IN5060

Performance in distributed systems

User studies

UiO : University of Oslo

Why user studies?

- Just because something is technically possible doesn't mean it improves human experiences.
  - 8K video on a 2015 iPhone?
- You cannot be sure that a new technology can rely on old assumptions.
  - in games, higher frame rates are good for fluid gameplay
  - but the actual reason is that processing loops are tied to frame rate, so higher frame rate leads to faster rendering
- You cannot be sure that your own intuition holds for the majority of humankind.
  - timed text must scroll from right to left
  - Powerpoint menus should be at the top of the window, independent of OS style guide and screen aspect ratio

Why user studies?

- A classical multimedia example

Peak Signal-to-Noise Ratio
A prevalent video quality metric

$$PSNR = 10\log_{10}\frac{(2^B-1)^2}{MSE}$$

where:

$$MSE = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}[Im_a(x,y) - Im_b(x,y)]^2$$

M, N = image dimensions
Im_a , Im_b = pictures to compare
B= bit depth

Why user studies?

Reference

Example from
Prof. Touradj Ebrahimi,
ACM MM'09 keynote

PSNR = 24.9 dB

PSNR = 24.9 dB

PSNR = 24.9 dB

Why user studies?

Peak Signal-to-Noise Ratio
A prevalent video quality metric

In addition to this:
- several different PSNR computations for color images
- different PSNR for different color spaces (RGB,YUV)
- visible influence of the encoding format

These problems hurts all metrics that are based on PSNR

Improved by image quality metrics such as
- SSIM variants
- rate distortion metrics

never believe a statement where PSNR is used for video quality estimation

Quality assessment methods

most of these are described and named in
Recommendations (standards) of the ITU

Types

- Single Stimulus methods
  - ACR: Absolute Category Rating
    - each sample separately, no reference
    - rating on 5-point Likert scale
      - possibly named categories: intolerable … excellent
      - possibly numbered categories: 1 … 5
    - video sample should not be 8-12 seconds long
  - ACR-HR: Absolute Category Rating with Hidden Reference
    - start like ACR
    - calculate ratings as differences between reference rating and sample rating
  - SSCQE: Single Stimulus Continuous Quality Evaluation
    - watch a single (long) sample with quality that varies over time
    - use a slider (0-100) for continuous rating

Types

- Double Stimulus methods
  - DSCQS: Double Stimulus Continuous Quality Scale
    - watch unimpaired reference and impaired sample in random order
    - repeat watching as long as desired
    - rate quality of both on continuous scale 1-5
  - DSIS: Double Stimulus Impairment Scale / DCR: Degradation Category Rating
    - watch unimpaired reference followed by impaired sample
    - use categories to rate
      (impairment imperceptible … impairment very annoying)
  - PC: Pair Comparison
    - watch two impaired samples
    - rate which one was better
    - randomness is extremely important

Types

- Other methods
  - SDSCE: Simultaneous Double Stimulus for Continuous Evaluation
    - double stimulus method where two samples are shown side-by-side
    - rating on continuous scale 0-100
  - SAMVIQ: Subjective Assessment Methodology for Video Quality
    - explicit reference, hidden reference, up to 10 measured samples
    - participant may repeat watching, last score stands
    - continuous scale 0-100

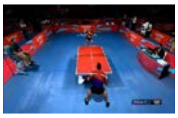
User studies and human memory

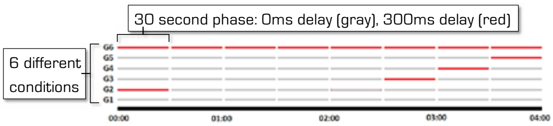
"Influence of Primacy, Recency and Peak effects on the Game Experience Questionnaire"

paper by Saeed Shafiee (Simula) et al.

