The Difference Matters: Benchmarking Visual Performance of a Cognitive Pilot Model

Florian Frische, Andreas Lüdtke

OFFIS Institute for Computer Science, Escherweg 2,
26121 Oldenburg, Germany
{frische, luedtke}@offis.de

Abstract. In this paper we introduce an approach to objectively validate visual performance of a cognitive pilot model using benchmarks of human performance. A study with 16 human airline pilots and two competing models has been conducted in order to validate visual performance of the models applying these benchmarks. The study shows that human performance benchmarks can support analysts with a powerful and easy to use method for validation of human performance models. The benchmark is part of a larger-scale method, which will be developed in order to evaluate human factors issues of future HCI-concepts in early stages of system design.

Keywords. Human Performance Modelling, -Validation, -Analysis, Goodness-of-Fit Measures

1 Introduction

The European project HUMAN[1] aims at developing virtual pilots, in order to improve the human error analysis of new cockpit systems. The virtual pilots should allow simulation-based testing of new cockpit system designs in early design phases. This enables simulating a huge number of scenarios in accelerated time in order to identify potential problems in the operation of new systems (e.g. human errors due to *clumsy automation* [9]) and to derive necessary improvements. The virtual pilots in HUMAN are based on a cognitive architecture named CASCaS (Cognitive Architecture for Safety Critical Task Simulation) [6] which is an implementation of the information processing paradigm of human cognition (similar to ACT-R [1] and MIDAS [5]).

Pilots flying modern aircrafts are confronted with many information (e.g. weather conditions, flight plans, air traffic and automation modes) that are mainly presented on visual displays. Consequently, pilots' vision is the main cognitive resource (apart from auditive and haptic resources) for gathering information about the aircraft and the outside world. Thus, a sophisticated and thoroughly validated visual perception model is a pre-condition for simulating pilot-like interaction with visual interfaces.

[1] 7th Framework Programme, see http://www.human.aero for further information about the project

Evaluating the predictive power of visual performance models is very complex and time consuming because human visual performance (1) is very variable between different pilots as well as within a single pilot and (2) is a combination of different aspects, such as glance duration, glance frequency and scanpaths. Methods, techniques and tools are needed that can easily be applied by analysts to evaluate model performance.

In HUMAN we validated the performance of CASCaS by comparing data produced by the model with data produced by human pilots in experimental simulator studies, both flying the same scenarios. We developed a benchmark approach that is used in combination with traditional validation techniques. Benchmarks are commonly understood as an analysis method to objectively compare characteristics of subjects with characteristics of a reference subject. The measures considered for the analysis usually characterize the overall power of the system with regard to a target question [10]. Our benchmark approach provides a catalogue of objective measurement criteria allowing (1) to compare the fitness of competing models, (2) to find the best fitting model and (3) to decide objectively if the predictive power of a model is sufficient for the desired application.

In this paper we will briefly introduce our benchmark approach, which we call *Human Performance Benchmarking* and present results of two CASCaS versions that have been compared in order to find the best fitting model. The paper is organized as follows: In Section 2 we present related work in the area of human model validation. Next, in Section 3 our approach will be introduced. In Section 4 we present exemplary results of a first application of the Human Performance Benchmarking. Finally, in Section 5, we will summarize the paper and point out potential improvements of our approach, identified based on the study results. The Human Performance Benchmarks are part of a larger-scale method, on which we are currently working in order to evaluate human factors issues of future HCI-concepts in early stages of system design.

2 Related Work

Different measures, mainly of descriptive statistics, have been used in order to demonstrate the fit of model performance to human performance based on quantitative data (see e.g. [3], [7] and [2]). These measures can be categorized in three types:

1. Measures of central tendency
2. Measures of dispersion
3. Measures of association

The most common type of measure used in the area of model validation are measures of central tendency, e.g. mean, median and mode. Measures of central tendency describe the center or middle of a given data distribution by mapping multiple performance values on a single value. Thus, measures of central tendency are not sufficient to describe the variability in human performance.

