PA198
Augmented Reality Interfaces

Lecture 8

Evaluating Augmented Reality Interfaces

Fotis Liarokapis

23rd November 2015

Introduction

Evaluating User Interfaces

- Assess effect of interface on user performance and satisfaction
- Identify specific usability problems
- Evaluate users' access to functionality of system
- Compare alternative systems/designs

Major Parameters

- The major parameters in the user interface evaluation activities are:
  - Stage of the design
  - Inspection methods vs. usability testing
  - Formative vs. summative

Developer watching videotape of usability test.

http://www.topdesignmag.com/navigate-usability-evaluation/

Influence of the Parameters

- These parameters influence:
  - How the design is represented to evaluators
  - Documents/deliverables required
  - Need for resources (personnel, equipment, lab)
  - Methodology
    - For data gathering
    - For analysis of results

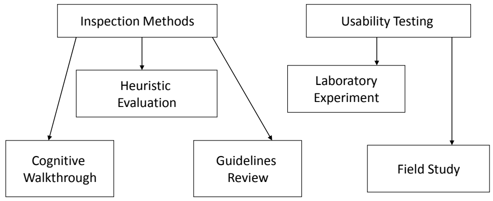
Methodologies for Data Gathering

- Structured inspection
- Interviews
- Focus groups
- Questionnaires
- Field studies
- Controlled Experiments
  - Quantitative metrics
  - Thinking aloud, cooperative evaluation

Evaluating User Interface Designs

- Stage of the design process
  - Early design (prototype)
  - Intermediate
  - Full design
  - After deployment
- Evaluation should be done throughout the usability life cycle – not just at the end
  - Called iterative design
- Different evaluation methods appropriate at different stages of the cycle

Evaluating User Interface Designs .



Inspection Methods

Heuristic Evaluation

Cognitive Walkthrough

Guidelines Review

Usability Testing

Laboratory Experiment

Field Study

Formative vs. Summative Evaluation

- Formative evaluation
  - Identify usability problems
    - Qualitative measures
    - Ethnographic methods
- Summative evaluation
  - Measure/compare user performance
    - Quantitative measures
    - Statistical methods

Participatory or User-centered Design

- Users are active members of the design team
- Characteristics
  - Context and task oriented rather than system oriented
  - Collaborative
  - Iterative
- Methods
  - Brain-storming ("focus groups")
  - Storyboarding
  - Workshops
  - Pencil and paper exercises

Cognitive Walkthrough

- Evaluates design on how well it supports user in learning task
- Usually performed by expert in cognitive psychology
- Expert `walks though' design to identify potential problems using psychological principles
- Scenarios may be used to guide analysis

Cognitive Walkthrough .

- For each task, walkthrough considers:
  - What impact will interaction have on user?
  - What cognitive processes are required?
  - What learning problems may occur?
- Analysis focuses on users goals and knowledge
  - Does the design lead the user to generate the correct goals?

Cognitive Walkthrough Video

https://www.youtube.com/watch?v=bzvQY68lm8c

Heuristic Evaluation

- Usability criteria (heuristics) are identified
- Design examined by experts to see if these are violated
- Example heuristics
  - System behavior is consistent
  - Feedback is provided
- Heuristic evaluation debugs design

Heuristic Evaluation Video

https://www.youtube.com/watch?v=lkbBc4aF5FA

Guidelines Inspection

- A usability group should have a designated inspector!
- Written guidelines recommended for larger projects:
  - Screen layout
  - Appearance of objects
  - Terminology
  - Wording of prompts and error messages
  - Menu's
  - Direct manipulation actions and feedback
  - On-line help and other documentation

Usability Experiment

What is a Usability Experiment?

- Usability testing in a controlled environment
  - There is a test set of users
  - They perform pre-specified tasks
  - Data is collected (quantitative and qualitative)
  - Take mean and/or median value of measured attributes
  - Compare to goal or another system
- Contrasted with expert review and field study evaluation methodologies
- Note the growth of usability groups and usability laboratories

Experimental Factors

- Subjects
  - Representative
  - Sufficient sample
- Variables
  - Independent variable (IV)
    - Characteristic changed to produce different conditions
      - i.e. Interface style, number of menu items
  - Dependent variable (DV)
    - Characteristics measured in the experiment
      - i.e. Time to perform task, number of errors

Experimental Factors .