Example: delay in cloud games

"Influence of Primacy, Recency and Peak effects on the Game Experience Questionnaire"





Example: delay in cloud games

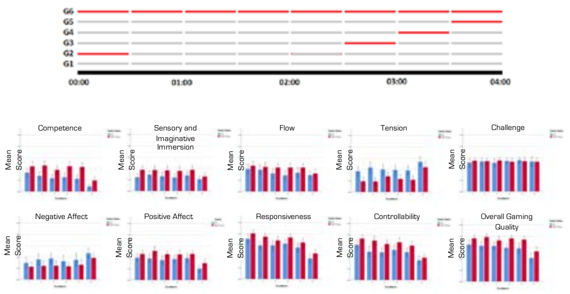
"Influence of Primacy, Recency and Peak effects on the Game Experience Questionnaire"

- GEQ – game experience questionnaire
  - 33 Questions
  - Assessing seven aspects of gaming QoE
  - Peak Effect
  - Very popular and widely used
  - ITU-T P.Game
- Additional questions
  - How do you rate the overall quality of your gaming experience?
  - The game has responded as expected to my inputs.
  - I had control over the game.

	not at all	slightly	moderately	fairly	extremely
I felt content					
I felt skilful					
I was interested in the game's story					
I thought it was fun					
I was fully occupied with the game					
I felt happy					
It gave me a bad mood					
I thought about other things					
I found it tiresome					
I felt competent					
I thought it was hard					
It was aesthetically pleasing					
I forgot everything around me					
I felt good					
I was good at it					

Example: delay in cloud games

"Influence of Primacy, Recency and Peak effects on the Game Experience Questionnaire"

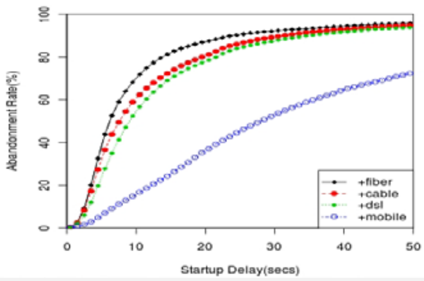


How tolerant are video users to startup delay?

paper at IMC 2012 by
Ramesh K. Sitaraman
(UMass Amherst & Akamai) and
S. Shunmuga Krishnan (Akamai)

Main result

Viewers with better connectivity have less patience for startup delay and abandon sooner.



Slides by Prof. Ramesh Sitaranam, Umass, Amherst (shown with permission)
"Video Stream Quality Impacts Viewer Behavior: Inferring Causality using Quasi-Experimental Designs", S. S. Krishnan and R. Sitaraman, ACM Internet Measurement Conference (IMC), Boston, MA, Nov 2012

Data set

- One of the most extensive data sets to that date
- analyzed data from a widely deployed Akamai client-side plug-in
  - 10 days
  - 12 content providers
  - 23 million views
  - 216 million minutes of video played
  - 102.000 videos
  - 1431 TB of video bytes
  - 3 continents
  - VoD only

Flickering in video streaming

by Pengpeng Ni (Simula) et al., 2011

Image-based metrics can fail badly:
Flickering

3 origins of flicker

Flicker arises from recurrent changes in spatial or temporal quality, some so rapid that the human visual system only perceives fluctuations within the video.

Noise flicker	Blur flicker	Motion flicker

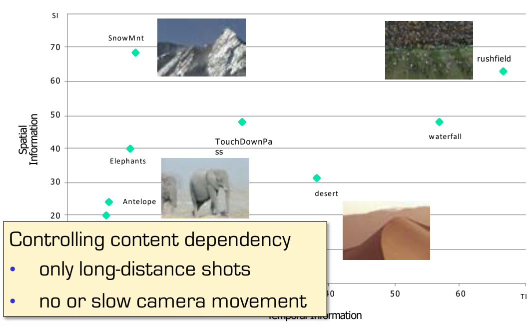
Compression scaling	Resolution scaling	Frame rate scaling

Assessment of video adaptation strategies

To cope with the bandwidth fluctuation, which scalability dimension is generally preferable for video adaptation?

- Within each dimension, which scaling pattern generates the least annoying flicker effect?
- Is it possible to control the annoyance of flicker effects?
- How is subjective video quality related to other factors, such as content, devices?

