Pick-by-Vision: A First Stress Test

Björn Schwerdtfeger* Rupert Reif† Willibald A. Günthner‡ Gudrun Klinker§
Technische Universität München

Daniel Hamacher¶ Lutz Schega‖ Irina Böckelmann**
Otto-von-Guericke Universität Magdeburg

Fabian Doil†† Johannes Tümler‡‡
Volkswagen AG

ABSTRACT

In this paper we report on our ongoing studies around the application of Augmented Reality methods to support the order picking process of logistics applications. Order picking is the gathering of goods out of a prepared range of items following some customer orders. We named the visual support of this order picking process using Head-mounted Displays "Pick-by-Vision". This work presents the case study of bringing our previously developed Pick-by-Vision system from the lab to an experimental factory hall to evaluate it under more realistic conditions. This includes the execution of two user studies. In the first one we compared our Pick-by-Vision system with and without tracking to picking using a paper list to check picking performance and quality in general. In a second test we had subjects using the Pick-by-Vision system continuously for two hours to gain in-depth insight into the longer use of our system, checking user strain besides the general performance. Furthermore, we report on the general obstacles of trying to use HMD-based AR in an industrial setup and discuss our observations of user behaviour.

Index Terms: H.5.1 [INFORMATION INTERFACES AND PRESENTATION]: Multimedia Information Systems—Artificial, augmented, and virtual realities; Evaluation/methodology; H.5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces —User-centered design;

1 INTRODUCTION

The basic conditions in the field of logistics have changed rapidly over the last years, with the market demanding customized products. As an example, 20 years ago automotive manufacturers offered three model series, while nowadays they are offering nearly ten. This is attended by an increase of the variants within one series. Thus, production and logistics systems have to become *supra-adaptive* [8].This requires that production processes as well as supporting IT systems be designed in a manner that enables workers to quickly handle new working conditions and environments. Thereby exists the need to make workers improve under such working conditions, without increasing their stress level, and while preventing them from making errors. This requires systems that support the worker with just the right information at exactly the right time. Such supporting systems have to provide detailed working instructions which have to be presented in a highly intuitive and precise way. As a result, workers can then start executing arbitrary jobs efficiently and error-free – and with minimal prior training. Augmented Reality (AR) is adjudged to have the potential to provide this very functionally.

Our use case in this context is the order picking process of logistics applications.To this end, we investigate the application of Head-mounted Display (HMD)-based visualizations (both AR and non-AR) to support the order picking process. We named this HMD-based support of the order picking process "Pick-by-Vision".

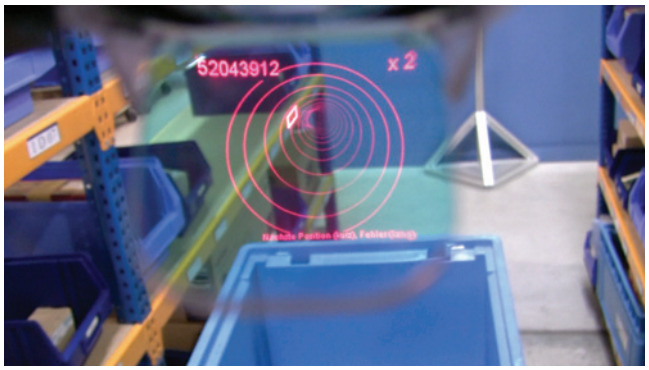

Figure 1: A photograph through the HMD of the Pick-by-Vision (AR) system (compare fig. 4). The visualization shows a Tunnel twisting to the left, with a square Frame at the end highlighting the label under the box in a warehouse from which a worker has to pick items. The article number and the amount to pick are shown at the top of the display.

1.1 The Problem of Executing AR Evaluations

There exist several obstacles when trying to apply AR to real life scenarios. In fact, most AR systems remain laboratory prototypes, mainly due to the lack of adequate hardware, but also due to not yet resolved usability issues [14]. Furthermore, we have to determine where the user can benefit from AR, while at the same time not knowing how good our AR user interface is. Thus, we must verify usability to determine if the system is effective [14], which is tough as we develop user interfaces at the limit of what is known or common practise in our field.

This goes along with the problem, that the utility of an AR application heavily depends on how good it fulfills the user's needs.

*schwerdt@in.tum.de
†Reif@fml.mw.tum.de
‡kontakt@fml.mw.tum.de
§klinker@in.tum.de
¶Daniel.Hamacher@gmx.net
‖lutz.schega@med.ovgu.de
**irina.boeckelmann@med.ovgu.de
††fabian.doil@volkswagen.de
‡‡johannes.tuemler@volkswagen.de

IEEE International Symposium on Mixed and Augmented Reality 2009
Science and Technology Proceedings
19 -22 October, Orlando, Florida, USA
978-1-4244-5419-8/09/$25.00 ©2009 IEEE

The problem is, to be able to elicit such user needs from future users, we need high-fidelity AR prototypes [1]. Those mature AR-systems help elicit much more realistic user needs than low-fidelity prototypes.

Due to those facts it is challenging to develop and evaluate such new AR user interfaces and applications. The 2D information visualization community has similar problems and thus promotes the report on long-term usage and field studies in natural settings and the investigation of new evaluation methods [18].

Furthermore, the execution of experiments in an industrial environment under realistic conditions with real tasks is complex, time consuming, and expensive, especially when incorporating real factory workers. It is still an open question how to execute such user studies in the field of AR. Dünser et al [5] state that, until now, it is partially unclear how to analyse the different user-related questions. In addition to that, the funding of such experiments usually comes from industry partners that expect results in a meaningful and summative manner. But the development of nowaday's AR systems has not yet matured enough so that today no well-developed and perfect systems are available. This means that we have to compare our not yet finished AR prototypes with established technologies in comparative experiments.

In summary, we address the typical situation of an AR researcher: a) we have a somehow mature AR prototype but b) it still has unresolved usability problems, c) it does not yet fulfill all of the important user's needs, d) we have to show the benefits of our AR system over an established technology in comparative experiments to give incentives for further investment, e) and it is expensive to solve all these problems one after each other.

1.2 Contribution

In this paper we contribute insights about the various obstacles of bringing our mobile Augmented Reality system from the lab to a realistic industrial environment. In particular we adress the midterm work with head-mounted AR systems. Moreover, we report on two tests comparing our Pick-by-Vision system with traditional technologies. In the first test, we compared two different configurations of our Pick-by-Vision system with a paper based list for the first time in a realistic industrial scenario. In the second test, we aimed at analyzing strain of the users when working with our system over a longer period of time. Basically we tried to reproduce the results from our previous tests executed in a completely different setup (different warehouse, different AR system, see [28]). To this end, we let subjects work for two hours with our Pick-by-Vision system and compared this with two hours working with a paper-based system. At the same time, we give a case study of how to "carefully" drive a comparative experiment with a not yet finished prototype.

A side effect of this evaluation was the development of a well-engineered Pick-by-Vision visualization[1], which we fully present in [26]. However, this visualization alone is not enough to support the picking process, the open problems of the system itself will be adressed in this paper.

2 Order Picking in Logistics Applications

Before we start explaining our evaluations, it is necessary to understand the environment in which our actions take place.