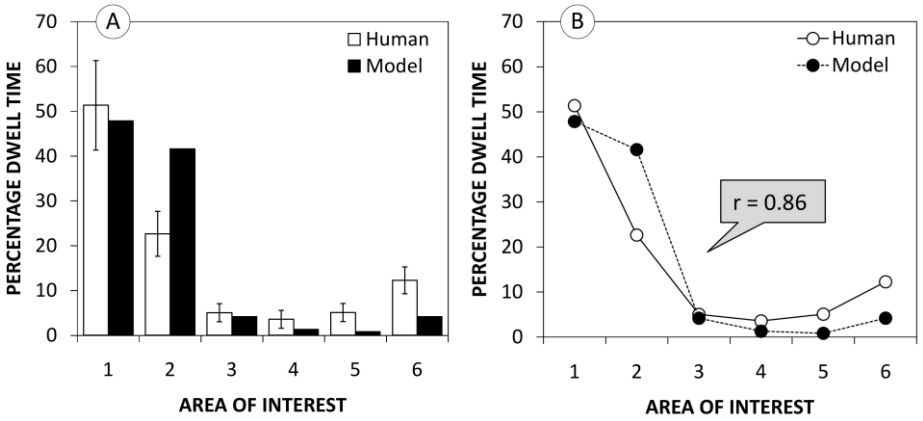

Fig. 1. Comparison of glance distribution of pilots in a flight simulator study conducted at German Aerospace Center (DLR). Chart A depitcts a common bar chart used to visualize local fitness including the mean and standard deviation for each area of interest. Chart B depicts the mean glance distribution as line chart (which is more intuitive for analysing trend consistency) including Pearson (r) as a quantification of trend fitness.

However, variability of pilot behavior has to be considered during cockpit design, because the design has to be safe and usable for different types of pilots. Measures of dispersion are used to describe the between-subjects variability and the within-subject variability. Typical measures of dispersion are range, standard deviation and confidence intervals. Finally, measures of association are used (1) to describe the relation of data points of a sample to data points of another sample, or (2) to describe the relation between parameters within a sample. Typical measures of association are Pearson's correlation coefficient (in the following referred to as *Pearson*) and Spearman's or Kendall's rank order correlation. All three types of measures are frequently used to validate models of human performance with different strength and weaknesses.

Measures of central tendency in combination with measures of dispersion are used to describe the local fit between human and model performance data, see e.g. [7] and [3]. Combining mean (as measure of central tendency) and standard deviation (as measure of dispersion) allows to quantify the average case and variability of human and model behavior based on given data samples. Fig. 1A shows an example for measuring distribution of gaze on areas of interest. The combination of mean and standard deviation provides a *qualification* of local fitness between datasets. Nevertheless, they do not *quantifiy* fitness. Quantification is especially needed if analysts have to validate different versions of a model in order to find the best fitting one.

Measures of association are used to quantify the relation between model data and human data by calculating the trend fitness between the two datasets. Fig. 1B shows an example for comparing the rank order of the pilots' gaze distribution. Here, Pearson is used to quantify the consistency of the rank order for the two data sets. Nevertheless, measures of association do not take into account the variability within human performance. Quantifications calculated based on measures of association provide only poor evaluation of the fitness between model and human data because

they do not consider behavioural variability. It is not possible to draw any conclusion about the final predictive power of the model. Thus, a measure is needed which on the one hand considers variability in human performance and on the other hand quantifies model performance in comparison to human performance.

First approaches to provide such a measure have been researched by Schunn and Wallach, and by Gluck. Schunn and Wallach provided a measure for quantification of local fitness in [8]. They recommend using the root mean square deviation (RMSD) as a measure of local fitness. RMSD calculates the difference between a set of related performance values and merges them on a single value of predictive power. Thus, RMSD allows comparing performance of different variants of a model on a quantitative basis. However, the problem analysts are faced with when using this approach for validation of local fitness is analog to the problem when using Pearson, that is the absence of taking into account human performance variability.