- Hypothesis
  - Prediction of outcome framed in terms of IV and DV
  - Null hypothesis: states no difference between conditions and the aim is to disprove this
- Experimental design
  - Within groups design
  - Between groups design

Within Groups Design

- Each subject performs experiment under each condition
- Advantages
  - Fewer subjects needed
  - Less likely to suffer from user variation
- Disadvantages
  - Transfer of learning possible

Between Groups Design

- Each subject performs under only one condition
- Advantages
  - No transfer of learning
- Disadvantages
  - More subjects required (therefore more costly)
  - User variation can bias results

How Many Test Users?

- Problems-found (i) = N (1 - (1 - l)i)
  - i = number of test users
  - N = number of existing problems
  - l = probability of finding a single problem with a single user

Data Collection Techniques

- Paper and pencil
  - Cheap, limited to writing speed
- Audio
  - Good for think aloud, difficult to match with other protocols
- Video
  - Accurate and realistic, needs special equipment, obtrusive
- Computer logging
  - Automatic and unobtrusive
  - Large amounts of data difficult to analyze

Data Collection Techniques .

- User notebooks
  - – Coarse and subjective, useful insights
  - – Good for longitudinal studies
- Brain logging
  - – More difficult technique

Summative Evaluation

- What to measure?
  - – Total task time
  - – User “think time” (dead time??)
  - – Time spent not moving toward goal
  - – Ratio of successful actions/errors
  - – Commands used/not used
  - – Frequency of user expression of:
    - Confusion, frustration, satisfaction
  - – Frequency of reference to manuals/help system
  - – Percent of time such reference provided the needed answer

Measuring User Performance

- Measuring learnability
  - – Time to complete a set of tasks by novice
  - – Learnability/efficiency trade-off
- Measuring efficiency
  - – Time to complete a set of tasks by expert
  - – How to define and locate ‘experienced’ users
- Measuring memorability
  - – The most difficult, since ‘casual’ users are hard to find for experiments
  - – Memory quizzes may be misleading

Measuring User Performance .

- Measuring user satisfaction
  - – Likert scale (agree or disagree)
  - – Semantic differential scale
  - – Physiological measure of stress
  - – EEG measures
- Measuring errors
  - – Classification of minor vs. serious
  - – Removing noise

Reliability and Validity

- Reliability means repeatability
  - – Statistical significance is a measure of reliability
  - – Difficult to achieve because of high variability in individual user performance
- Validity means will the results transfer into a real-life situation
  - – Depends on matching the users, task, environment
  - – Difficult to achieve because real-world users, environment and tasks difficult to duplicate in laboratory

Formative Evaluation

- What is a Usability Problem?
  - – Unclear
    - The planned method for using the system is not readily understood or remembered (task, mechanism, visual)
  - – Error-prone
    - The design leads users to stray from the correct operation of the system (task, mechanism, visual)

Formative Evaluation .

- What is a Usability Problem?
  - Mechanism overhead
    - The mechanism design creates awkward work flow patterns that slow down or distract users
  - Environment clash
    - The design of the system does not fit well with the users' overall work processes (task, mechanism, visual)
      - i.e. Incomplete transaction cannot be saved

Formative vs Summative

https://www.youtube.com/watch?v=bTGnJnuVNt8

Methods

Qualitative Methods for Collecting Usability Problems

- Thinking aloud method and related alternatives:
  - Constructive interaction, coaching method, retrospective walkthrough
- Output: Notes on what users did and expressed:
  - Goals, confusions or misunderstandings, errors, reactions expressed
- Questionnaires
  - Focus groups, interviews

Observational Methods - Think Aloud

- User observed performing task
  - User asked to describe what he is doing and why, what he thinks is happening etc.
- Advantages
  - Simplicity - requires little expertise
  - Can provide useful insight
  - Can show how system is actually use
- Disadvantages
  - Subjective
  - Difficult to conduct
  - Act of describing may alter task performance

Observational Methods - Cooperative evaluation

- Variation on think aloud
- User collaborates in evaluation
- Both user and evaluator can ask each other questions throughout
- Additional advantages
  - Less constrained and easier to use
  - User is encouraged to criticize system
  - Clarification possible