2.1 Order Picking and State-of-the-Art

Order picking is the gathering of goods out of a prepared range of items following customer orders [29]. In automotive industry, order picking is used to collect parts that are required for one customer's specific car. After picking these items into a basket it will be brought to the according car and all parts from the basket will be mounted to the car.

[1] and probably the most outworn HMD

Errors have a strong influence on the quality of delivery and the relationship between clients and suppliers. Thus, zero-defect picking is one major goal. Flexibility and fine motor skills of human beings are needed because the product range, and thus the variety of items, steadily increases while, by contrast, the size of orders is decreasing. Thus, complete process automation (e.g. the use of robots) is not the appropriate answer and humans are often the best solution for order picking [7].

Figure 2: differrent common order picking technologies: MDT with scanner (1), Pick-by-Light (2) and Pick-by-Voice (3)

Conventionally, workers execute their orders with paper lists which are intuitive for human beings but laborious to handle. In modern warehouses, worker support based on usual paper lists is often replaced by Mobile Data Terminals (MDT), Pick-by-Voice (PbV) or Pick-by-Light (PbL) systems [7] (fig. 2), .

All these technologies have specific advantages as well as disadvantages. PbV supports the worker by providing all instructions through the computer's speech output. Unfortunately, these systems face difficulties in noisy industrial environments. Furthermore, it is questionable whether the order picking worker likes being bossed by a monotone voice the whole day. Compared to voice support systems, PbL offers visual aid for the worker by installing small lamps on each storage compartment. PbL systems have the problem that the displays have to be elaborately integrated into the shelf construction and thus are very expensive and inflexible. PbL is most suitable for order picking stations with a high throughput. Brynzer et. al.[3] state the a better information system can save a lot of time in the order picking process.

Logistic Figures The quality of a Pick-by-Vision system is mainly measured according to the time and error rate.

The order picking time is divided into four interleaved tasks. The *base time* includes all tasks at the beginning and the end of one order (e.g. login at the system, pick-up and delivery of the collecting unit or the paper list). The *way time* is the time the user physically moves through the storage area, and the *picking time* consists of actually grabbing the item with the following delivery in the collecting unit or on a conveyor. Finally, the *dead time* includes the search for information and the human information processing and all process steps that are not necessary for the real task (e.g. open the packages). Comparing the order picking time with the

picked items, the order picking performance can be calculated in order lines per hour.

Picking errors can be divided according to their causes: a) wrong amount, b) wrong item / article, c) missing article / missing order line (as one line can consist of several articles) d) damaged article. The amount of errors is referred to in terms of the whole amount of picked items. This is called the order picking rate. Among other things, the error rate depends on the order picking technology. For a paper list it is normally 0.35%, for Pick-by-Light 0.40% or Pick-by-Voice 0.08% [27]. Even one error within 1,000 picked parts is not acceptable because, for example, in automotive industry each mistake can lead to halting the production line.

2.2 Pick-by-Vision - Definition

Using the analogy of the state-of-the-art systems, we name all systems using HMDs to support the order picking process *Pick-by-Vision* systems, as they primarily provide information via the visual sense. However, we have to further subdivide this definition, as we have systems making use of tracking technologies and those which do not. The systems which do not make use of tracking technologies to estimate the users position mainly present textual information in the form of a list of items or images, etc. Due to this, we will name such systems *Pick-by-Vision (2D)* systems. We call other systems, which use tracking and make explicit use of AR, *Pick-by-Vision (AR)*.

3 Related Work

We investigated the use of AR in industrial environments using the example of order picking (paper list and a simple Pick-by-Vision (AR) system) [28]. This evaluations focused on user-related issues in long-term use, mainly measuring strain by analyzing "heart rate variability" (HRV) during a two hour work phase. Since the heartbeat sequence is subject to a variability that can represent current strain of a person, the analysis of HRV can be used to analyze workplaces in the context of ergonomic examinations [22]. This has already been proven in the field of AR [17]. Next to analysis of HRV, the EZ-Scale [15] and discomfort questionnaire were used to examine subjective user strain. As a result, the overall strain did not differ significantly between paper list and Pick-by-Vision (AR). Nevertheless, a higher strain for uses' eyes was generated by the AR system.

In another study we evaluated the use of Pick-by-Vision (2D) systems in industrial environments [19]. In short-term studies the subjects performed better than with other conventional order picking technologies and they had not more physical strain.

To support order picking by appropriate visualizations we iteratively developed and improved visualisations for Pick-by-Vision (AR) systems [23, 25, 26].

Brau et al. did a long-term study, testing the general use of HMDs in an industrial setup (40 hour workweek) using the example of order picking [2]. They used the Microvision Nomad HMD (compare fig. 4) and provided textual order picking information, or what we call Pick-by-Vison (2D). The study had to be aborted because the users reported heavy headaches and eye strain. To find out about reasons for these problems a second study was started by Brau and Fritzsche [12, 6]. In their study they had three different setups: a) a Pick-by-Vision System, b) a paper-based list system, and c) a paper-based list system with subjects wearing a switched-off HMD. The experiment took place under HMD-friendly conditions, with controlled artificial illumination. The result was that the test subjects mainly complained about wearing an HMD in general, resulting in headaches due to the weight they had to carry on their head. Physiological problems for the eyes could not be found, but 20% complained about having to read from the HMD. Therefore it was proposed that each potential HMD user should do a suitability test before actually working with the system.

4 The Pick-by-Vision Systems and Setup

This section explains the setup we used for our evaluations. We start by illustrating the warehouse and the articles we used. Afterwards we present the different Pick-by-Vision Systems. Finally, we discuss the adjustment of the HMD's focal plane.

Figure 3: The warehouse with 4 shelves. The test subject wears a HMD with attached tracking markers. The WiFi-connected wireless PC is carried in a small backpack. As system control, we mounted a game show-like buzzer (adjusting knob) on the subject's belt. The tracking is done via 6 ART DTrack1/2 Cameras.

4.1 The Experimental Warehouse

Our warehouse (Fig. 3(4)) consists of two aisles, each 3.4 meters long and 1 meter wide. To the left and right of each aisle we placed shelves consisting of 5 layers (distance 0.4 meters) and 14 columns (distance 0.22 meters). Altogether we had 280 stock locations, which were filled up to 98%. The highest layer was at 1.8 meters and the bottom-most at 0.2 meters, forcing the users to make large body movements.

We used a standard and optimized convention of naming the item location with a number for the shelf, a character for the layer and a number for the column. This resulted, for example, in names like "3 A 13" (3rd shelf, first layer, 13th column). The warehouse was filled up - i.e. *chaotic* - instead of placing related articles close to each other. The assortment itself ranged from little drug boxes to heavy bolts requiring two-handed interaction, as well as paper ware (boxed and unboxed).

The base - the place in an order picking system for order receipt and delivery - consisted of a table approximately three meters from the storage area. The collecting units (boxes) were stored to the right side of the base.

The Picking Process At the beginning of an order, the subject takes a picking list from the stack Fig.3(2)) and takes one collecting unit from the station Fig.3(3)) and puts it on a picking trolley (Fig.3(1)) with four steerable wheels and dimensions of 400 x 600 mm. In the case of a worker using a Pick-by-Vision system, the step of picking up the list is dropped from the process.