Video content selection



Controlling content dependency
- only long-distance shots
- no or slow camera movement

Noise flicker example

Noise flicker
Amplitude: QP24 – QP40
Frequency: 10f / 3 Hz

Blurriness flicker example

Blur flicker
Amplitude: 480x320px – 120x80px
Frequency: 15f / 2 Hz

Motion flicker example

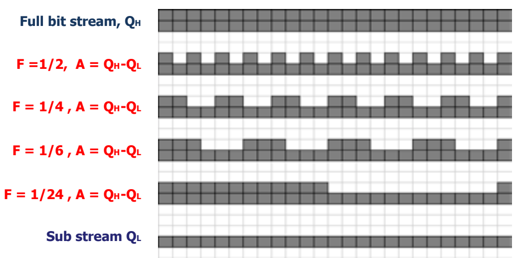
Motion flicker
Amplitude: 30fps – 3fps
Frequency: 6f / 5 Hz

How to describe different layer fluctuations?

- Layer fluctuation pattern
  - Frequency: The time interval it takes for a video sequence return to its previous status
  - Amplitude: The quality difference between the two layers being switched
  - Dimension: Spatial or temporal, artifact type

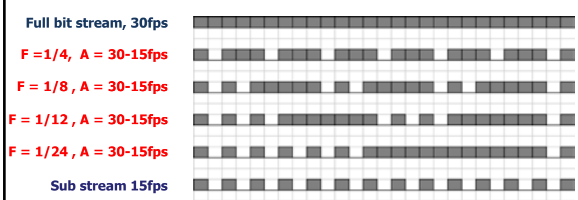
Layer Frequency and Amplitude are the interesting factors in our subjective test

Layer fluctuation pattern in Spatial dimension



Bandwidth consumption in all of these patterns is the same, due to the same amplitude.

Layer fluctuation pattern in Temporal dimension



Although the average bit-rate is the same, the visual experience of different patterns may not be identical.

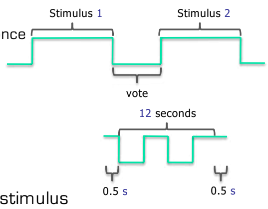
Method

Participants
- 28 paid, voluntary participants
- 9 females, 19 males
- Age 19 – 41 years (mean 24)
- Self-reported normal hearing, and normal/corrected vision

Procedure
- Field study at university library
- Presented on iPod touch devices
  - Resolution 480x320
  - Frame rate 30 fps
- 12 sec video duration
- Random presentations
- Optional number of blocks

Test procedure

We use the Single Stimulus (SS) method to collect responses from subjects
- Each test stimulus is displayed only once

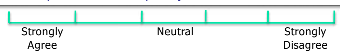


Each stimulus lasts for 12 seconds
based on previous study about memory effect

Two responses collected after each stimulus

I think the video quality was at a stable level: Yes or No

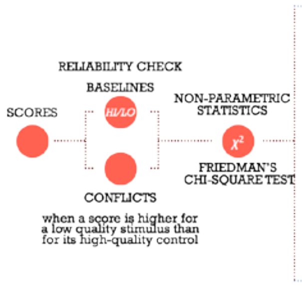
I accept the overall quality of the video: 5-likert scale



Design & Analysis

- Repeated measures
- Friedman's Chi-square test
- Stimuli blocked by flicker and amplitude
- Responses to stability measure converted to binomial scores
- Quality ratings converted to ordinal scores ranging from -2 (least acceptable) to 2 (most acceptable)
  - we can assume ORDER between scores
  - we cannot assume equidistance between scores
- Results for experimental stimuli assessed relative to control stimuli of constant high or low quality

Analysis



Main Effects

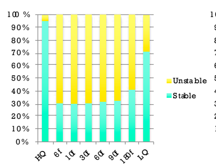
effect of an independent variable on a dependent variable, averaged across the other independent variables

Interaction Effects

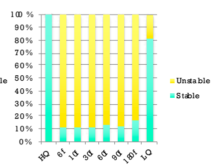
effect of the levels of one independent variable, across the levels of another independent variable, on a dependent variable

Stability scores - Period

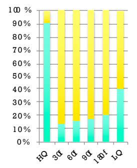
I think the video quality was at a stable level: Yes or No



Perceived quality stability across period levels for *Noise flicker*



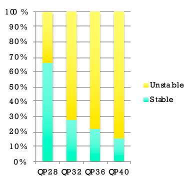
Perceived quality stability across period levels for *Blur flicker*



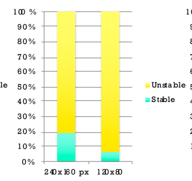
Perceived quality stability across period levels for *Motion flicker*