Observational Methods

- Post task walkthrough
  - User reacts on action after the event
  - Used to fill in intention
- Advantages
  - Analyst has time to focus on relevant incidents
  - Avoid excessive interruption of task
- Disadvantages
  - Lack of freshness
  - May be post-hoc interpretation of events

Query Techniques - Interviews

- Analyst questions user on one to one basis
- Usually based on prepared questions
- Informal, subjective and relatively cheap
- Advantages
  - Can be varied to suit context
  - Issues can be explored more fully
  - Can elicit user views and identify unanticipated problems
- Disadvantages
  - Very subjective
  - Time consuming

Query Techniques - Questionnaires

- Set of fixed questions given to users
- Advantages
  - Quick and reaches large user group
  - Can be analyzed quantitatively
- Disadvantages
  - Less flexible
  - Less probing

SIR, MAY I HAVE FIVE MINUTES OF YOUR TIME?
SURE, I CAN DO 9:30 TO 9:35 NEXT TUESDAY...

Query Techniques - Questionnaires .

- Need careful design
  - What information is required?
  - How are answers to be analyzed?
- Should be pilot tested for usability!
- Styles of question
  - General
  - Open-ended
  - Scalar
  - Multi-choice
  - Ranked

☑ Excellent
□ Very good
□ Good
□ Average
□ Poor

Laboratory studies: Pros and Cons

- Advantages:
  - Specialist equipment available
  - Uninterrupted environment
- Disadvantages:
  - Lack of context
  - Difficult to observe several users cooperating
- Appropriate
  - If actual system location is dangerous or impractical for to allow controlled manipulation of use

Conducting A Usability Experiment

Main Steps

- The planning phase
- The execution phase
- Data collection techniques
- Data analysis

The Planning Phase

- Who, what, where, when and how much?
  - Who are test users, and how will they be recruited?
  - Who are the experimenters?
  - When, where, and how long will the test take?
  - What equipment/software is needed?
  - How much will the experiment cost?
  - Outline of test protocol

Outline of Test Protocol

- What tasks?
- Criteria for completion?
- User aids
- What will users be asked to do
  - i.e. Thinking aloud studies
- Interaction with experimenter
- What data will be collected?

Designing Test Tasks

- Tasks:
  - Are representative
  - Cover most important parts of UI
  - Don’t take too long to complete
  - Goal or result oriented (possibly with scenario)
- Tips:
  - First task should build confidence
  - Last task should create a sense of accomplishment

Detailed Test Protocol

- All materials to be given to users as part of the test, including detailed description of the tasks
- Deliverables from detailed test protocol
  - What test tasks? (written task sheets)
  - What user aids? (written manual)
  - What data collected? (include questionnaire)
  - How will results be analyzed/evaluated? (sample tables/charts)
- Then do a pilot with a few users

Pilot Studies

- A small trial run of the main study
  - Can identify majority of issues with interface design
- Pilot studies check:
  - That the evaluation plan is viable
  - You can conduct the procedure
  - That interview scripts, questionnaires, experiments, etc. work appropriately
- Iron out problems before doing the main study

Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

The Execution Phase

- Prepare environment, materials, software
- Introduction should include:
  - Purpose (evaluating software)
  - Voluntary and confidential
  - Explain all procedures
    - i.e. Recording, question-handling
  - Invite questions
- During experiment
  - Give user written task description(s), one at a time only one experimenter should talk
- De-briefing

Ethics of Human Experimentation

- Users feel exposed using unfamiliar tools and making errors
- Guidelines:
  - Re-assure that individual results not revealed
  - Re-assure that user can stop any time
  - Provide comfortable environment
  - Don't laugh or refer to users as subjects or guinea pigs
  - Don't volunteer help, but don't allow user to struggle too long
  - In de-briefing
    - Answer all questions
    - Reveal any deception
    - Thanks for helping

Data Collection

- Pad and paper the only absolutely necessary data collection tool!
- Observation areas (for other experimenters, developers, customer reps, etc.) - should be shown to users
- Videotape (may be overrated) - users must sign a release
- Video display capture
- Portable usability labs
- Usability kiosks

Data Analysis

- Before you start to do any statistics:
  - Look at data
  - Save original data
- Choice of statistical technique depends on
  - Type of data
  - Information required
- Type of data
  - Discrete - finite number of values
  - Continuous - any value

Statistics

- The mean time to perform a task (or mean no. of errors or other event type)
- Measures of variance – standard deviation
- For a normal distribution:
  - 1 standard deviation covers ~ 2/3 of the cases
  - In usability studies:
    - Expert time SD ~ 33% of mean
    - Novice time SD ~ 46% of mean
    - Error rate SD ~ 59% of mean

Statistics .