During the actual picking process, the worker moves the trolley (Fig.3(1)) through the warehouse. The worker handles the order lines from his list sequentially. Each line contains information about the location of the item, the amount to pick and the article number. Before picking the item, the subjects have to make sure that they pick the right item by comparing the article numbers on the list with those on the label of the item. This control process has to be done for each of the different picking technologies.

To finish an order, the subjects deliver the collecting unit to the station (Fig.3(5)). In the case of pick by paper lists, the list has to be signed and must be put in to the collection unit. If subjects had discovered picking errors, they had to be noted on the list, while they had to be entered in the Pick-by-Vision system when it was used. All delivered articles were controlled and presorted (for the refill of the warehouse) at station 6 (Fig.3(6)).

4.2 The Pick-by-Vison Systems.

The Pick-by-Vision equipment, which was used for both Pick-by-Vision systems AR and 2D, can be seen in Fig.4. Even though our prototype seems quite large, the subjects did not complain about the system weight or dimensions during the tests. The system consists of the Microvision Nomad ND2000 HMD, with an attached A.R.T. marker target, which was aligned over the head of the user. For one thing, this prevents the user from sticking with the target in the shelf, and that also allowed the user to be tracked even when bending over. The Pick-by-Vision software itself runs at a 13-inch, 2kg tablet PC carried in a small backpack connected to the tracking server via wifi. The control unit of the Nomad display is mounted at the backpacks' belt. On top of the unit we mounted a click-turn-wheel adjusting knob, which is the only input device to the system, allowing for four inputs: turn (left/right), short click and long click. The user mainly has to enter the following system options: requesting an order, getting the next/last order line, annotating an error (including its kind) and terminating an order.

Figure 4: The Augmented Reality Equipment.

4.2.1 Pick-by-Vision (AR)

In several iterations, we have continuously developed and enhanced our Pick-by-Vision (AR) visualization, which is shown in Fig.1. The whole history of this evaluation is described in[26]. We have to explain some steps of this history as the iterative improvements were part of the experiments described in Sec. 5 and Sec. 6.

The picking process consists of two navigation phases. Both have to be somehow supported by the system:

- In phase A, the *coarse navigation*, the worker has to find the way to the right shelf.
- In phase B, the *fine navigation*, the worker has to find the specific box (to pick from) on this shelf.

Phase A is not the critical task. Most logistics experts agree that the number of shelves is generally manageable. Due to that and the fact that our warehouse only consisted of two aisles, we provide this information only textually (e.g.: "Go to aisle 1")

Much more critical is phase B, the fine navigation to one of the large number of boxes to pick from. This visualization consists of two elements, shown in Fig.1. The first element is the square highlighting the label under the box. We decided not to highlight the box itself as boxes can vary in size. However, it is not enough to just highlight the box/label with an augmentation. Due to the small field of view of current optical see-through HMDs (somewhere between 15 to 40 degrees) such an augmentation only rarely appears on the display. More often, it is outside the field of view because users are not looking in the direction of the target box. We then have to extend the actual visualization by a meta visualization to guide the users' gaze toward the box, thereby covering the entire range of 4π steradians. To this end we developed a visual tunnel, which can be best imagined like a hose of a vacuum cleaner starting a few centimeters in front of the eye and ending at the box.

The maturity level of our Pick-by-Vision (AR) system strongly depending on the maturity of the Tunnel visualization. The Pick-by-Vision (AR) 1.0 system, which was used in the first evaluation (Sec. 5), performed poorly when the user was turned away from the box by more than 80 degres. The Pick-by-Vision (AR) 2.0 System, which was used in the second evaluation (Sec. 6), performed good until about 120 degrees. Our current version, 3.0, performs well over the entire range of 4π steradians.

4.2.2 Pick-by-Vision (2D)

The Pick-by-Vision (2D) system is the same system as the Pick-by-Vision (AR) system, except that we display the location to go as text instead of making use of augmented reality. Due to this the user is able to find the right box by following the labels on the shelves.

4.3 Adjusting the Focus of the HMD

The problem of the accommodation discrepancy for the use of optical see-through displays (OST) (like the Nomad HMD) for displaying Augmented Reality information is widely known [13, 11, 10]. The user of an HMD can only focus on the virtual image of the display and the real world simultaneously when both are at the same distance [4], which is rarely the case. Rolland et. al. [20] stated that it can be shown, for example, that rendered depth errors are minimized when the virtual image plane is located in the average plane of the 3-D virtual object visualized. As a solution to the various conflicts in accommodation, Rolland et. al. [21] suggest to allow autofocus of the virtual image plane as a function of the location of the user's gaze point in the virtual environment, or to implement multifocal planes. A frequent change of accommodation produces fatigue, but placing the virtual focus in typical handling distances of 0.6 meters forces a continuous contraction of the ciliary muscle [16].

So what are the consequences for our application? Since currently none of the proposed autofocus / multifocal plane HMDs are available, we have to deal with the single feature of our Nomad

HMD allowing to manually adjust the virtual image plane between 0.3 meters to infinity. The users of our order picking system have to regard two types of visualizations: 1) the 3D augmentation and 2) 2D textual information. The 3D augmentation consists of the tunnel and square in front of the box and can stretch over a focus area of about 0.4-3.5 meters. The user needs to read text from the HMD in different situations, but the most relevant situation seems to be when the user holds the picked item in his / her hand, at a distance of about 0.6m. At that point the user has to compare the article number shown on the HMD with the article number printed on the article itself. This requires a multiple change of focus (between HMD and reality) in a short time periode, as the number is mostly quite long. Finally, we put the focal distance of the HMD in the middle of the discussed values, somewhere between 1.0 and 1.5 meters, as the Nomad does not allow an exact adjustment.

5 Evaluation 1

Up to that point of this evaluation, we could only evaluate our Pick-by-Vision (2D) system in an industrial setup [19]. So far our Pick-by-Vision (AR) system was just tested in the lab [23, 25]. Having the chance of equipping a real warehouse for a limited timeframe with tracking cameras, we brought our Pick-by-Vision (AR) system to the warehouse, to be able to compare it in an experiment with the other technologies. Due to the limited timeframe, we had to find an efficient way to adapt the visualization of our Pick-by-Vision (AR) system to the real warehouse. To this end we did an informal pre-evaluation with 8 participants described in [26]. This helped us remove the most important usability problems and to find good initial parameters for our visualization.

In the following we describe the experiment in which we compare our Pick-by-Vision (2D), our Pick-by-Vision (AR) (at version 1.0 and, as a reference, a common paper-based system.

5.1 Experimental Setup and Hypotheses

The only independent variable of this experiment was the picking system used at the three levels described above. The main dependent variables we discuss here were picking performance (time), errors and strain, as a subjective measurement. To analyze the latter variable (strain), we used the common NASA Task Load Index (TLX) test [9].

For our testing environment, we used the warehouse described in Sec. 4.1. The experimental design was of type with-in subject. Which means that each subject had to carry out the test with each of the three technologies, whereby the start sequence of the technologies was permuted. The subjects had to fulfill six orders with all in all 30 order lines (5.0 order lines/order) and 61 items (2.03 items/order line) for each technology. To reduce learning effects the orders were different between the three picking technologies. Within one order the order lines are different but the whole amount of items, their weight, and the ranges to go and to pick were the same. To compensate for learning effects, in particular when using an HMD for the first time, we followed the important approach of letting people try-and-ask, as long as they really felt comfortable and (we thought) they fully understand the system [25].

