Putting Gaze into Context: A Framework for Analyzing Gaze Behavior in Interactive and Dynamic Environments

Thomas Bader
Lehrstuhl für Interaktive Echtzeitsysteme
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT),
Germany
bader@kit.edu

Jürgen Beyerer
Fraunhofer IOSB
Institut für Optronik, Systemtechnik und
Bildauswertung
Germany
juergen.beyerer@iosb.fraunhofer.de

ABSTRACT
Gaze data contains valuable information about user's cognitive processes during execution of a task. In order to use this information, e.g., for studying user's strategies or for designing new gaze-based interaction techniques for HCI, gaze data needs to be aligned with the task executed by the user.

In this paper we propose a novel framework based on the theory of Markov Decision Processes for putting gaze data into context, allowing for automated interpretation of gaze position and movement with respect to the task performed by the user. The model can be used for both, offline and online analysis of gaze data. We evaluate the proposed model with an indirect object manipulation task and demonstrate how it can be used for intention recognition and/or detection of a mismatch of the user's mental model.

Author Keywords
gaze-based interaction, cognitive model, multimodal integration

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces

INTRODUCTION
Visual perception is an important information channel during manipulation of real or virtual objects (e.g., icons on a graphical user interface). It allows for perceiving the current state of manipulated objects and/or for visuomotoric control of manipulators like our hands or a computer mouse. During manipulation tasks, our gaze behavior is mainly controlled top-down and subconsciously by cognitive processes which are responsible for task execution. Therefore, natural gaze behavior provides a window into the human mind and allows a conclusion to be drawn about user's intentions and cognitive processes. This information could be valuable for future adaptive and proactive human-computer interfaces (HCIs).

However, natural gaze behavior is very complex and is influenced by a vast amount of factors. Therefore, in order to allow for integration of natural gaze as an additional modality into HCIs the following steps need to be undertaken:

1. A thorough understanding of natural gaze behavior in dynamic environments needs to be established.
2. Methods for automated online analysis of gaze data in the context of dynamic environments need to be developed.
3. New multimodal gaze-supported HCIs need to be designed, implemented, and evaluated.

All of the three points above have been already covered by a large body of research. *Natural gaze behavior* has been studied in different natural environments (e.g., during block-copying [11], basic object manipulation [6], driving [8], and playing cricket [9]) as well as during human-computer interaction (e.g., [15, 1]) and on the field of psychology and physiology (e.g., [4, 3]). However, the results of these studies like different gaze behaviors observed during task execution mostly are reported in an informal way, e.g., as verbal descriptions or as plots of gaze data. Such descriptions can help to improve the principal understanding of gaze, however, a more formal description of different gaze behaviors in different contexts is required in order to make results comparable and accessible for automated interpretation.

General *methods for automated online analysis* of gaze data currently are mainly limited to fixation detection [14, 13] and analysis of fixation frequency, duration, and position (e.g., [12]). Alignment of gaze data with the task or cognitive processes often is done manually or is restricted to static environments (e.g., [14, 5, 13]). In order to develop new methods for interpretation of natural gaze behavior in arbitrary interactive and dynamic environments a common formal framework would be very helpful.

Most state-of-the-art *gaze-based interfaces* use gaze as an explicit pointing device, e.g., as replacement for a mouse [10]. This requires gaze to be used for manipulation (e.g., for pressing keys on a virtual keyboard [10]) in addition to its natural purpose, namely visual perception. Such interaction techniques might be useful for certain applications, e.g., when hands are not available as an input modality. However,

Presented at International IUI 2010 Workshop on Eye Gaze in Intelligent Human Machine Interaction.

using gaze-based pointing as a general input technique for human computer interaction has many limitations (see [1]). Promising examples for gaze-supported interaction techniques are presented in [5] and [16]. In both approaches *natural* gaze behavior is analyzed and the user is not forced to diverge from that natural behavior for interaction purposes. *iDict* [5] analyzes the duration of fixations while the user reads a text in a foreign language and automatically provides a translation of the fixated word if a longer fixation is detected. In the approach "Manual And Gaze Input Cascaded (MAGIC) Pointing"[16] the mouse pointer is placed close to the currently fixated object in order to eliminate a large portion of the cursor movement. Both approaches do not use gaze directly as pointing or input device, but interpret gaze data in the context of the task (reading, pointing). However, the link between gaze and cognition in none of the two approaches is made explicit in form of a model. This limits generalization and development of a deeper understanding of the underlying principles of such techniques. The need for modeling the dependencies between gaze and the task was also already stated by other researchers (e.g., [12] suggest to use tools from cognitive modeling).

