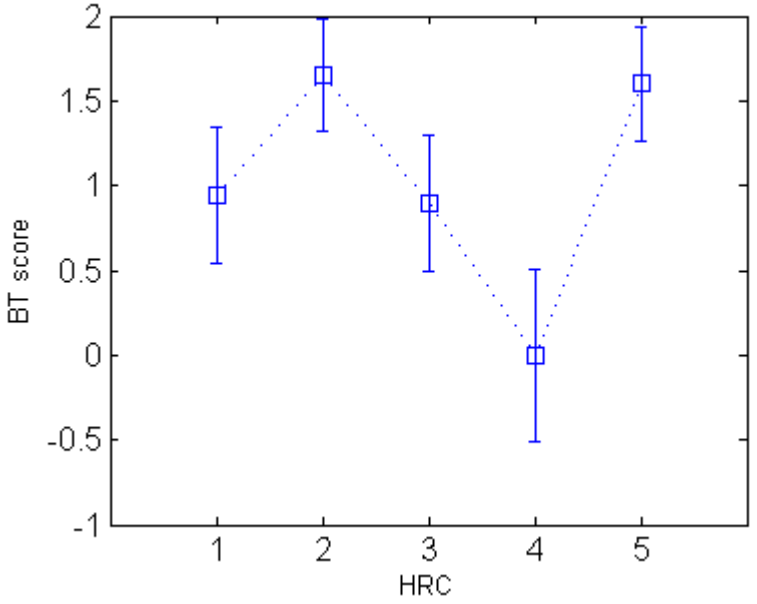
Data analysis

- Bradley-Terry (BT) model

→ convert Pair Comparison data to scale values

HRC	1	2	3	4	5
1	0	7	10	15	6
2	13	0	14	16	11
3	10	6	0	14	7
4	5	4	6	0	3
5	14	9	13	17	0

BT model



For pair HRC1-HRC2,
7 observers chose HRC1
13 observers chose HRC2

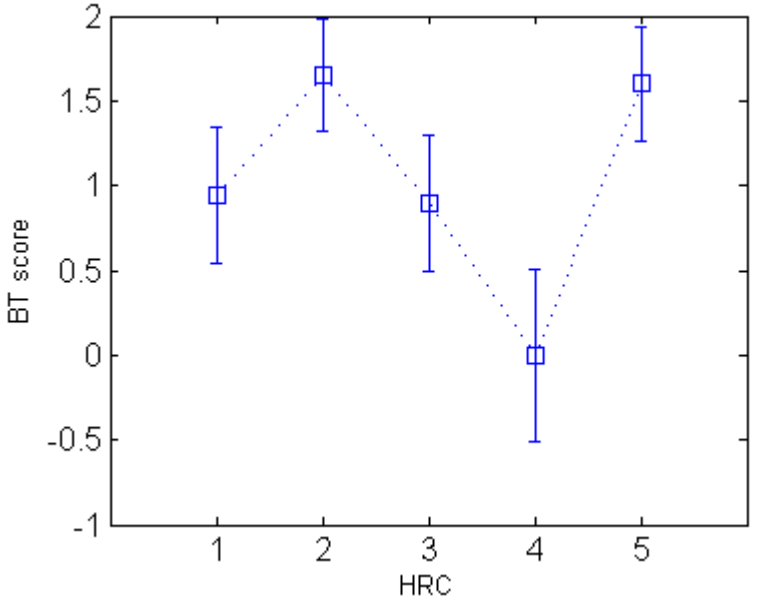
Data analysis

➢ Bradley-Terry (BT) model

→ convert Pair Comparison data to scale values

Evaluation/Comparison:

- PLCC
- SROCC
- RMSE
- …



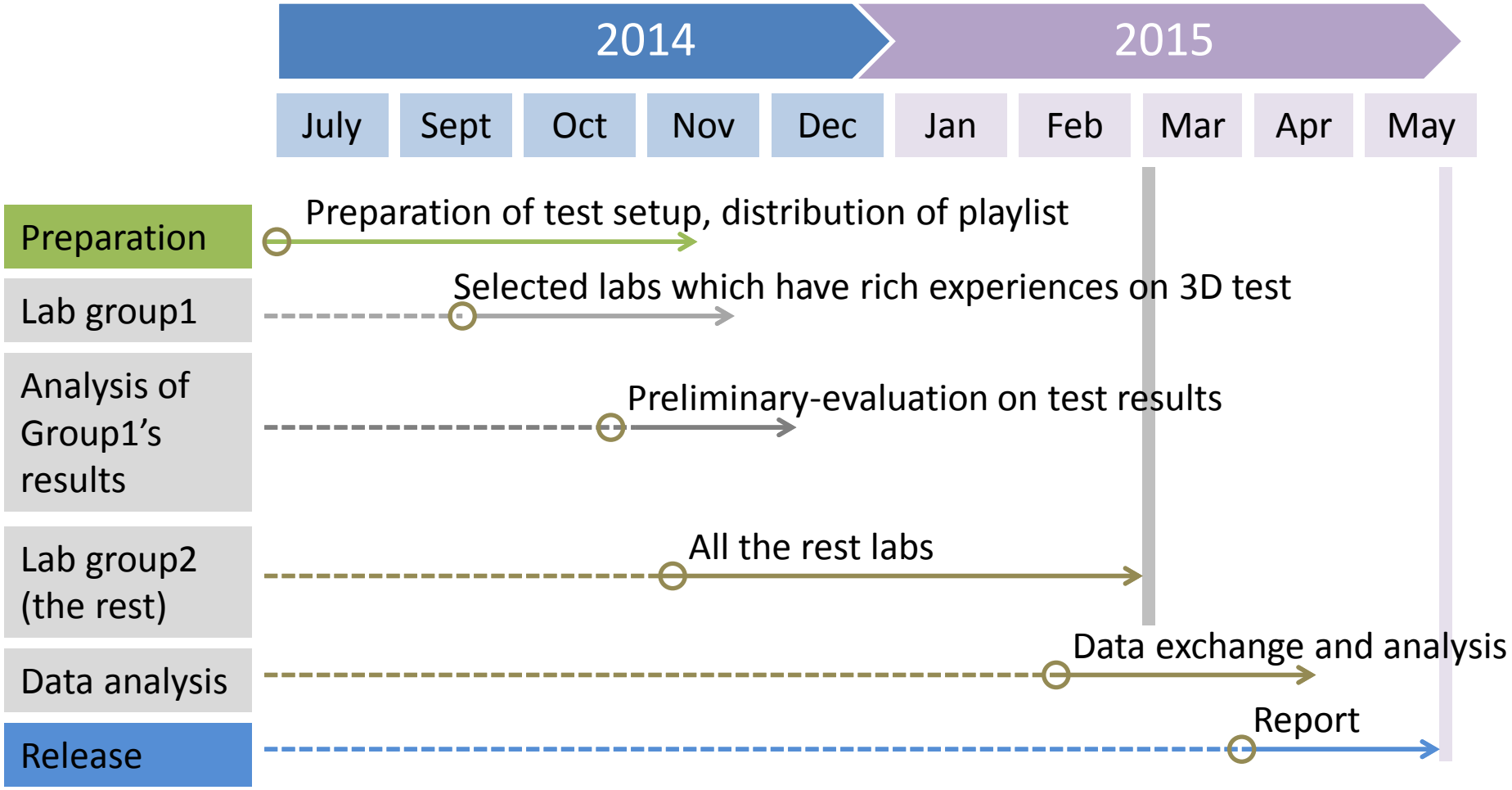
The BT score generated by the combined pair comparison data is the subjective score of the database

Summary

- Subjective methodology → BT500, P910
- "…However, as QoE of 3DTV is a multidimensional concept, how to measure it is still a question…"
- " …Pair Comparison is one of the optimal solutions, however, it is time infeasible…"
- …

We have ORD (and ARD) to reduce the number of pairs for Pair Comparison !

Schedule


2014
2015
July
Sept
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Preparation
Lab group1
Analysis of Group1's results
Lab group2 (the rest)
Data analysis
Release
Preparation of test setup, distribution of playlist
Selected labs which have rich experiences on 3D test
Preliminary-evaluation on test results
All the rest labs
Data exchange and analysis
Report

