Objective and Subjective Measurements

Kai Römer
kai.roemer@gmx.de

Three dimensional user interfaces can help users in time-critical situations to improve their ability to solve problems. But they also may distract them from fulfilling their tasks with efficiency and effectiveness. This requests to control and limit the distraction caused by usage of those interfaces. Therefore it is necessary to measure and compare distraction caused by three dimensional user interfaces in time-critical environments. In this document ways are provided to measure main causes of distraction. Further on mathematical methods to statistically evaluate the measurement results are presented.

Contents

1 Introduction

In time-time critical situations user interfaces can provide additional information in a fast and direct way. But they may also distract the user from efficiently and effectively fulfilling the time-critical main task, like driving a car, as they introduce a new, secondary task: the usage and interaction with the user interface. This leads to the necessity to measure the distraction and influence caused by the usage (secondary task) of the three dimensional user interface on the main task.

2 Types of Distraction

To find out about the influence of the usage of the three dimensional user interface on the main task the main reasons for distraction need to be defined. There are a lot of publications about measuring usability. They are mainly based on the ISO 9241-11 norm provided by the International Standardization Organization [1]. The ISO defines usability as:

> The usability of a product is the degree to which specific users can achieve specific goals in a particular environment with effectiveness, efficiency and satisfaction. [4]

This provides no useful hints about measuring usability for three dimensional user interfaces. But Tönnis et al provide the following types of distractions:

- Information Overload
- Change Blindness
- Perceptual Tunneling
- Cognitive Capture

For these kinds of distraction the following section will provide methods to measure and discover the impact of 3D user interfaces on the ability to fulfill a time-critical task.

3 Measurement of Distraction

3.1 Objective Measurement

Objective measurement tries to find out, how strong the distraction is measured by means of times, distances, failures and successes. The following subsections provide methods to gain values for those dimensions.

3.1.1 Task Time

The task time is a general indicator for distraction. As well the time needed for the secondary task $t_{secondary}$ as the time spent on the main task t_{main}can be measured. $t_{secondary}$ provides the time that is not available for the main task and leads to a decrease in situational awareness. t_{main} is interesting in ratio to $t_{secondary}$:

$$ratio = \frac{t_{secondary}}{t_{main}}$$

With theses values the influence of the distraction can be estimated.

3.1.2 Eye Tracking: Glance-Off Time

In order to receive reliable values for the time used for the secondary task, the glance-off time can be measured. The glance-off time is split into three parts as Tönnis describes in [5]:

- time to focus on the user interface
- time to interact with the user interface
- time to focus back on the main task

The glance-off time is used to find out, how long the user is out of the loop of the main task and is not aware of what is happening around him. The time to mentally process the information might not be measured by this method. This is because the user may still be thinking about the usage of the user interface after he focused back on the main task with his eyes. Therefore subjective measurement methods are useful. We will talk about them later.

3.1.3 Occlusion Test

The occlusion test is another eyes-based test. There are two versions of the test. One focuses on the interruptibility of the main task (figure 1), the second type focuses on the interruptibility of the secondary task, like the usage of the user interface. For 3D time-critical user interfaces the second type is of interest. With this test user interfaces can be evaluated for interruptibility of the usage of the interface, for resumability of the structure of the interface, and for time dependence. This kind of test should find out, how fast a user's perception of the user interface is, how long he needs to understand, what is going on in the interface and how long he needs to interact with the system. This test has a flaw. Participants might