In this paper we propose a new framework based on Markov Decision Processes (MDPs), which can provide a common ground for all of the three above mentioned steps towards adequate interpretation of natural gaze behavior in interactive environments and usage as additional modality in future HCIs. In particular we consider the following aspects as important to be covered by such a framework:

- *Uncertainty of knowledge* of the user about the system which he/she interacts with seems to play an important role for explaining and interpreting natural gaze behavior [1]. In contrast to traditional approaches to cognitive modeling of tasks like GOMS [7] the proposed framework explicitly allows for modeling uncertain knowledge.
- *Multiple strategies* may lead to a desired goal when interacting with a system. Natural gaze behavior therefore can be influenced by more than one of those strategies and by the choice between them, respectively. In the proposed framework multiple possible strategies can be explicitly modeled and/or generated automatically.
- *Simplicity* of the framework is important for keeping it applicable for the above mentioned steps. Even if important components of cognition like short-time memory are not explicitely covered by the framework, it provides a good basis for investigating and interpreting natural gaze behavior for many tasks.

The framework proposed here is an extension of the work presented in [1], where fundamental causal relations between task execution and gaze behavior are discussed and modeled in a probabilistic framework. However, in contrast to [1] the model presented here is more generic and allows for modeling of more complex tasks.

FRAMEWORK FOR ALIGNMENT OF GAZE DATA

In this section we propose a formal framework for modeling the interdependencies between gaze and the task executed by the user. We first derive a formal model, then show how gaze behavior can be classified into different categories related to the current state of task execution. Finally, we illustrate what can be further inferred from the model regarding user intention and mismatches between the mental model and the real system.

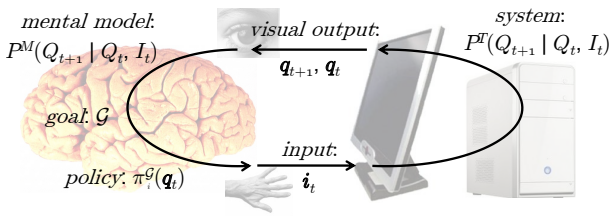


Figure 1. Basic components of the model.

A formal model of an interactive system and its mental representation

The human decision process during interaction with an interactive system consisting of a number of interactive objects can be modeled by the 4-tupel $(\mathcal{Q}^o, \mathcal{I}, P(Q^o_{t+1}|Q^o_t, I_t), c(\cdot)^o)$, which defines an MDP. We denote $\mathcal{Q}^o$ as the set of all possible states of an object. $\mathcal{I}$ is the set of possible inputs or actions which can be performed by the user in order to change the system state. $P(Q^o_{t+1} = \boldsymbol{q}^o_{t+1}|Q^o_t = \boldsymbol{q}^o_t, I_t = \boldsymbol{i}_t)$ denotes the probability of a state transition of object o from state $\boldsymbol{q}^o_t$ to state $\boldsymbol{q}^o_{t+1}$ if input $\boldsymbol{i}_t \in \mathcal{I}$ is performed by the user. The function c defines costs for each system state and state transition, respectively. These costs can reflect relevance of certain system states for the success of the whole task, but also cognitive or physical load induced by certain inputs. Without loss of generality, in the following we will focus on one single object and therefore omit the index o in the formulas above.

Usually, state transitions in a technical interactive system, in contrast to those in natural ones, are deterministic. Hence, the state transition probabilities P^T for most technical systems reduce to

$$P^T(Q_{t+1} = \boldsymbol{q}^i_{t+1} \mid Q_t = \boldsymbol{q}_t, I_t = \boldsymbol{i}_t) = 1 \text{ and} \quad (1)$$

$$P^T(Q_{t+1} \neq \boldsymbol{q}^i_{t+1} \mid Q_t = \boldsymbol{q}_t, I_t = \boldsymbol{i}_t) = 0, \quad (2)$$

where $\boldsymbol{q}^i_{t+1}$ denotes the state induced by input $\boldsymbol{i}_t$. However, the mental model of such a deterministic system can be incomplete and/or uncertain. Since gaze behavior is determined by this mental representation of the interactive system P^M and not by the real one (P^T) we chose to use a probabilistic framework in order to allow for modeling of effects such as a model mismatch ($P^M \neq P^T$). The different basic components of the model are illustrated in Figure 1.