We set up the following hypotheses, all while keeping in mind that our systems are far from being perfect.

In previous tests we improved our Pick-by-Vision (AR) System to support an efficient and error free-picking [24] , and for the Pick-by-Vision (2D) system we knew that it performs better than a paper based system [19], so we set up the following hypothesis:

> **E1_H1:** a) Picking with a Pick-by-Vision (AR/2D) system produces less errors than picking with a paper list, and b) Picking with the Pick-by-Vision (AR) system produces less errors as picking with Pick-by-Vision (2D).

> **E1_H2:** a) Picking with Pick-by-Vision (AR/2D) is faster than picking with a paper list, and b) Picking with Pick-by-Vision (AR) is faster than picking with Pick-by-Vision(2D).

Besides improving the logistic performance figures, one design goal is to not produce stress using the Pick-by-Vision systems. Nevertheless, especially because of the weight of hardware and the mentioned problem of switching between real and virtual focal planes, we assume that using the Pick-by-Vision systems results in larger strain than working with a paper list. So our third hypothesis is:

> **E1_H3:** The NASA TLX shows a lower strain for working with the paper list compared to the Pick-by-Vision systems.

Beside those metrics used in most 3D user interface evaluations, we explored some other aspects by using questionnaires after the test. We had conspicuousness between the general experience with 3D user interfaces and the performance in previous AR tests [23], and therefore asked the subjects about their experiences with 3D. Furthermore, we used three positive questions (Likert scale) to gain insight about the general comfort when wearing the Pick-by-Vision system, as well as whether or not the subjects felt constrained by the HMD or the visualization.

5.2 Results

We had 19 subjects (16 male/3 female) between the ages of 18 and 45 (mean age: 27.2, std dev: 6.78). The subjects were mainly individuals from all areas of the university, friends, as well as three professional order picking workers. As we said at the beginning, we compare our not yet perfect Pick-by-Vision System (here at Version 1.0) with an established technology. For that reason we expect to find several unlucky factors, which lead to a general bad performance of our Pick-by-Vision systems. By an in-depth observation of the experiment, we have to equalize the results, as shown in the following.

Error Rates On the first view, according to the error rate, subjects performed significantly worse using both Pick-by-Vision systems as compared to the paper-based order picking. However, most errors can be traced back to a bad system configuration. Sometimes subjects accidently skipped over an order line and had no possibility to go back in the system. Hence, errors of this category recognized by the subjects were subtracted. However, two missing order lines within Pick-by-Vision (AR) were not recognized by the subjects. After this correction, the most common error was the picking of the wrong amount. Neither the paper list nor either of the Pick-by-Vision systems had a function to avoid this error. Furthermore, a wrong item was picked using the paper list and also using the Pick-by-Vision (2D) system, but no wrong items were picked using the Pick-by-Vision (AR) system. Finally, Pick-by-Vision (AR) has, in average, the lowest error rate (0.7%), followed by Pick-by-Vision (2D) (1.23%) and then paper list (1.4%). This result could not be shown to be significant using the Friedman Test[2] ($\alpha = 5\%$) and thus **E1_H1 a/b** cannot be proved. Figure 5 and 6 show the error rates of the three systems.

Time The picking time performance for the Pick-by-Vision systems seems to be about 10% better than that of the paper list, as shown in Fig.7. However, we could not find significant differences (ANOVA, $\alpha = 5\%$) Thus **E1_H2 a/b** could not be proven. Furthermore, we observed that the distributions of the Pick-by-Vision systems are skewed left because 12 of the 19 subjects were faster

[2]We use the Friedman Test instead of ANOVA, as the sample is not normally distributed.

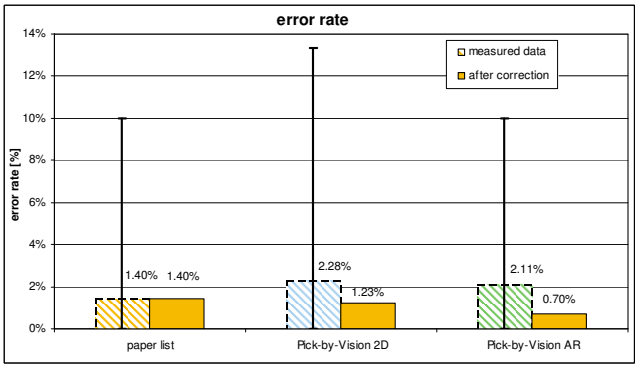

Figure 5: Picking error rates for the three technologies before and after the correction of systematic errors.

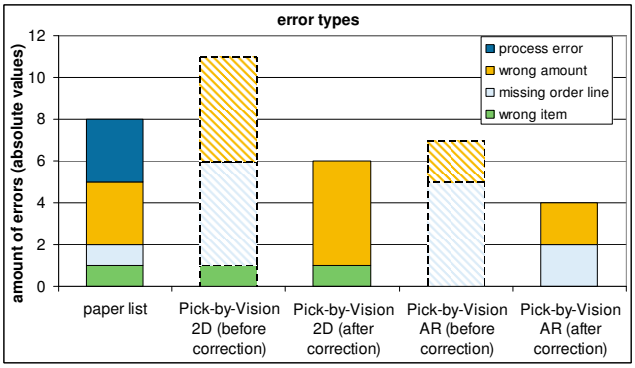

Figure 6: Kinds of picking the overall errors for the three technologies before and after the correction of systematic errors.

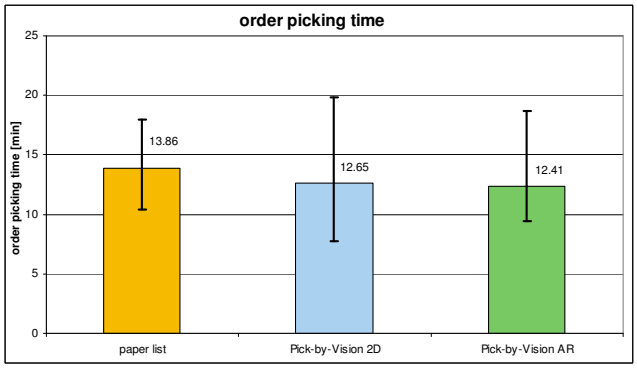

Figure 7: mean values of the order picking time for the three technologies

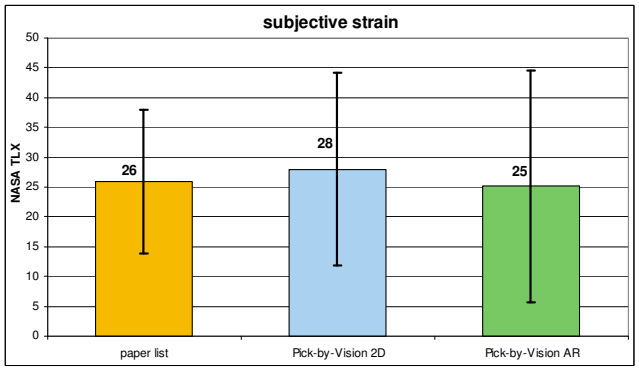

Figure 8: NASA TLX for the three technologies. Values can be between 0 (no task load) - 100 (full task load).

than the mean value. Which, on the other hand, means that one third of the participants performed really slow. But we could not trace this back to any specific reason.