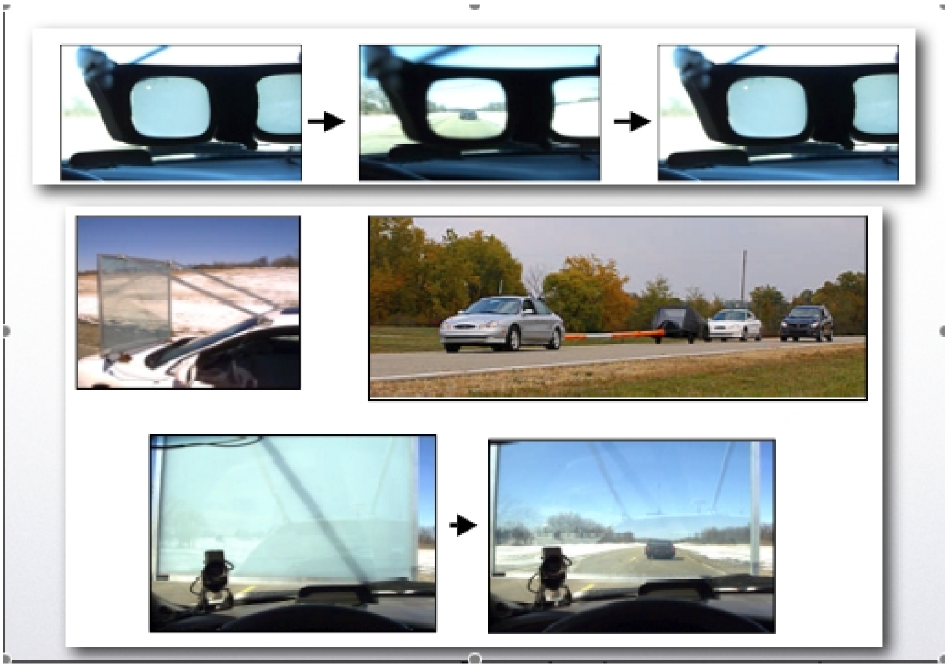

Figure 1: Application of Occlusion Test for Main Task: A drivers sight is occluded

tend to stay focused on the actual task of using the interface, while the shutter glasses are nontransparent. This means, that the results may be adulterated, as the time for focusing on the interface, for finding back into the display position and understanding, whats going on, are left out.

3.1.4 Quality of Main Task

Another group are tests, that measure the distraction of the user by aspects of the quality of the main task. These tests are normally used by the car industry.

- Speed Keeping Ability Test
  In this test a participant has to drive along a straight lane and interact with the user interface. The speed and the user interaction is recorded. Then the correlation between the usage of the user interface and changes in speed are evaluated. This test is an indicator for cognitive capture, as it indicates when the participant is no more able to realize the speed he is going and becomes mostly slower. This although indicates perceptual tunneling, as the participant is too focused on the interface that he leaves the speedometer out of view.

- Lane Keeping Ability Test
  This test is similar to the Speed Keeping Ability Test, but instead of the speed the variance in keeping of the lane is recorded.

- Lane Change Task Test
  The lane change task test is usually done inside a car driving simulator. The scenario is a three lane street with signs at the sides of the street that indicate on which lane the participant has to drive on. Have a look at figure 2, which shows a picture of the scenario of the Daimler-Chrysler lane change test. The results of this test are the differences between a reference trajectory and the recorded trajectory as shown in figure 3. With this test the impact of the usage of the interface on the parallel task execution

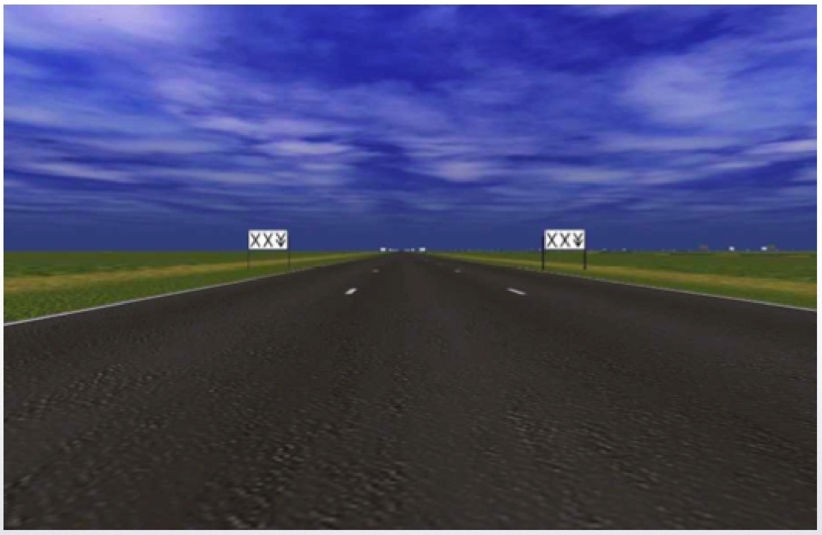

Figure 2: Lane Change Test Scenario by Daimler-Chrysler

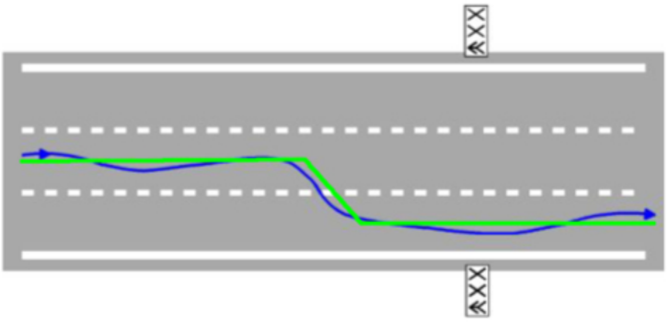

Figure 3: Trajectories of the car in a lane change test: blue: recorded trajectory; green:reference trajectory

ability of the participant can be measured. This also provides an indicator for stress and workload, while performing and additional task. The main disadvantage of this test is that it is unrealistic, as changing the lane due to signs does not happen very often in real world, and unnatural, as the test is performed inside a simulator. On the other hand this test provides a simple, quick and standardized way of evaluation. Simple because it can be done in a existing simulator, quick as no addition hardware needs to be constructed and standardized, as the test can be used for a lot of different kind of interfaces. This provides the ability to compare results between different types of interfaces.