When interacting with a system, the user mostly has a certain goal in mind. He/she wants the system to take a certain target state, which has to fulfill some requirements. We therefore model a goal as a subset of system states $\mathcal{G} \subseteq \mathcal{Q}$ with the

indicator function

$$\mu_{\mathcal{G}}(\boldsymbol{q}) = \begin{cases} 1 & \text{if } \boldsymbol{q} \in \mathcal{G}, \\ 0 & \text{if } \boldsymbol{q} \notin \mathcal{G}. \end{cases} \quad (3)$$

Such a goal for example can be the selection of a set of objects or their positioning in a target area on the screen. A complete task can consist of multiple subgoals, which have to be reached subsequently.

To convey an interactive system from its initial state $\boldsymbol{q}^0$ into a target state $\boldsymbol{q}^{\mathcal{G}} \in \mathcal{G}$, the user has to execute a sequence of actions. In most tasks not only one action sequence leads to the target state, but many different "ways" can be chosen by the user. The different *policies* the user can follow are described as a set of functions

$$\pi_i^{\mathcal{G}}(\boldsymbol{q}) : \mathcal{Q} \rightarrow \mathcal{I}, \quad i = 1, ..., N_\pi, \quad (4)$$

where the function $\pi_i^{\mathcal{G}}(\boldsymbol{q})$ specifies the action the user will choose according to policy i when the system is in state $\boldsymbol{q}$ and the goal is $\mathcal{G}$. N_π is the number of possible policies.

Given a certain task with an initial state $\boldsymbol{q}^0$ and a goal $\mathcal{G}$ the user has to decide, which actions are to be executed in order to reach the goal. $P^M(Q_{t+1}|Q_t = \boldsymbol{q}_t, I_t = \boldsymbol{i}_t)$ reflects the knowledge the user has about the system, namely which effects a certain input $\boldsymbol{i}_t$ has on the state of an object and the system, respectively. If the user had perfect knowledge of the system ($P^M = P^T$), he could calculate the set of optimal policies for reaching the goal according to his internal value function c, describing costs for executing the different inputs and the amount of reward for reaching certain system states. If the user had no knowledge of the system, he could have the same value function c but would not be able to calculate any policy, since the effect of actions to the system state would be unknown. Therefore, in this case the user would have to build a mental model of the system previous or in parallel to the execution of the primary task.

Note that the framework described above can not only be used for modeling interactive systems, but also for dynamic environments the user has no influence on ($\mathcal{I} = \{\}$).

Two reasons why we look where we look

Given the above model of an interactive system we have two reasons to draw our visual attention on a certain location in an interactive environment.

The first one is *control of input* $\boldsymbol{i}$. If the human operator once has decided to execute a certain action in order to reach the goal, it is important to perform this input as accurate and fast as possible. Since absolute positioning of limbs by only proprioceptive feedback is not very accurate [2], vision is needed as additional feedback channel for accurate positioning, e.g., of the hand. An input can be either controlled *directly* by observing the input device or body part, or *indirectly* by observing the system reaction.

The second reason for drawing our visual attention on a certain location is the *verification of system reactions or states*. If we had a perfect mental model of the system and complete and accurate non-visual feedback about our actions $\boldsymbol{i}$, we could execute the task successfully with closed eyes. However, neither the first nor the second assumption is realistic. In most cases the mental model is incomplete or uncertain, and, as already mentioned above, non-visual feedback channels (e.g., proprioception or touch) are not accurate enough. Therefore, in order to create, improve or verify the mental model of the system the user has to check permanently, whether a certain input leads to the anticipated system reaction or not. The more confident the mental model is, the less verification of the system reaction is required.

For further investigations of these two processes, which concurrently influence natural gaze behavior during human-computer interaction, alignment of gaze data with the task and system states is required.

Methods for alignment of gaze data

To align gaze data with the model presented above we use both, absolute position and movement of gaze. Typically gaze movements are separated into two different components: *fixations* and *saccades*. While saccades are rapid eye movements used to locate the gaze at a certain position, gaze remains almost still during fixations to enable retrieval of visual information. Two different algorithms have been implemented for automated fixation detection. Both algorithms "I-DT" (Dispersion-Threshold Identification) and "I-VT" (Velocity-Threshold Identification) are taken from [14]. The first algorithm clusters gaze points according to their spatial distribution, the second one according to the velocity of gaze movements.