This problem has been tackled by Gluck et al. in [4]. They proposed an approach to validate a pilot model, taking into account the between-subject *differences* of a set of human performance datasets instead of absolute performance values. The between-subject differences served as a reference in order to actually evaluate the goodness-of-fit of a pilot model *m*. For a group of human subjects *S*, they calculated the difference in performance of each individual subject s_k to the average performance of $S \backslash s_k$ using RMSD. The result was a set of difference values describing the natural variability of human performance which has been used to evaluate the difference between *m* and *S*.

When analysing visual data, one of the characteristics most frequently described is the glance distribution on a set of pre-defined areas of interest. Here, RMSD for local fitness and Pearson for trend fitness can be used to calculate the difference between two subjects. Consequently, for a group of *n* human subjects, the result can be described as a bi-directional complete graph K_n (see Fig. 2) where *K* denotes the graph and *n* depicts the number of nodes. Edges $i \leftrightarrow j$ are labeled with a value v_{ij} representing the difference (RMSD or Pearson) between s_i and s_j. Thus, the graph completely describes the differences between human subjects (difference graph). The same approach can be used for measuring differences between human subjects and human performance models.

Human Performance Benchmarks use difference graphs as the basis in order to validate visual performance of human models tackling quantification of visual performance and taking behavioural variability into account.

3 Human Performance Benchmarking

Benchmarks are commonly understood as an analysis method objectively comparing characteristics of subjects with characteristics of a reference subject. The measures considered for the analysis usually characterize the overall power of the system with regard to a target question [10]. A benchmark is a quantitative test evaluating the characteristics observed and matching results on a scale which is independent from the characteristics. Thus, results of different characteristics can be aggregated and evaluated on different levels of abstraction. This allows analysts to identify strengths and weaknesses of systems objectively based on data.

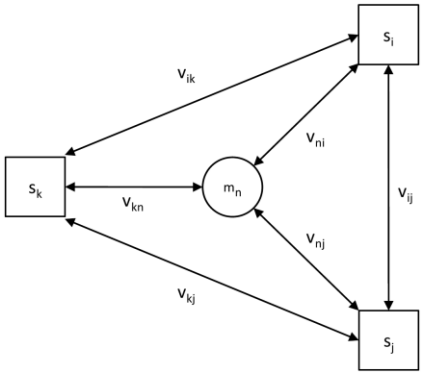

Fig. 2. Difference graph describing the individual performance differences of human subjects (square) to each other as well as the differences between human subjects and a model (circle), or between competing model versions.

The intention of our benchmark for visual performance is providing a goodness-of-fit evaluation for different characteristics C (such as glance distribution, glance frequency or scanpaths) for a set of competing models M with regard to a set of human subjects S. Our benchmark algorithm calculates an individual fitness value for each model $m \in M$. Fitness values can be used (1) to compare the fitness of competing models, (2) to find the best fitting model and (3) to decide objectively if the predictive power of a certain model is sufficient for the desired application. In the following, the algorithm used to calculate a model's fitness value is presented:

1. For all $c \in C$, calculate difference graph for m and S
2. For all $c \in C$, calculate confidence interval CI_S of individual difference values in S
3. For all $c \in C$, calculate the mean $mean_m$ of differences between m and S
4. For all $c \in C$, evaluate similarity between m and S
5. Calculate overall fitness value fit_m of m based on similarity evaluations of each c

The similarity evaluation in step 4 can be defined as a function $sim_c(m,S)$ evaluating for each c, if the performance of a given m is human-like, where human-like performance is determined by S. Our first version of this function is a naive boolean function returning *true* if $mean_m$ is covered by CI_S and *false* if $mean_m$ is not covered by CI_S:

$$sim_c(m,s) = \begin{cases} true \mid mean_m \in CI_S \\ false \mid else \end{cases} \quad (1)$$