Subjective Strain The NASA TLX, test resulted in nearly the same values (25-28) for all systems (see Fig.8). We could not find (ANOVA, $\alpha = 5\%$) any significant difference in the subjective strain and therefore could not proof hypothesis **E1_H3**.

Results from Questionnaire We could find a tendency that people with previous experience with 3D user interfaces performed better in both logistics operating figures (fewer errors and faster). An analysis of the questionnaire did not show that subjects felt uncomfortable or constrained by using the Pick-by-Vision system (it should be noted that subjects had to only work for about half an hour with the system).

5.3 Conclusion of Evaluation 1

Our first field trial showed slight benefits for the Pick-by-Vison systems over paper-based picking. Even though none of the results is significant, we could at least deliver the first positive results to prove the concept. If one takes into account that our technologies, in particular the Pick-by-Vision (AR) system (at version 1.0), still had several open usability issues leading to bad performance in terms of time, this is a quite good result for Pick-by-Vision. If we take a deeper look into the structure of errors made we can see that, with Pick-by-Vision (AR), none picked an item out of the wrong box. Moreover, we could not find an increased level of strain produced by the Pick-by-Vision system, but we just did a "short" test, where subjects wore the HMD only for a short time period. However, we figured out that some people had trouble reading the information from the HMD, and one third performed quite bad with the system in general. Whereby we could not find a significant correlation between both groups. Apart from that, we identified several aspects of our AR visualization to be optimized, leading to the fundamentally improved Pick-by-Vision (AR) 2.0 system.

6 Evaluation 2

The previous experiment showed that Pick-by-Vision works in general and is already comparable to the established paper-based picking support. That's why we decided to test our Pick-by-Vision system in an in-depth evaluation over a "longer" period of time. On the one hand we wanted to get insight about how learning effects affect errors and performance. Especially when comparing our error rates with common error rates (see [27]), we could see a lot of potential for learning effects to influence the results after a longer period of use. On the other hand, we knew from [12] that long-term use of an HMD can be a load for the user resulting in headaches, eye fatigue and discomfort.

To this end, we designed a new experiment following our first study that analyzed user strain [28]. As the execution of a mid-term study goes along with a hugh expense, especially if users have to do real tasks, we had to determine some limitations. We decided to use only a few participants, and follow a rather in-depth qualitative approach, than a quantitative one. Furthermore, we expected the Pick-by-Vision (AR) system to deliver more interesting results than the Pick-by-Vision (2D) system. Finally, we set up an experiment comparing our Pick-by-Vision (AR) system in a mid-term study with a common paper based system.

6.1 Experimental Setup and Hypotheses

Experimental Setup The only independent variable of this experiment was the type of the picking system used: Pick-by-Vision (AR) or a paper-based list. We designed a with-in subject experiment in which each subjects had to work with each technology for two hours, whereby the start sequence of each technology was permuted. A single test session for a technology lasts about four hours, including the actual two hours picking, pre- and post-tests, recovery lying (resting) phases and up to half an hour try-and-ask introduction into the technology and a structured interview at the end of the test. Each test session for each subject was executed on two different days at the same time of day, to get comparable results for the strain parameters. Within the two hours, subjects should fulfill at maximum 100 orders (2.9 order lines/order, 2.2 items/order line). Three rack compartments contained wrong articles. The wrong articles were supposed to be discovered by the users during the picking.

We measured the logistic operating figures (order picking time and errors) for a whole run and for the single orders.

The analysis of user strain was done in the same way as we did in [28]. This means that we analysed the heart rate variability (HRV), used an EZ-Scale and a discomfort questionnaire. In addition to that, we again used the NASA TLX. The subjects wore a "Polar RS 800 CX Multi" pulse recorder for analysis of HRV, the software used to retrieve the data from the recorder was "Polar ProTrainer 5".

Even if there was no difference in strain in the first evaluation, we expected the Pick-by-Vision system to produce a higher strain in this experiment, This was because this time the subjects had two work with the system for two hours. That is why we set up the following hypotheses:

> **E2_H1:** The analysis of HRV data reveals a higher user strain for the Pick-by-Vision (AR) system than for the paper list.
>
> **E2_H2:** The analysis of EZ-Scale data reveals a higher user strain for the Pick-by-Vision (AR) system than for the paper list.
>
> **E2_H3:** The analysis of discomfort questionnaires reveals a higher user strain for the Pick-by-Vision (AR) system than for the paper list.
>
> **E2_H4:** The NASA TLX score is larger for Pick-by-Vision (AR) than for the paper list.

The Pick-by-Vision (AR) system already showed good results in the first evaluation and moreover we expect learning effects from using the AR system over a longer periode of time. Because of that, we set up the hypotheses that the Pick-by-Vision (AR) system should perform better according to the logistic figures:

> **E2_H5:** The error rate is lower or the same with the Pick-by-Vision (AR) system compared to the paper list.
>
> **E2_H6:** The order picking time with the Pick-by-Vision (AR) system is equal or better compared to the paper list.

Observation Strategies To get a deep insight into the experiment, we had several assistants besides the actual investigator, mainly to help with checking and sorting the picked articles. Moreover, we created the role of a special observer, who was placed on an observation deck a few meters above the actual experimentation area. This physical position had three benefits: the observer could perfectly observe, was not recognised by the test subject (like behind a traditional half-mirror) and, furthermore, was allowed to give comments to the subject and even to interrupt the expriment.The observer observed the user behaviour by mainly focussing on user strategies (how the user handles the system or how the user tries to conquer usability problems). To this end, the observer tried to analyze each little step / movement a user did. While observing the user, the observer noted questions for an interview after the test. In cases where the observer observed the test subject having severe trouble, in particular leading to unnecessary bad performance, the observer could intervene in the experiment. We stopped the time during the breaks and solved the problem by giving a comment or e.g. adjusting the HMD. We think this policy falsifies the results less than not interrupting the test subject.

After each test the subjects were confronted with the observations in a semi-structured interview. Such interviews typically lasted between fifteen and thirty minutes. The interviews were recorded using a tape recorder.

6.2 Results

We had 8 subjects (4 male/4 female) between the ages of 18 and 37 (mean age: 26, std dev: 5.37) performing the test. Subjects were, again, from all around the university, friends and as well as two professional order picking workers.

6.2.1 Analysis of User Strain

The objective user strain was measured by analyzing the HRV data from the pulse recorder. In general, it can be said that the frequency of the heart rate is not an indicator for the stress level, but rather the variability in the heart rate. A high variability in heart rate indicates a low stress level, whereas a constant heart beat is an indicator for stress. For the analysis of HRV the standard deviation (SD) was chosen as a parameter of the time domain which is used as a marker for short term changes of the sympathetic and parasympathetic nervous system indicating changes in user strain.