Stability scores - Amplitude

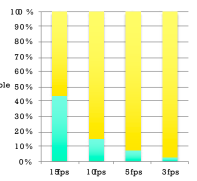
I think the video quality was at a stable level: Yes or No



Perceived quality stability across amplitude levels for *Noise flicker*



Perceived quality stability across amplitude levels for *Blur flicker*



Perceived quality stability across amplitude levels for *Motion flicker*

Significance of results

I think the video quality was at a stable level: Yes or No

noise

Options	Stable	Unstable	P-value	Signif.
QP28	65.8%	34.2%	3.66e-12	+
QP32	27.7%	72.3%	4.49e-23	–
QP36	21.7%	78.3%	3.51e-37	–
QP40	15.6%	84.4%	8.74e-56	–

blur

Options	Stable	Unstable	P-value	Signif.
240x160	19.3%	80.7%	4.89e-31	–
120x80	06.6%	93.5%	2.57e-67	–

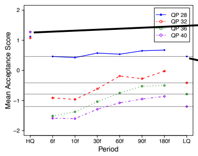
motion

Options	Stable	Unstable	P-value	Signif.
15fps	43.8%	56.2%	0.045	(*)
10fps	15.1%	84.9%	2.62e-33	–
5fps	07.4%	92.6%	2.82e-52	–
3fps	02.9%	97.1%	1.82e-67	–

+ stable, significant
- unstable, significant
(*) not significant

Video quality

I accept the overall quality of the video: 5-likert scale



Constant high quality references

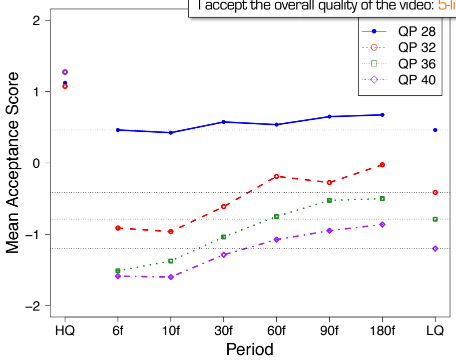
Constant low quality reference, QP28

Not investigated here: relation between qualities

Noise	
L1	QP24
L0	QP28, QP32, QP36, QP40
Period	1/5s, 1/3s, 1s, 2s, 3s, 6s
Content	4 mid/long distance shots