In the following we present methods for interpretation of gaze data in an interactive environment with visual representation of object and system states on a 2-dimensional planar display. The state of an object in such an environment can be described by $\boldsymbol{q}_t = (\boldsymbol{p}_t, \boldsymbol{\alpha}_t)$, where $\boldsymbol{p}_t \in \mathbb{N}^2$ is the position of the visual representation of the object on the display and $\boldsymbol{\alpha}_t$ describes further attributes of the object. Fixation positions are further denoted with $\boldsymbol{f}_t \in \mathbb{N}^2$. In order to analyze gaze positions in the context of a task, we calculate the difference vector $\boldsymbol{v}_t = \boldsymbol{f}_t - \boldsymbol{p}_t$. Additionally, we define the vector $\boldsymbol{w}_t^i = \hat{\boldsymbol{p}}_{t+1}^{\pi_i} - \boldsymbol{p}_t$ with

$$\hat{\boldsymbol{p}}_{t+1}^{\pi_i} = \underset{\boldsymbol{p}_{t+1}}{\operatorname{argmax}} \quad P^T(Q_{t+1} = (\boldsymbol{p}_{t+1}, \boldsymbol{\alpha}_{t+1}) | Q_t = \boldsymbol{q}_t, I_t = \pi_i(\boldsymbol{q}_t)). \quad (5)$$

as the most probable next object state given a certain policy π_i. $\boldsymbol{w}_t^i$ also could be calculated by not only considering one but cumulating multiple steps of the policy until a certain horizon of prediction.

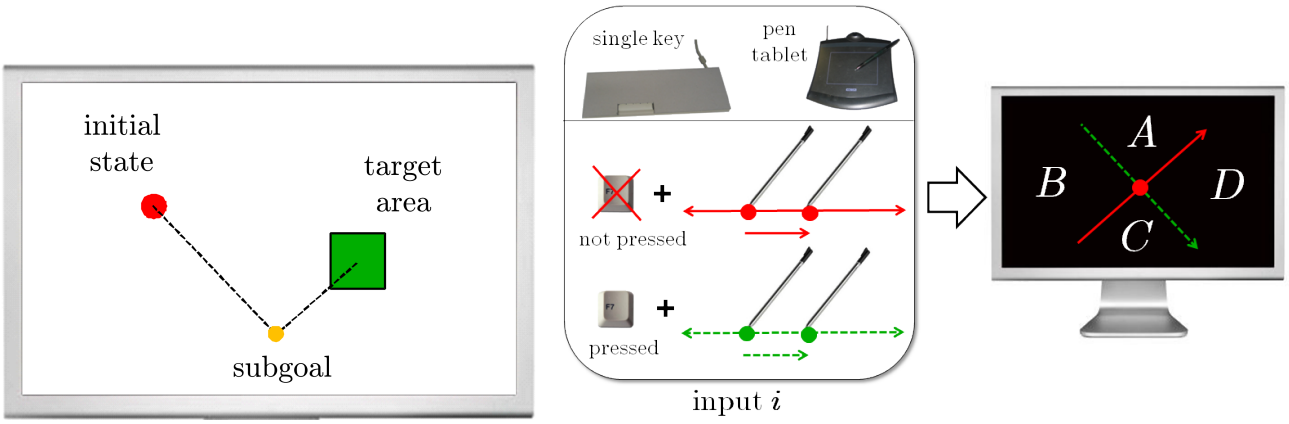
Every fixation further is classified into one of the following categories:

$$P^{f,i} : (\|\boldsymbol{v}_t\| > v_o) \wedge (\measuredangle(\boldsymbol{v}_t, \boldsymbol{w}_t^i) \leq \beta_{max}) \quad (6)$$

$$O^{f,i} : (\|\boldsymbol{v}_t\| \leq v_o) \wedge (\boldsymbol{w}_t^i = \boldsymbol{0}) \quad (7)$$

$$N^f : (\|\boldsymbol{v}_t\| > v_o) \wedge (\forall i : \measuredangle(\boldsymbol{v}_t, \boldsymbol{w}_t^i) > \beta_{max}) \quad (8)$$

$$O^f : (\|\boldsymbol{v}_t\| \leq v_o) \wedge (\forall i : \boldsymbol{w}_t^i \neq \boldsymbol{0}) \quad (9)$$



(a) Task (b) Mapping of inputs to system reactions and target quadrants A,B,C and D

Figure 2. Experimental task (a) and mapping of input devices (b).