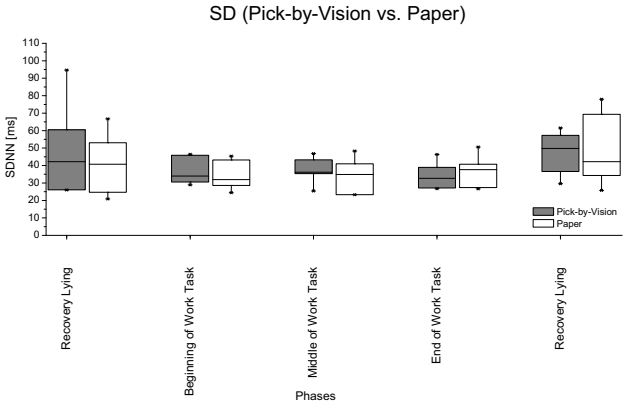

Figure 9: Development of SD through the test phases

Fig. 9 shows the development of SD through the test phases. The analysis of SD does not show a significant difference (Wilcoxon Test) between paper list and Pick-by-Vision (AR). This means that no higher physiological strain can be seen between both systems, **E2_H1** thus can be rejected.

The EZ-Scale describes the subjective state of well-being of a person by Stanine values. Altogether changes in different factors can be seen for both systems indicating a rise in strain. Significant differences before and after the test can be found (Wilcoxon Test) for the Pick-by-Vision (AR) system in a reduction of activation ($p = .020$), rise in fatigue ($p = .027$) and decreasing relaxation ($p = .046$). For the paper list the willingness for exertion is reduced ($p = .039$) as well as social acceptance ($p = .034$) and sleepyness ($p = .041$). From this data it is obvious that both systems result in an increase of strain significantly influencing different factors of

personal well-being. Again it was not indicated that working with the Pick-by-Vision (AR) system caused a higher strain than working without AR, so that **E2_H2** can be rejected.

The discomfort questionnaire asked for current physical complaints mostly focused on the visual system of the user. The analysis of the data revealed significant differences between Pick-by-Vision (AR) and paper list only for the factor "headache" ($p = .034$, fig. 10). On the one hand, this corresponds to the results presented

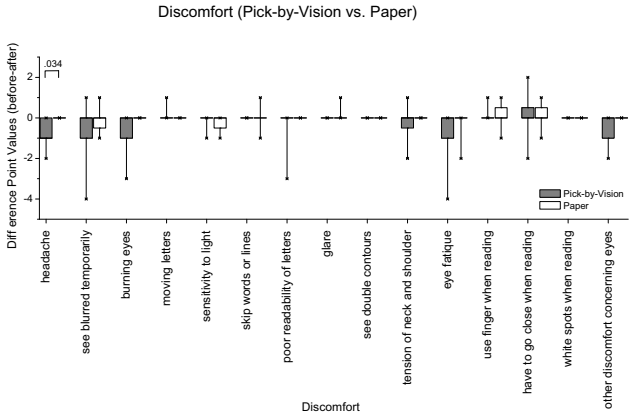

Figure 10: Result of discomfort questionnaire comparing differences between Pick-by-Vision (AR) and paper list before and after testing

in [28], because only one out of 15 factors showed significant differences. But on the other hand, we found a change in another factor (headache with $p = .034$ instead of burning eyes, Wilcoxon Test). Regarding **E2_H3** we cannot fully agree or disagree since only one out of 15 items revealed a significant difference.

The NASA TLX shows a task load of 87.81 for Pick-by-Vision (AR) and 71.98 for the paper list. However, there is no significant difference (ANOVA, α = 5%) (compare Fig.11). So **E2_H4** can neither be accepted nor rejected. Compared to the first experiment, both values are higher, even though the actual task was the same. We lead this back to the fact that two hours of manual work without a break can be straining in general.

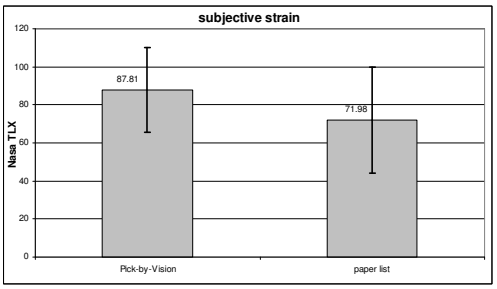

Figure 11: NASA TLX for both technologies. Values can be between 0 (no task load) - 100 (full task load).

6.2.2 Logistic Figures

Error Rates In this test series the error rate for the Pick-by-Vision (AR) was higher than for the paper list (figure 12) and thus **E2_H5** can be rejected. Within 5,873 order lines for the paper list 29 (1.37% $\pm$ 1.09%) were picked wrong, compared to 58 out of the 5,519 order lines (3.03% $\pm$ 2.24%) for the Pick-by-Vision (AR). We corrected the value of the Pick-by-Vision system to 2.34% $\pm$ 2.08%, as subjects had some problems with the adjusting knob. The knob was not suited for long term use, and subjects sometimes pressed the button twice accidentally, and just told us about the mistakes rather than use the go-back function.The error rates are shown in Fig.12

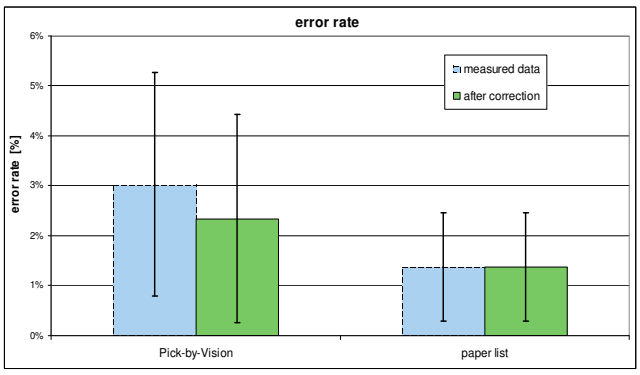

Figure 12: picking error rates for both technologies before and after the correction of the systematic errors

There are several reasons for the picking errors (compare Fig.13). With the paper list most of the errors (14) were based on the wrong amount, seven on the wrong article and six on a missing order line. There were also two processing errors because the subjects forgot to sign the list. With Pick-by-Vision (AR) the wrong amount was also the most frequent error (23).

Despite of reducing errors caused by the systematic system error, there are still 15 missing order lines. For example the subjects clicked the button when they entered the aisle. This was not necessaray and the first order line in this aisle was missing. The process flow of the Pick-by-Vision (AR) system has to be improved in this case. For AR the picking of wrong items is the most interessting error. The subjects picked five wrong articles, three of which were due to the wrong items we had integrated. These three errors were not recognized by the subjects because they did not check the article number. The subjects did, however, pick from the right storage compartment, therefore this error does not directly depend on the AR visualization. One subject picked two items of one order line right and one item from the storing compartment beneath. He did not recognize the error because the items had the same shape. A more precise inspection of this leads to the following explanation: The subject first took two items (three were not possible at the same time because of the weight). When he took the third item, he did not move his head back to see the augmentation highlighting, but rather saw the box, while picking, only in the corner of his eye. One subject picked an item far away from the actual box. As we had to refill the warehouse manually and could not be 100% certain that all articles were placed correctly we think this mistake was due to an incorrect refillment.

When we compared the distribution of errors over the time, we figured out that the errors for the Pick-by-Vision (AR) system decreased during runtime and increased for the paper list.