- Steering Wheel Reversal Rate Test
  The last presented test in this row monitors the handling of the wheel. It checks if the participant has to do strong steering wheel corrections. These strong corrections can be found by a peak detection algorithm and can be counted. While concentrating on driving the user normally handles the wheel smooth, but tends to need more often stronger corrections of the steering wheel, when demands and workload increase. These steering wheel reversals usually occur after or while interacting with the user interface.

3.1.5 Peripheral Detection Task Test (PDT)

The usage of a 3D user interface often leads to a reduced field of view, as the user is too concentrated on interacting with the interface. How strong this decrease is, can be measured with Peripheral Detection Task (PDT). Therefore the participant has to execute the main task, like driving a car, a secondary task, like interacting with the user interface and an additional third task, which is to respond to the blinking of a LED that is positioned on a horizontal line from the center of the view to the outside. Figure 4 shows one possible test arrangement. As far as the participant realizes that one LED is blinking, he has to press the button.

Another way is to mount the LEDs on a horizontal line to a baseball cap. This has the advantage, that the LEDs are always in the same angle to the center of view of the participant. For the test the different LEDs blink randomly. Then it is recorded if the participant reacts to the LED (failure rate), how long does he need to react (reaction time) and how do those both values perform for different distances of the blinking LEDs to the center of the view.

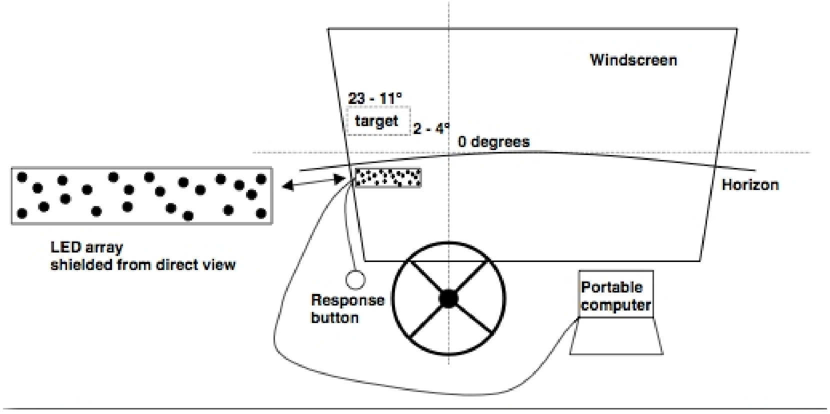

Figure 4: Peripheral Detection Task test arrangement

The error rate especially in correlation to the distance of the LED to the center of view is a good indicator for perceptual tunneling and change blindness, because if the user is tunneled on something he wont realize the blinking of the LED. The reaction time is an indicator for cognitive capture. Additional this test can be used to monitor the workload of the user. This is done by evaluation of the average reaction time and the failure rate. Both will increase for higher workload. This seems to be test with big potential, but it is not yet standardized.

3.1.6 Summary of Objective Tests

test name	gained values	indication
Task Time	t_{main}, $t_{secondary}$, $\frac{t_{secondary}}{t_{main}}$	length/ratio of distraction
Glance-Off Time	$t_{focusoninstrument}$, $t_{readinstrument}$, $t_{focusbackonmaintask}$	out of the loop effect
Occlusion Test	interruptibility, resumability, time dependence	speed of perception, understanding and transcription
QoMT Lane/Speed Keeping Ability	speed variance, driving lane variance	cognitive capture, perceptual tunneling
QoMT Lane Change Test	driving trajectory deflection	workload, stress level
QoMT Steering Wheel Reversal Rate Test	count of steering wheel reversals	workload
Peripheral Detection Task Test	error rate, reaction time	change blindness, perceptual tunneling, cognitive capture, workload

Table 1: Summary of objective measurement methods

Table 1 presents a summary of the gained values and results of the objective tests.

3.2 Subjective Measurement

Additional to the objective measurement methods there are the subjective methods. Subjective means asking the user about his personal opinion.

3.2.1 NASA-TLX

Mostly this is done by questionnaires, where NASA-TLX and SWAT are the most common ones. The NASA-TLX is a questionnaire developed by NASA. It provides an overall index for the workload while fulfilling a job. This test is applyable to 3D user interfaces in a time-critical environment [2]. It measures by questioning users for their mental demand, physical demand, temporal demand, performance, effort level, and frustration level and calculates a single value from these results. In a normal driving situation a person should have a value around 30 out of an interval ranging from 0 to 100. With heavier load by using an interface this index will increase.