Participating labs

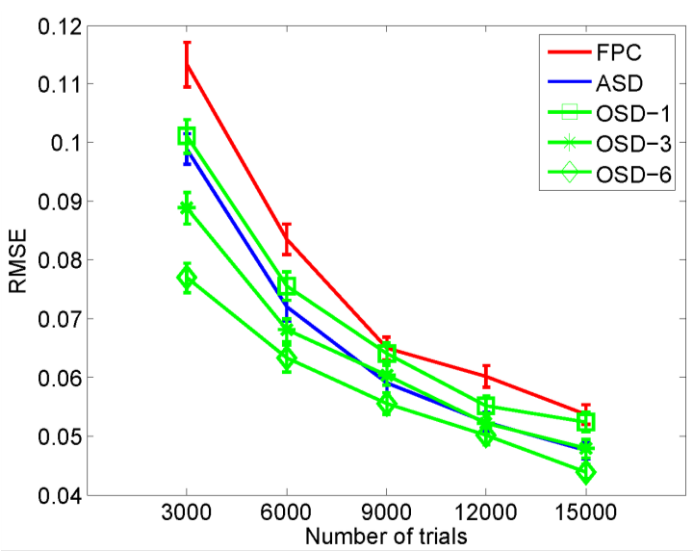
LAB	YES/NO	Method*	Group1	Group2
Acreo				
NTIA				
Vittorio				
Intel				
KDDI				
AGH				
Yonsei				
Orange				
UPM	Y	P	X	
INSA	Y	S		X
IRCCyN	Y	P	X	

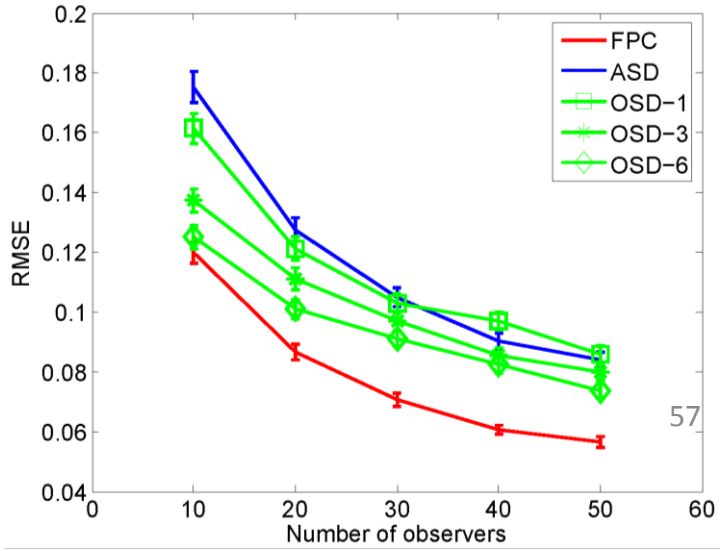
*Method: time-parallel (P) or time-sequential (S)