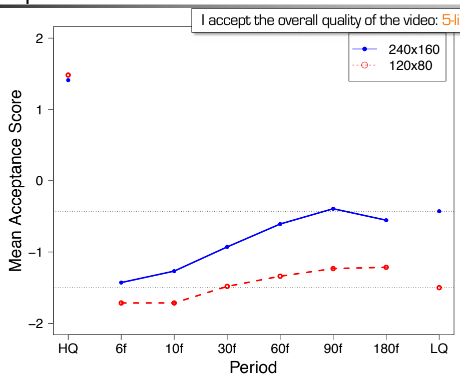
Acceptance - Noise flicker

I accept the overall quality of the video: 5-likert scale



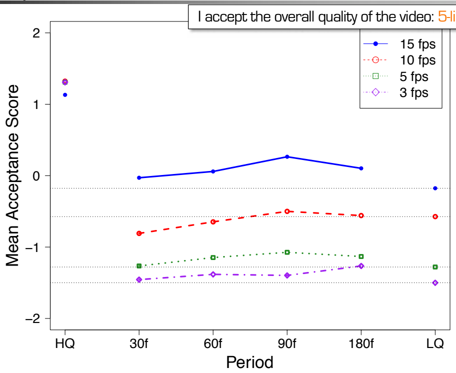
Acceptance – Blur flicker

I accept the overall quality of the video: 5-likert scale



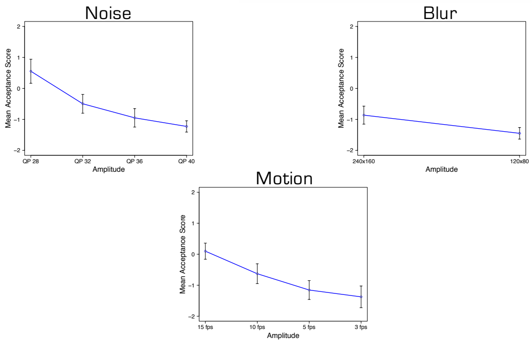
Acceptance – Motion flicker

I accept the overall quality of the video: 5-likert scale



Acceptance

I accept the overall quality of the video: 5-likert scale



Conclusions

- With longer flicker frequencies (high periods), acceptance of video quality increases in the spatial dimension
- Amplitude (quality difference) has larger effect than frequency, both for stability and acceptance
- For noise flicker, large quality differences are rated more acceptable with less frequent quality shifts.
- For blur flicker, improved acceptance with less frequent shifts is more pronounced for the smallest quality difference.
- The flicker effect varies across contents, particularly for motion flicker.
- The three types of flicker have different influences on stability and quality acceptance scores. Scores are generally lower for blur flicker.

Friedman's Chi^2 (or X^2) test

Friedman's X^2 test

- This is a test to verify the relevance of categorical data
- That means that you can use it when you cannot (or should not) compute distances between the possible values of the responses
- Examples:
  - did you like it / not like it
  - did it look red / green / blue
  - was is stable / unstable

Noise flicker example – separate relevance tests

\ settings(k) participants(n)	QP 28	QP 32	QP 36	QP 40	Σ
#1	$r_{1,1}$	$r_{1,2}$	$r_{1,3}$	$r_{1,4}$	$r_{1\cdot}$
...	...	...	...	...	...
#28	$r_{28,1}$	$r_{28,2}$	$r_{28,3}$	$r_{28,4}$	$r_{28\cdot}$
Σ	$r_{\cdot 1}$	$r_{\cdot 2}$	$r_{\cdot 3}$	$r_{\cdot 4}$	

ranks for quality ratings (how often was it stable) average if equal

compute Q :

$$Q = \frac{12}{nk(k+1)} \sum_{i=1}^{k} (r_{\cdot i})^2 - (3n(k+1))$$

If the sum Q is larger than the tabulated lookup value for the X^2 distribution, the result is relevant

For k=4 and p=0.001, the limit for X^2_{k-1} is 16.27

If the X^2 succeeds (Q>16.27), you can say that the ranking determined by the values $\overline{r_{\cdot j}}$ is **relevant**.

You must **never** interpret p for anything more.

Relevance tables for X^2

- https://web.ma.utexas.edu/users/davis/375/popecol/tables/chisq.html
- Some tools, like SPSS, can compute the result from the tables

Does blur hide asynchrony?

study by Ragnhild Eg (Simula) et al., 2011

Perception of synchrony

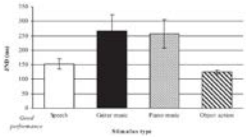
Sensitivity for perceptual synchrony is subjective and depends on the content

Spoken sentences (Grant et al., 2003)

- Discrimination thresholds: ≈50 ms audio lead, ≈200 ms aud io lag

Hitting table with wand (Levitin et al., 2000)

- Synchrony thresholds set to 75 %: 41 ms Alead to 45 ms Alag



Music, baseball, speech (Vatakis & Spence, 2006)

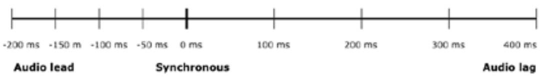
- Temporal order judgements (audio/video first)

Stimuli

3 content types

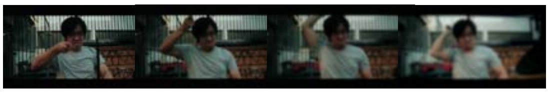
Chess game News broadcast Drummer

9 asynchrony levels



Stimuli

Visual distortion, 4 levels, Gaussian blur filter



Undistorted Blur 2x2 pixels Blur 4x4 pixels Blur 6x6 pixels

Procedure

- Carried out at the Speech Lab, NTNU

Chess content - 200 ms audio lead

Chess content - 200 ms audio lag, blurred

News - 300 ms audio lag, blurred

UiO : University of Oslo IN5060

Drums - 100 ms audio lag, blurred

UiO : University of Oslo IN5060

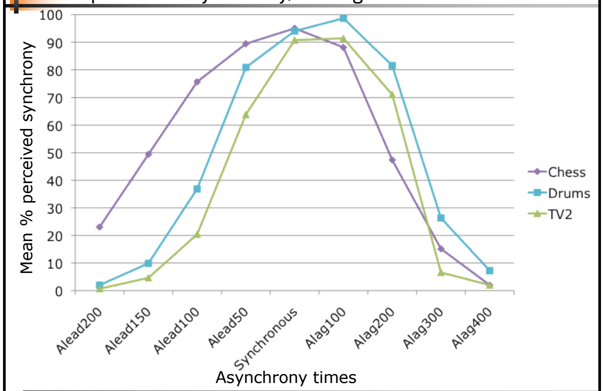
Drums - 150 ms audio lead, slightly blurred