In step 5, fit_m is calculated. We define fit_m as the sum of *true* similarity evaluations divided by the sum of all (*true* or *false*) similarity evaluations. Thus, the range of fit_m is [0;1], where 0 means no fitness and 1 means total fitness. Based on fit_m, analysts

can decide if a model's performance is sufficient for the desired application or not. A model is sufficient, if $fit_m \geq thres$, where *thres* is a pre-defined threshold parameter.
The application of this algorithm allows considering performance variability in terms of between-subject differences provided by difference graphs (step 1) which are used to calculate an interval CI_S which represents the variability (step 2). In addition, the algorithm allows quantification of fitness by applying $sim_c(m,S)$ (step 4) and calculating fit_m (step 5) as a measure of model fitness. In the following Section, we will present exemplary validation results of an application of Human Performance Benchmarks in order to preliminary assess the approach.

4 Results

Experiments with 16 human airline pilots have been conducted in order to collect reference data for the validation of the visual performance data of our pilot model.

Performance data of human pilots and two competing model versions (m_1 and m_2) have been validated based on data of glance distributions during three flight phases (cruise (1), approach (2) and final approach (3)), see Fig. 3A.
Accordingly, the benchmark has been used to quantify performance and to objectively evaluate fitness of m_1 and m_2 for three performance characteristics c_1 (glance distribution during cruise phase), c_2 (glance distribution during approach phase) and c_3 (glance distribution during final approach phase):

Step 1: We computed the corresponding difference graph for c_1, c_2 and c_3 taking into account the 16 subject pilots *S* and m_1 and m_2. Due to size limitations the graphs are not shown here.

Step 2: We computed CI_S for c_1, c_2 and c_3. The results are shown in Fig. 3B.

Step 3: We calculated $mean_{m1}$ and $mean_{m2}$. The results are shown in Fig. 3B.

Step 4: We evaluated the similarity between m_1 and *S*, and between m_2 and *S*. The results are shown in Table 1. According to the similarity function defined in Section 3, Fig. 3B shows that only m_1 receives a *true* evaluation ($sim_{c2}(m_1,S) = true$).

Step 5: Results of step 4 have been used to calculate the overall fitness value of m_1 and m_2. The results are shown in Table 1. For the datasets analysed in this study, the overall fitness value of model 1 is .33 and the evaluation of model 2 is 0. The fitness values can be used to evaluate if these models are sufficient for the desired application based on a comparison of each fitness value to *thres*. (e.g., if analysts define *thres* = .3, then m_1 would be sufficient).