Fixations of category $P^{f,i}$ are proactive with respect to policy π_i. Category $O^{f,i}$ contains fixations on objects along a certain policy with no changes in object position, while O^f describes all other fixations on an object. N^f contains all fixations belonging to none of the other categories.

The criteria for the different categories only depend on system states at one single point in time. The time delay for fixation classification induced by a previous approach presented in [1] can therefore be avoided.

Taking only absolute fixation positions into account for gaze analysis may lead to problems if gaze position measurement is subject to drift, as it is often the case with current eye tracking hardware. Therefore, we also classify gaze movements according to their spatial relation to object and target positions as well as to possible policies. We define saccades as $\Delta \boldsymbol{f}_t = \boldsymbol{f}_t - \boldsymbol{f}_{t-1}$ and assign them to one of the following classes:

$$P^{s,i} : (\|\boldsymbol{v}_t\| > \|\boldsymbol{v}_{t-1}\|) \wedge (\measuredangle(\Delta \boldsymbol{f}_t, \boldsymbol{w}_t^i) \leq \beta_{max}) \quad (10)$$

$$P^{s} : (\|\boldsymbol{v}_t\| > \|\boldsymbol{v}_{t-1}\|) \wedge (\forall i : \measuredangle(\Delta \boldsymbol{f}_t, \boldsymbol{w}_t^i) > \beta_{max}) \quad (11)$$

$$R^{s} : (\|\boldsymbol{v}_t\| \leq \|\boldsymbol{v}_{t-1}\|) \quad (12)$$

Categories $P^{s,i}$ and P^s contain proactive, R^s reactive saccades. Too small saccades or those too far away from an object are filtered out by the conditions $\|\Delta \boldsymbol{f}_t\| > \Delta f_{min}$ and $\|\boldsymbol{v}_t\| \leq v_s \vee \|\boldsymbol{v}_{t-1}\| \leq v_s$, where Δf_{min} is the minimal length of a saccade and v_s the maximal distance to an object.

Proactive gaze behavior which can be assigned to a certain policy (categories $P^{f,i}$, $O^{f,i}$ and $P^{s,i}$) can be used for estimating user's intention. Especially for tasks with multiple possible policies the policy which will be chosen by the user could be identified previous to any object movement, allowing for designing new proactive interaction techniques. Additionally, by knowing which policy is chosen by the user according to his/her gaze movements, a model mismatch can be detected if a different policy is actually executed. In order to reduce training time and to resolve the model mismatch a hint could be displayed to the user in such a case. More generally speaking, the interface can automatically adapt to novice or expert users.

Reactive gaze movements are not of value for intention recognition, since they do not convey any additional information not already available through observation of system state changes. However, a high amount of reactive fixations might indicate uncertainty of the user and of the mental model, respectively.

EVALUATION

Participants and Task

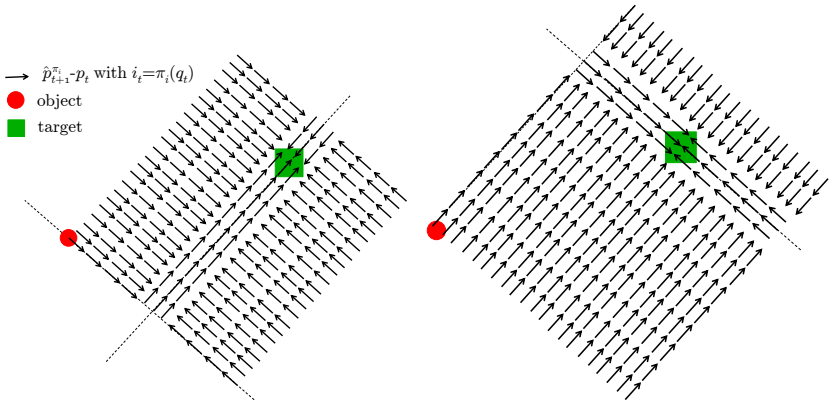
We evaluated the proposed model with a preliminary user study. Four participants were asked to perform an indirect object manipulation task as fast as possible. The goal was to move a point from its initial position into a target area as shown in Figure 2(a). We distinguish between four types of tasks (A,B,C,D) depending on the location of the target area on the screen (see Figure 2(b)).