UiO : University of Oslo IN5060

Design & Analysis

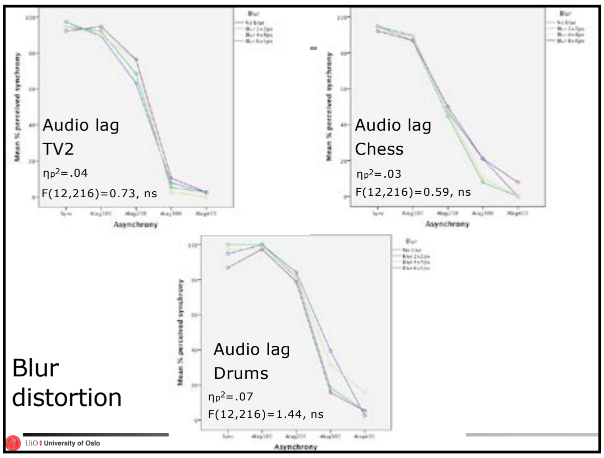
- 2 independent studies
- Full-factorial design
- 2 repetitions of each condition
- Binomial responses converted to percentages
- Repeated-measures ANOVAs
- Separate analyses for:
  - Audio lag and audio lead (different scales)
  - Content types (different response patterns)

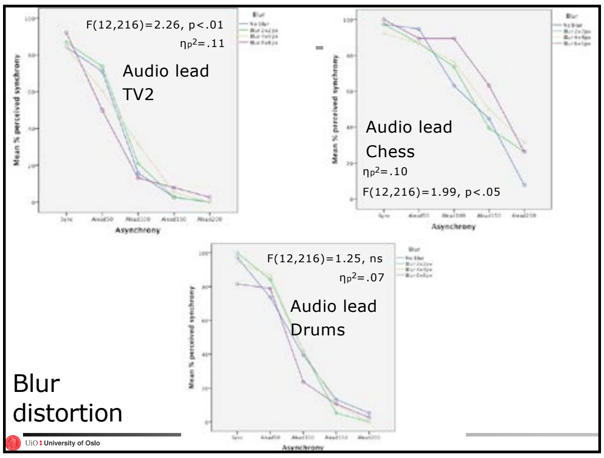
Mean perceived synchrony, averaged across **blur** levels



Assessment of relevance

	Visual distortion			Auditory distortion		
	Content	F-statistics	η_p^2	Content	F-statistics	η_p^2
Audio lag	Chess	F(4,72)=88.79, p<.001	0.83	Chess	F(4,48)=64.28, p<.001	0.84
	TV2	F(4,72)=232.54, p<.001	0.93	TV2	F(4,48)=80.50, p<.001	0.87
	Drums	F(4,72)=197.57, p<.001	0.92			
Audio lead	Chess	F(4,72)=71.77, p<.001	0.80	Chess	F(4,48)=55.16, p<.001	0.82
	TV2	F(4,72)=100.26, p<.001	0.85	TV2	F(4,48)=108.54, p<.001	0.90
	Drums	F(4,72)=126.31, p<.001	0.88			


Audio lag
TV2
η_p^2=.04
$F(12,216)=0.73$, ns
Audio lag
Chess
η_p^2=.03
$F(12,216)=0.59$, ns
Audio lag
Drums
η_p^2=.07
$F(12,216)=1.44$, ns
Blur
distortion
UiO : University of Oslo


$F(12,216)=2.26$, $p<.01$
η_p^2=.11
Audio lead
TV2
Audio lead
Chess
η_p^2=.10
$F(12,216)=1.99$, $p<.05$
$F(12,216)=1.25$, ns
η_p^2=.07
Audio lead
Drums
Blur
distortion
UiO : University of Oslo