- Confidence intervals (the smaller the better)
  - The "true mean" is within N of the observed
  - Mean, with confidence level (probability) .95
- Since confidence interval gets smaller as the number of users grow:
  - How many test users required to get a given
  - Confidence interval and confidence level

Testing Usability in the Field

- Direct observation in actual use
  - Discover new uses
  - Take notes, don’t help, chat later
- Logging actual use
  - Objective, not intrusive
  - Great for identifying errors
  - Which features are/are not used
  - Privacy concerns
- Bulletin boards and user groups

Testing Usability in the Field .

- Questionnaires and interviews with real users
  - Ask users to recall critical incidents
  - Questionnaires must be short and easy to return
- Focus groups
  - 6-9 users
  - Skilled moderator with pre-planned script
  - Computer conferencing
  - Virtual environments
- On-line direct feedback mechanisms
  - Initiated by users
  - May signal change in user needs
  - Trust but verify

Field Studies: Pros and Cons

- Advantages:
  - Natural environment
  - Context retained (though observation may alter it)
  - Longitudinal studies possible
- Disadvantages:
  - Distractions
  - Noise
- Appropriate:
  - For beta testing
  - Where context is crucial for longitudinal studies

Choosing an Evaluation Method

- When in process
  - Design vs. implementation
- Style of evaluation
  - Laboratory vs. field
- How objective
  - Subjective vs. objective
- Type of measures
  - Qualitative vs. quantitative

Choosing an Evaluation Method .

- Level of information
  - High level vs. low level
- Level of interference
  - Obtrusive vs. unobtrusive
- Resources available
  - Time
  - Subjects
  - Equipment
  - Expertise

Subjects

- The choice of subjects is critical to the validity of the results of an experiment
  - Subjects group should be representative of the expected user population
- In selecting the subjects it is important to consider things such as their
  - Age group, education, skills, culture
  - How does the sample influence the results?
- Report the selection criteria and give relevant demographic information in your publication

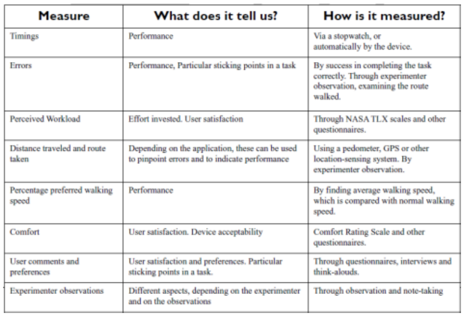
Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

Subjects .

- How many participants depends on how big is the effect you want to measure?
  - Large effects can be detected with smaller samples
    - i.e. Small n needed to discriminate speed between turtles and a rabbits
  - The more participants the “smoother” the data
  - Central Limit Theorem - as n increases (n>30) the sample mean approaches a normal distribution
  - Extreme data has less influence (e.g. one sleepy participants does not mess up the results that much)
- For quantitative analysis:
  - Min 15-20 or more per group/cell

Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

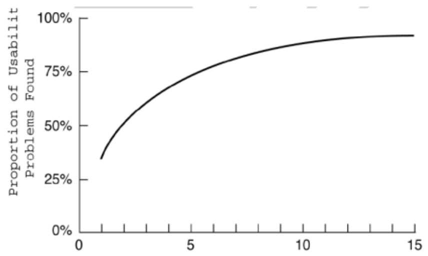
Experimental Measures



Measure	What does it tell us?	How is it measured?
Timings	Performance	Via a stopwatch, or automatically by the device.
Errors	Performance, Particular sticking points in a task	By success in completing the task correctly. Through experimenter observation, examining the route walked.
Perceived Workload	Effort invested. User satisfaction	Through NASA TLX scales and other questionnaires.
Distance traveled and route taken	Depending on the application, these can be used to pinpoint errors and to indicate performance	Using a pedometer, GPS or other location-sensing system. By experimenter observation.
Percentage preferred walking speed	Performance	By finding average walking speed, which is compared with normal walking speed.
Comfort	User satisfaction. Device acceptability	Comfort Rating Scale and other questionnaires.
User comments and preferences	User satisfaction and preferences. Particular sticking points in a task.	Through questionnaires, interviews and think-alouds.
Experimenter observations	Different aspects, depending on the experimenter and on the observations	Through observation and note-taking

Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

Evaluators & Problems



Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

Evaluate AR Apps

Why Evaluate AR Applications

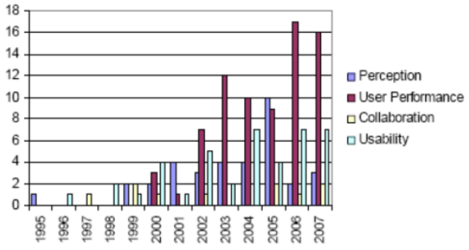
- To test and compare interfaces, new technologies, interaction techniques
- Test Usability
  - Learnability, efficiency, satisfaction,...
- Get user feedback
- Refine interface design
- Better understand your end users

Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

Types of User Studies in AR

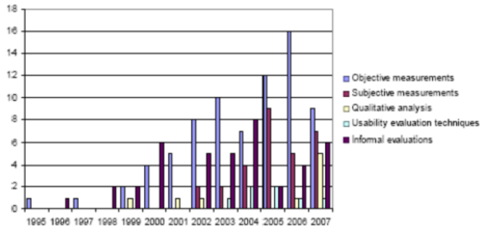
- Perception
- User Performance
- Collaboration
- Usability of Complete Systems
- Brain Analysis

Types of AR User Studies



Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

Types of Experimental Measures Used



Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

Typical Hardware

- Eye Tracking
- HMDs
- Physiological devices

Eye Tracking

- Head or desk mounted equipment tracks the position of the eye
- Eye movement reflects the amount of cognitive processing a display requires
- Measurements include
  - Fixations: eye maintains stable position. Number and duration indicate level of difficulty with display
  - Saccades: rapid eye movement from one point of interest to another
  - Scan paths: moving straight to a target with a short fixation at the target is optimal

Physiological Measurements

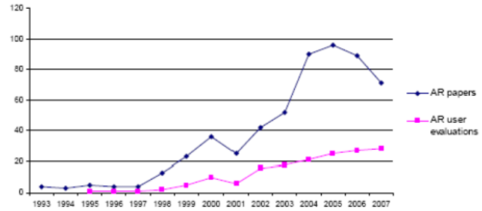
- Emotional response linked to physical changes
  - May help determine a user’s reaction to an interface
- Measurements include:
  - heart activity, including blood pressure, volume and pulse
  - activity of sweat glands: Galvanic Skin Response (GSR)
  - electrical activity in muscle: electromyogram (EMG)
  - electrical activity in brain: electroencephalogram (EEG)
- Some difficulty in interpreting these physiological responses
  - More research needed

Survey of AR Papers

- Edward Swan (2005)
  - Surveyed major conference/journals (1992-2004)
    - Presence, ISMAR, ISWC, IEEE VR
- Summary
  - 1104 total papers
  - 266 AR papers
  - 38 AR HCI papers (Interaction)
  - 21 AR user studies
- Only 21 from 266 AR papers had a formal user study
  - Less than 8% of all AR papers

Billinghurst, M. Evaluating AR Applications, HIT Lab NZ, University of Canterbury

AR User Evaluations



Perceptual Evaluation of Photo-Realism AR

https://www.youtube.com/watch?v=IrtUZKl9v34

User Experiences with AR Mobile Navigation

https://www.youtube.com/watch?v=qoOMDP2uHq0

Conclusions

- Very extensive field
- Not easy to select the best approach
- Biggest problems:
  - Understand the problem
  - Get a large sample
  - Analyse the data properly
- Still AR is not properly explored
  - Need for more research

Questions