Time Subjects were supposed to work with both techniques for two hours. They performed 133 ($\pm$ 35) order lines using the list and 124 ($\pm$ 55) with Pick-by-Vision (AR). However, some subject fulfilled the 100 given orders in less time, some subjects had to take a break and we had a few breakdowns. After taking these corrections into account, subjects performed about 7.6% faster using Pick-by-Vision (AR): 145 order lines per hour compared to 134 order lines per hour using the paper list (see Fig. 14). Thus, order picking with Pick-by-Vision (AR) is faster and **E2_H6** was found to be true.

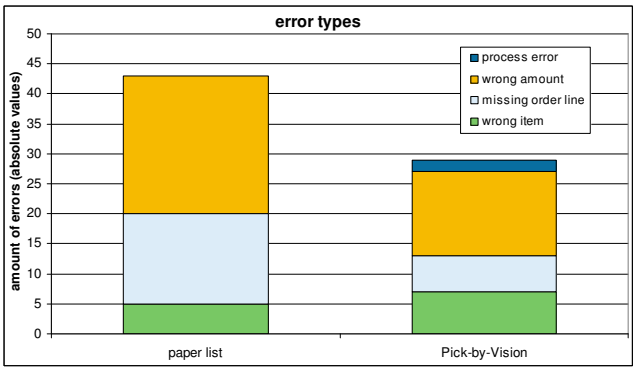

Figure 13: kinds of picking errors for both technologies before and after the correction of the systematic errors

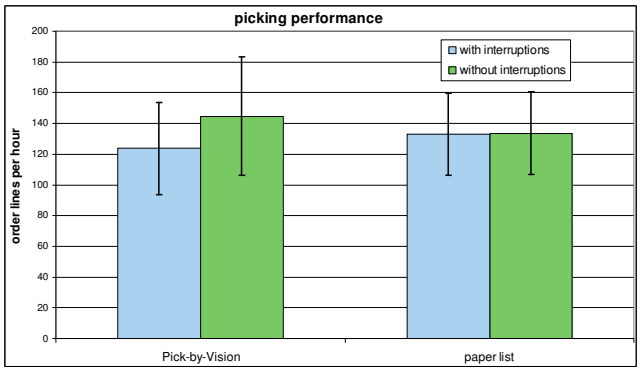

Figure 14: mean values of the order picking performance for both technologies with and without regarding the interruptions

6.2.3 Problems using the HMD

There were several problems directly related to the HMD. Two people complained about headaches, which they traced back to the HMD headband. Three people complained about pressure in their eyes. Both observations were also made by [12, 6]. Two of our subjects had serious problems focussing on the HMD and stopped regularly during the experiment to try to focus the display. One of them even needed a 15-minute break because otherwise "the eye would have jumped out". We thought about psychological reasons, like being afraid of the technology, but the subject had a high affinity with new technologies. The other subject said that as long as the background was in the focal distance she could focus on the HMD. We tested both subjects using a Landolt-C-Ring test, both had neither perfect nor particularly bad eyes.

In general people said that they sometimes needed some time to focus on the HMD. However, subjects only complained about reading the text from the HMD, which was, in our opinion, displayed in an appropriate size (compare Fig.1). None of the subject had problems seeing the actual 3d augmentation. We reason that one does not need to focus on the 3d augmentation to be able to interpret and work with it. On the other hand, 50 % of the subjects did not have problems with the focus.

6.2.4 Other Observations

Against our expectations, the real picking workers performed slow compared to the other subjects. This has several reasons: They tried to do an error-free job because they know about the consequences of false delivered items. They placed all items in an organized way with the label side up in the sump, to have a fast overview before delivery, while other subjects just throw the items in the box. Furthermore, we realized that this is their real job and they actually do not do it in a rush, as they usually have to do it eight hours a day.

Besides this observation, we found out that there are two groups of HMD users. One group works parallel: they make user input while walking. The other ones stop walking for each user input. Even though we interrupted people to suggest that they work parallel, they did not change their behaviour. For that reason, that group performed slower than those that worked using the parallel method.

6.3 Discussion of the Results

Even if it does not look like it on the first view, the results of this first endurance test are quite satisfying:

Subjects performed slightly faster using the Pick-by-Vison (AR) system. We could identify several issues at Pick-by-Vision (AR) 2.0 system, which have slowed down the usage of the system in this experiment: bad input device, too many system states to click through, no direct back button, late display of the aisle to go to, partial occlusion of the text by the augmentation and imperfect guiding by the augmented reality visualization. We could address most of these usability issues in the Pick-by-Vision (AR) 3.0 system, which we developed after this experiment.

According to the error rate, subjects performed three times worse than in the first evaluation. Even if the AR visualization seems to be a perfect indicator for the right box, subjects could and did make other errors. Mainly they ignored to control the article numbers or the amount. On the one hand we traced this back to the fact that they had problems focusing on the text of the HMD, and some of them definitely ignored it. On the other hand, subjects probably grew tired and were not concentrating anymore. Those errors have to be avoided by the use of other mechanisms, for example a speech input[19]. In such systems users have to speak the amount and an additional error checking number.

Another important result of this study is that the results of HRV, EZ-Scale and discomfort questionnaire show no general differences in strain changes between paper list and Pick-by-Vision which is very similar to our first study of user strain [28]. This means the Pick-by-Vision system in general is no additional load to the user compared to a paper list. From the EZ-Scale we can see that both systems influence different subjective parameters. The results of the discomfort questionnaire point out that working with an HMD can result in a headache, even though the reason for that remains unclear. Either the headband of the display was set too tight or the changes between the virtual image plane and the surroundings are a main reason for this finding. Current research projects deal with investigations into that problem (e.g. [11]).

7 Conclusion

Doing research in industrial augmented reality is a challenging task, as industry partners want to know about the capability of a technology they invest in. AR still has many unsolved problems but we need to compare our AR applications with established technologies to motivate its capabilities. In this paper we show how to do such a comparison in a fair way and without whitewashing the results.

We present two user studies undertaken in an order picking scenario to analyze picking performance, error rates and user strain of our Pick-by-Vision systems. In summary, it can be stated that Pick-by-Vision (AR) can increase the performance of an order picking worker according to the main important logistic figures, namely time and error rate. However, there were still some errors made, using the Pick-by-Vision (AR) system. It was a perfect indicator, for the right box, but people picked the wrong amount, or did not look at the article number to see that a wrong article was in the box. Those problems have to be conquered by other control mechanisms.

Regarding user strain, we found that even though we have uncomfortable HMD headbands, a backpack to carry, and non-

addressable display focal planes, our system did not cause a higher general user strain than the conventional paper list. Nevertheless, the discomfort questionnaire shows that improvements of the display devices are necessary to reduce the potential for headaches.

The problem remains that about 20% of subjects had serious problems using the HMD. Brau and Fritzsche [12, 6], figured out the same problem for 20% of their subjects by also using the same Nomad HMD. In our case users did not have problems with the 3D augmentations, but just with reading the 2D text. Further investigations are necessary to find out if it takes some time for habituation or if it is directly related to the people or the type of HMD.