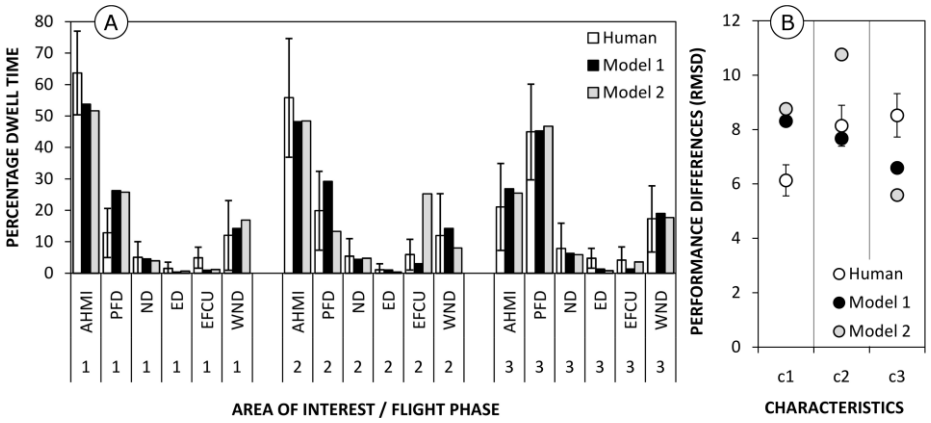

Fig. 3. Chart A depicts the glance distributions (AOIs: Advanced Human Machine Interface (AHMI), Primary Flight Display (PFD), Navigation Display (ND), Engine Display (ED), Electronic Flight Control Unit (EFCU) and Window (WND)) of a human population (mean and standard deviation) and two concurring models for three flight phases (cruise (C), approach (A) and final approach (F). Chart B depicts the mean values and confidence intervals (95\%) of human pilots and the mean values of model 1 and model 2, which have been calculated based on the difference graphs for the flight phases.

Table 1. Results of benchmarking reveal that m_1 receives a higher overall fitness value than m_2

Model	Cruise	Approach	Final Approach	Overall Fitness Value
m_1	false	true	false	.33
m_2	false	false	false	.0

5 Summary and Future Work

In this paper, we presented an approach (called *Human Performance Benchmarking*) for quantitative evaluation of goodness-of-fit for human performance models based on quantitative performance data. The approach considers variability in human performance and uses these performance differences as a model evaluation reference. A first naive evaluation function has been presented, which evaluates the model fitness in context of these reference data. The benchmarking approach has been applied within a study with aircraft pilots focusing on gaze data. The results have been presented and issues requiring improvements have been identified. The first issue is that the evaluation function is very simple and results may be misleading for models that perform just within the human performance intervals. Evaluation of these models would be similar to those that perform much better and would considerably differ from models that perform slightly worse. We plan to improve the evaluation function by weighting the distance to the center of the distribution. The second issue is that the function works fine for deterministic models (same input - same output) but models like CASCaS produce performance variance as well. We have not yet taken

variability in model performance into account. We plan solving this issue by considering CI_m instead of $mean_m$ for a set of model runs and to compare CI_m to CI_S.

Acknowledgments

The work described in this paper is funded by the European Commission in the 7th Framework Programme, Transportation under the number FP7 - 211988.

References

1. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111, 1036-1060 (2004)
2. Ball, J.T., Gluck, K.A., Krusmark, M.A., Rodgers, S.M.: 2003: Comparing three variants of a computational process model of basic aircraft maneuvering. In: Institute for Simulation and Training. pp. 87-98 (2003)
3. Frische, F., Osterloh, J.P., Luedtke, A.: Modelling and validating pilot visual attention allocation during the interaction with an advanced ight management system. In: Human Modelling in Assisted Transportation: Models, Tools and Risk Methods. pp. 165-172. Springer (2011)
4. Gluck, K.A., Ball, J.T., Krusmark, M.A., Rodgers, S.M., Purtee, M.D.: A computational process model of basic aircraft maneuvering. In: In Proceedings of International Conference on Cognitive Modelling (ICCM). pp. 117-122. Universitaets-Verlag Bamberg (2003)
5. Gore, B.F., Corker, K.M.: Increasing aviation safety using human performance modeling tools: An air man-machine design and analysis system application. In M. J. Chinni (Ed 2002, 183-188 (2002)
6. Lüdtke, A., Osterloh, J.P., Mioch, T., Rister, F., Looije, R.: Cognitive modelling of pilot errors and error recovery in ight management tasks. In: HESSD. pp. 54-67 (2009)
7. Salvucci, D.D.: Modeling driver behavior in a cognitive architecture. Human Factors 48 (2006)
8. Schunn, C.D., Wallach, D.: Evaluating goodness-of-fit in comparison of models to data (2005)
9. Wiener, E.L.: Human factors of advanced technology (glass cockpit) transport aircraft. Tech. rep., NASA Ames Research Center (1989)
10. Zhang, X.: Application-specific benchmarking. Tech. rep., Harvard University (2001)