Appendix I: Evaluation of ORD

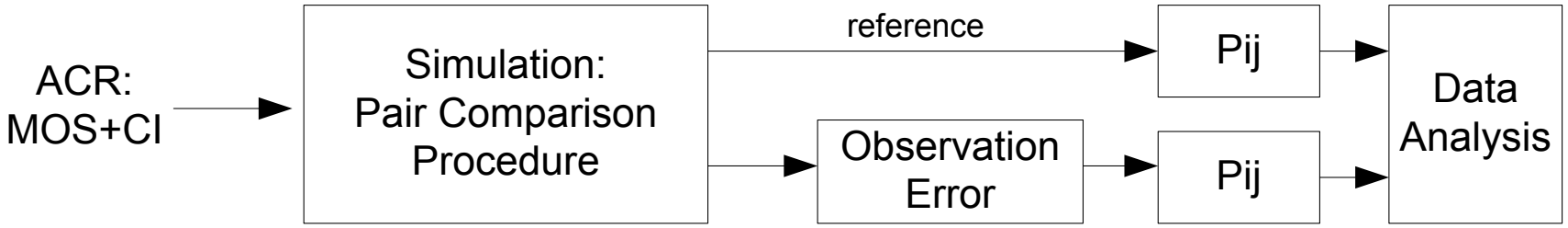
- Monte-Carlo simulation test

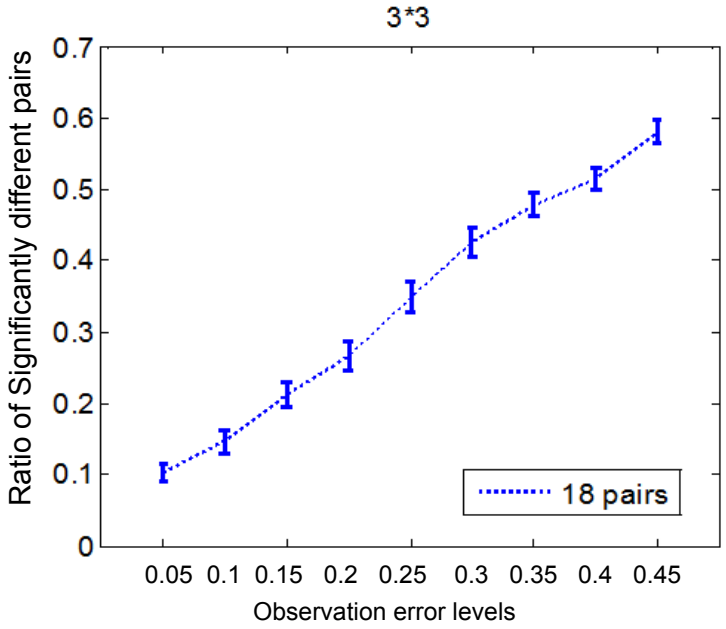
- Each stimulus has a single score;
- In each observation, the observed value follows a Gaussian distribution, the mean value is the stimulus score and the standard deviation is 0.7, which is obtained from the subjective scores of VQEG HDTV Final Report;
- Each observer has a 5% probability to make a mistake on an observation, i.e., inverting the vote;
- Each comparison is independent.





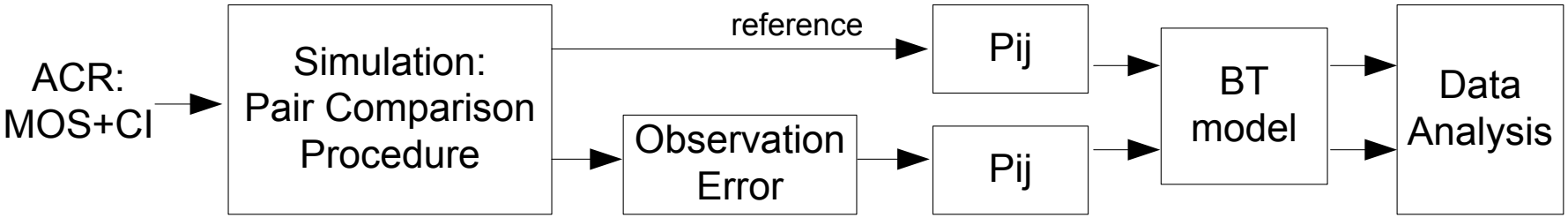
Appendix II: Barnard's exact test on Common set