Besides that, we showed the useful application of an active observer, who had a complete overview and was allowed to intervene in the experiment. This prevented subjects from performing poorly and distorting the results, just because they, for example, wore the HMD in a bad way. We draw a conclusion of following our philosophy of observing a few people in-depth rather than following an approach with a large number of users. If we had used more subjects in the tests, we probably would have had more statistically significant results, but the insights and conclusion would have been pretty much the same.

Even though the Pick-by-Vision (AR) system was slightly better than the paper list this improvement typically is not significant enough to introduce and apply the technology in industry. Thus further improvements and tests of our v3.0 system will follow.

Acknowledgements

This work was partially supported by the German Federal Ministry of Education and Research (AVILUS project, grant no. 01 IM 08 001 A) as well as the German Federal Ministry of Economics and Technology (AiF-FV 14756).

The authors wish to thank ART GmbH for supporting us with Tracking Cameras. Many students were involved in the project. Two should be mentioned here: Michael Stather and Max Meister. Thanks to Margarita Anastassova for the help with the evaluation strategies. Finally, thanks to all participants who took part in the different kinds of experiments and gave us feedback for optimizing the Pick-by-Vision system.

References

[1] M. Anastassova, C. Mégard, and J.-M. Burkhardt. Prototype evaluation and user-needs analysis in the early design of emerging technologies. In *Human-Computer Interaction. Interaction Design and Usability*, volume 4550 of *LNCS*. Springer Berlin / Heidelberg, 2007.

[2] H. Brau, C. Ullmann, M. Duthweiler, and H. Schulze. Gestaltung von augmented reality applikationen für kommissionieraufgaben. In L. Urbas and C. Steffens, editors, *Zustandserkennung und Systemgestaltung Bd. 19*. VDI-Verlag, 2005.

[3] H. Brynzer and M. I. Johannsonn. Design and performance of kitting and order picking systems. *International Journal of Production Economics*, 45:115–125, 1995.

[4] D. Drasic and P. Milgram. Perceptual issues in augmented reality. In *Proc. SPIE Vol. 2653*, 1996.

[5] A. Dünser, R. Grasset, and M. Billinghurst. A survey of evaluation techniques used in augmented reality studies (tr-2008-02). Technical report, University of Canterbury, HITLabNZ, 2008.

[6] L. Fritzsche. Eignung von augmented reality für den vollschichteinsatz in der automobilproduktion. Master's thesis, TU Dresden, 2006.

[7] T. Gudehus. *Logistik*. Springer, Berlin, 3. edition, 2005.

[8] W. A. Günthner. *Neue Wege in der Automobillogistik: Die Vision der Supra-Adaptivität*. Springer, 2007.

[9] S. Hart and L. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, pages 139–183. North-Holland, Amsterdam, 1988.

[10] A. Huckauf, M. H. Urbina, I. Böckelmann, L. Schega, R. Mecke, F. Doil, and J. Tümler. Perceptual issues in optical-see-through displays. In *submitted to eighth IEEE and ACM International Symposium on Mixed and Augmented reality*, 2009.

[11] A. Huckauf, M. H. Urbina, F. Doil, J. Tümler, and R. Mecke. Distribution of visual attention in head-worn displays. In *Proceedings of the ACM Symposium on Applied Perception in Graphics and Visualisation 2008 (APGV08)*, Los Angeles, California, USA, 2008. ACM.

[12] J. Kampmeier, A. Cucera, L. Fritzsche, H. Brau, M. Duthweiler, and L. G. K. Eignung monokularer augmented reality – technologien in der automobilproduktion. In *Tagung der Deutschen Ophthalmologischen Gesellschaft "Augenheilkunde in der alternden Gesellschaft - Herausforderung und Chance"*, 2006.

[13] S. Liu, D. Cheng, and H. Hua. An optical see-through head mounted display with addressable focal planes. In *7th IEEE/ACM International Symposium on Mixed and Augmented Reality, 2008. ISMAR 2008.*, 2008.

[14] M. A. Livingston. Evaluating human factors in augmented reality systems. *IEEE Comput. Graph. Appl.*, 25(6):6–9, 2005.

[15] J. R. Nitsch. Die Eigenzustandsskala (EZ-Skala) - Ein Verfahren zur hierarchisch-mehrdimensionalen Befindlichkeitsskalierung. In J. R. Nitsch and I. Udris, editors, *Beanspruchung im Sport. Beiträge zur psychologischen Analyse sportlicher Leistungssituationen*, pages 81–102, Bad Homburg, Germany, 1976. Limpert.

[16] O. Oehme. *Ergonomische Untersuchung von kopfbasierten Displays für Anwendungen der erweiterten Realität in Produktion und Service*. PhD thesis, RWTH Aachen, 2004.

[17] O. Oehme, L. Schmidt, and H. Luczak. Comparison between the strain indicator hrv of a head based virtual retinal display and lc-mounted displays for augmented reality. In *Proceedings of the Conference WWDU 2002 World Wide Work - 2002*, pages 387–389, Berchtesgaden, 2002. Abindgon, Oxon, UK : Taylor & Francis.

[18] C. Plaisant. The challenge of information visualization evaluation. In *AVI '04: Proceedings of the working conference on Advanced visual interfaces*, pages 109–116, New York, NY, USA, 2004. ACM.

[19] R. Reif, W. Günthner, B. Schwerdtfeger, and G. Klinker. Pick-by-vision comes on age: Evaluation of an augmented reality supported picking system in a real storage environment. In *6th International Conference on Computer Graphics, Virtual Reality, Visualisation and Interaction in Africa (Afrigraph 2009)*, 2009.

[20] J. Rolland, A. D., and G. W. Towards quantifying depth and size perception in virtual environments. *Presence: Teleoperators and Virtual Environments, 4(1)*, 1995.

[21] J. P. Rolland, M. Krueger, and A. Goon. Multifocus planes in head-mounted displays. *Applied Optics, 39(19)*, 2000.

[22] D. Rowe, J. Silbert, and D. Irwin. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *CHI98*, pages 480–487. ACM Press, 1998.

[23] B. Schwerdtfeger, T. Frimor, D. Pustka, and G. Klinker. Mobile information presentation schemes for logistics applications. In *Proc. 16th International Conference on Artificial Reality and Telexistence (ICAT 2006)*, November 2006.

[24] B. Schwerdtfeger and G. Klinker. An evaluation of augmented reality visualizations to support the order picking, 2008. Technische Universität München, Report TUM-I-08-19.

[25] B. Schwerdtfeger and G. Klinker. Supporting order picking with augmented reality. In *Proc. of the seventh IEEE and ACM International Symposium on Mixed and Augmented reality*, September 2008.

[26] B. Schwerdtfeger, G. Klinker, R. Reif, and W. Günthner. Pick-by-vision: There is something to pick at the end of the augmented tunnel. Technical Report TUM-I0921, TU München, 2009.

[27] M. ten Hompel and T. Schmidt. *Warehouse Management*. Springer, Berlin, 2004.

[28] J. Tümler, R. Mecke, M. Schenk, A. Huckauf, F. Doil, G. Paul, E. Pfister, I. Böckelmann, and A. Roggentin. Mobile augmented reality in industrial applications: Approaches for solution of user-related issues. In *Proc. of the seventh IEEE and ACM International Symposium on Mixed and Augmented reality*, 2008.

[29] VDI, Berlin. *VDI guideline 3590: Order picking systems*, 1994.