3.2.2 SWAT

Another test is SWAT. It means Subjective Workload Assessment Techniques. It provides an indication for loads for time, mental effort and psychological stress and calculates an index for that.

3.2.3 Summary

The questionnaire is normally answered directly after the participation in a test scenario or a certain variant of an interface, where the user has to interact with the 3D user interface. These questionnaires are very easy to implement, as they exist as a predesigned table of questions and evaluation methods, quite often available as programs that automatically provide results to the test crew. This brings the advantage of low costs. One big advantage of after-test questionnaires is, that they do not influence the main task performance. But on the other hand they have disadvantages like it might be problematic for the user to understand the questions designed by a developer or he might misinterpret them. If there are several types of a user interface to be tested, the participant might tend to reduce the precision of his answers due to repetition.
Usually these kind of tests are designed to measure workload and information overload.

4 Designing Valid Tests

This section describes, what is necessary to design a valid test for evaluation for 3D user interfaces. The following list sums up the requirements for a valid design.

- At first a group of participants needs to be found. They should represent the targeted user group.
- It is necessary to define a structure for the test in order for the tests to be comparable.

- Then the gathered values need to be defined.
- The group of participants can be used for all scenarios, or every participant is just used for one scenario. When all participant are involved in all test, it is most common to divide them into groups and let group A begin with scenario 1 and group B begin with scenario 2 as [3] suggests.

This leads to a big amount of data, that now needs to be evaluated. This is discussed in the following section.

5 Evaluation of Measurement Results

After measuring there is a lot of data that needs to be analyzed. This means, that it has to be found out if one system is significantly better than another one.

5.1 Basic Evaluation

For the basic evaluation the mean value can be build with formula 1, which provides an average of the gathered data.

$$\overline{X} = \frac{\sum X}{N} \tag{1}$$

The mean absolute deviation is calculated with formula 2, which provides the mean deviation of the data from the mean value.

$$mad = \frac{\sum |X - \overline{X}|}{N} \tag{2}$$

The variance is calculated in formula 3.

$$s^2 = \frac{\sum (X - \overline{X})^2}{N - 1} \tag{3}$$

Standard deviation:

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{N - 1}} \tag{4}$$

5.2 Hypothesis Testing

The goal of hypothesis testing is find out if there is a significant difference between two or more groups of collected data. This is done by generating a hypothesis H_1 that says group A is significantly better than group B. But we can only test the null hypothesis H_0 which is the logical opposite of H_1. This means we have to prove:

$$H_0 : p(X|H_0)$$

This mean that we want to find the probability that we got our test data under the condition that our hypothesis H_1 is wrong. This can be done on different significance levels like: $\alpha <= 0.05$ or $\alpha <= 0.01$
To calculate this significance value there are several methods. Most common are t-test and ANOVA, which will be discussed here.

5.2.1 t-test

The t-test compares two groups for metrical attributes. This requires that the mean values of the groups are normally distributed. The t-test is done by comparing the distribution of the collected data to a t-distribution. This is done by the following formula:

$$t = \frac{1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} * \frac{\overline{y}_1 - \overline{y}_2}{s}$$

with n_1,n_2: sample count; $\overline{y}_1$,$\overline{y}_2$: mean value; s: mean variance
t is then compared to a precalculated data table. There the value for the wanted significance is taken and compared to t. If t is smaller than that value the hypothesis is accepted, otherwise refused.

5.2.2 ANOVA

ANOVA works similar to t-test. The main differences are, that it compares not only two groups but a number of groups greater or equal to two and uses the f-distribution to calculate the significance correlation. With ANOVA for each pair of groups a value is calculated and compared to each other.

$$F = \frac{n_1 n_2 (\bar{x}_1 - \bar{x}_2)^2}{(n_1 + n_2) var_g}$$

var_g: variance over all
Both test are implemented in many data processing tools like Excel or SPSS, which provide help at evaluating test data.

6 Summary

In this document several surveys of testing 3D user interfaces in time critical environments were introduced and shown what they can provide and which data they return. After that there was a short introduction to processing those data with t-test and ANOVA.

References

[1] *ISO norm 9241-11.*

[2] FAA. Faa human factors awareness web course. http://www.hf.faa.gov/webtraining.

[3] Bernard D. Adelstein J. Edward Swan II, Stephan R. Ellis. *IEEE Virtual Reality 2006, Tutorial Proceedings, Conduction Human-Subject Experiments with Virtual and Augmented Reality.* 2006.

[4] R E Renger. *NPL Report DICT 163/90*, chapter A preliminary design for a methodology for experimentally measuring usability. 1990.

[5] Laura K. Thompson, Christian Lange, and Marcus Tönnis. Driver Visual Behaviour while Interacting with Adaptive Cruise Control. In *50th Annual Meeting of the Human Factors and Ergonomics Society (HFES)*, October 2006.