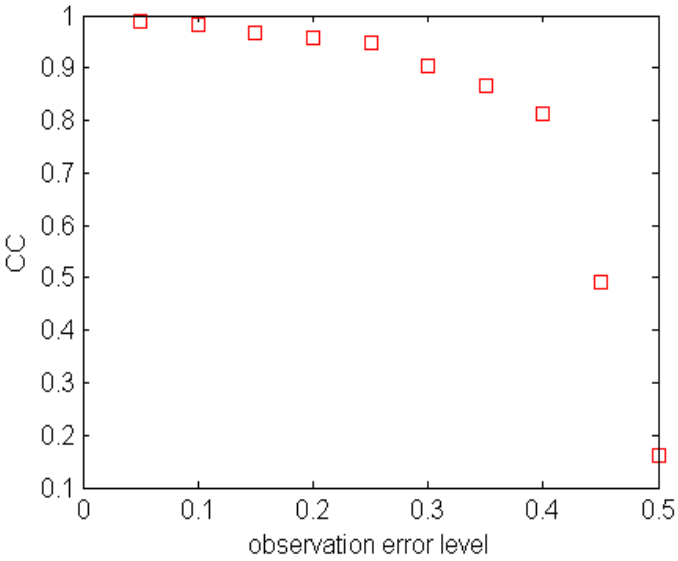


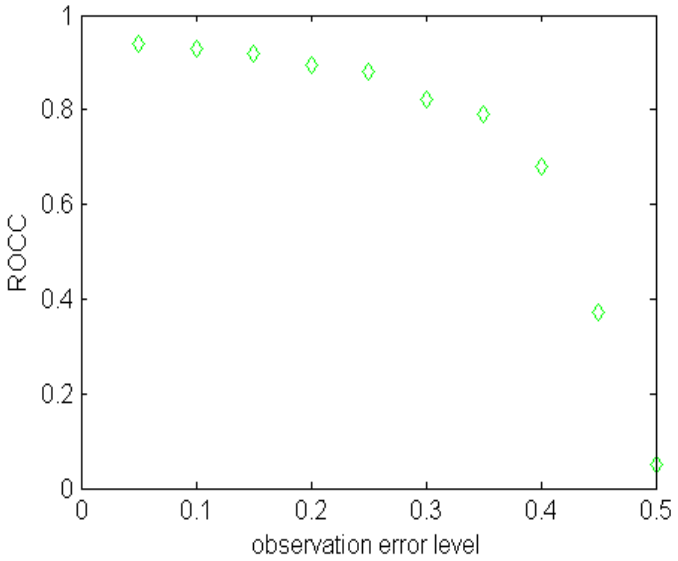


20 observations/pair
Obs error level = 0.05 → 0.1*20 = 2 sig. diff.
Obs error level = 0.1 → 0.15*20=3 sig. diff

Appendix II: BT score analysis on Common set

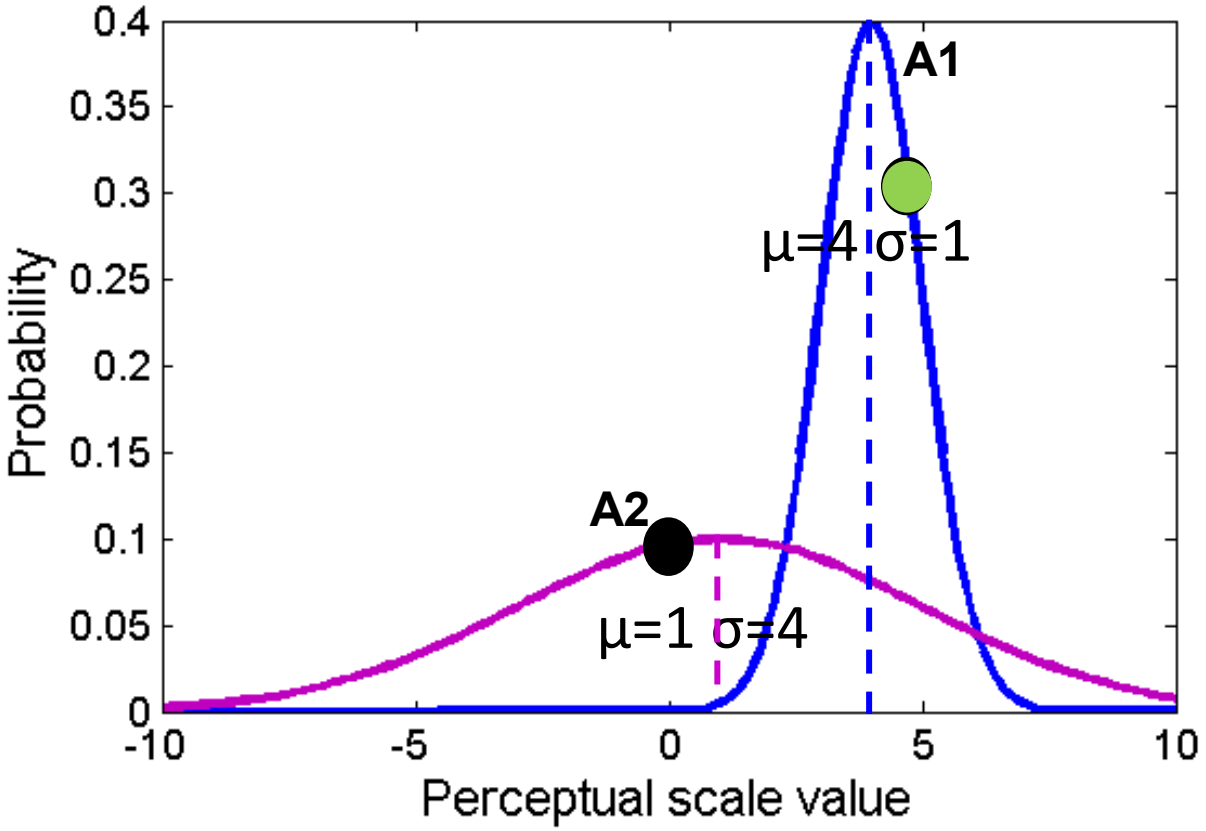






Pair Comparison procedure

- A given physical stimulus does not always produce the same psychological experience.



Pair Comparison procedure

- A given physical stimulus does not always produce the same psychological experience.