Apparatus and system model

The size of the display is 12.1 inches with a resolution of 1024×768 pixels. As input devices we use one single key of a keyboard and a pen tablet, while only horizontal movements of the pen on the tablet were interpreted by the system. Hence, we obtain the set of possible inputs $\mathcal{I} = \{0,1\} \times \mathbb{N}$ with elements $\boldsymbol{i} = (i^k, i^d)$, where $i^k \in \{0,1\}$ indicates whether the key is pressed (1) or not (0) and $i^d \in \mathbb{N}$ is the relative movement of the pen in horizontal direction to the left (< 0) or to the right (> 0). The position of the point after a state transition is defined by

$$\boldsymbol{p}_{t+1} = \begin{cases} \boldsymbol{p}_t + i_t^d \cdot \boldsymbol{a}_0 & \text{if } i_t^k = 0, \\ \boldsymbol{p}_t + i_t^d \cdot \boldsymbol{a}_1 & \text{if } i_t^k = 1, \end{cases} \quad (13)$$

where $\boldsymbol{a}_0$ and $\boldsymbol{a}_1$ are the two possible movement directions of the object. The mapping between inputs and system state transitions is graphically illustrated in Figure 2(b). The mapping was chosen to be distinct from any standard mapping potentially already known by the participants in order to be-



(a) Policy π_1 for task D (b) Policy π_2 for task D

Figure 3. Two policies for task of type D with minimal number of direction changes in the movement of the manipulated object.

ing able to study gaze behavior under uncertain mental models $P^M \neq P^T$ of the users.

Figure 3 shows the two policies π_1 and π_2 for a task of type D with minimal number of direction changes of the object. Note that the arrows in Figure 3 are not the inputs delivered by the policy functions, but indicate the direction of movement of the visual representation of a manipulated object induced by a certain input of the policy.

Subgoals for the two policies with minimal number of direction changes of the object movement were indicated for each task as small dots (see Figure 2(a)).

Participants' eye movements were captured by a SMI iViewX™ HED head mounted eye tracking system. For transforming gaze positions into screen coordinates a video-based marker detection and tracking system was used (see [1] for details).

Procedure

The participants had to execute 60 tasks of different types (A,B,C,D) in random order, while the order was equal for all participants. Previous to the first task, every user was told that only horizontal movements of the pen are interpreted by the system and the key of the keyboard has "some additional function". Additionally, users had time to practice interaction with the pen, however, in a separate application. Hence, no mental model could be build about the mapping between input and object movement during practice.

Data Analysis

In the study described here we focus on analyzing gaze movements, which occur previous to any object movement. Such *pre-object gaze movements* are particularly interesting for estimating user's intention and model mismatches.

Both, fixations and saccades are classified as described in previous section. In order to obtain an estimate of user's intention, information about all pre-object gaze movements is cumulated by

$$\hat{\Pi}^{\mathcal{G}} = \{\pi^{\mathcal{G}}_{j_1}, ..., \pi^{\mathcal{G}}_{j_M}\} \tag{14}$$

with

$$j_k \in \{i \in \{1, ..., N_\pi\} | \#P^{f,i} + \#P^{s,i} = \text{maximum}\}, \tag{15}$$

where $\#P^{f,i}$ and $\#P^{s,i}$ denote the number of pre-object proactive fixations/saccades assigned to the respective policy $\pi^{\mathcal{G}}_i$. In our experiments we only consider the two different policies shown in Figure 3 ($\Rightarrow N_\pi = 2$). If there are no proactive gaze movements, intention can not be estimated and $\hat{\Pi}^{\mathcal{G}} = \{\}$. If the number of fixations/saccades assigned to two or more distinct policies is equal ($|\hat{\Pi}^{\mathcal{G}}| > 1$), also no decision for one single policy is possible.

The policy actually taken by the user is determined by aligning the sequence of inputs $\boldsymbol{i}_{t_1}, ..., \boldsymbol{i}_{t_{N_I}}$ with the different policies. We calculate a score s_i for every policy π_i of a given task according to

$$s_i = \sum_{k=1}^{N_I} < \boldsymbol{p}_{t_k+1} - \boldsymbol{p}_{t_k}, \hat{\boldsymbol{p}}^{\pi_i}_{t_k+1} - \boldsymbol{p}_{t_k} > . \tag{16}$$

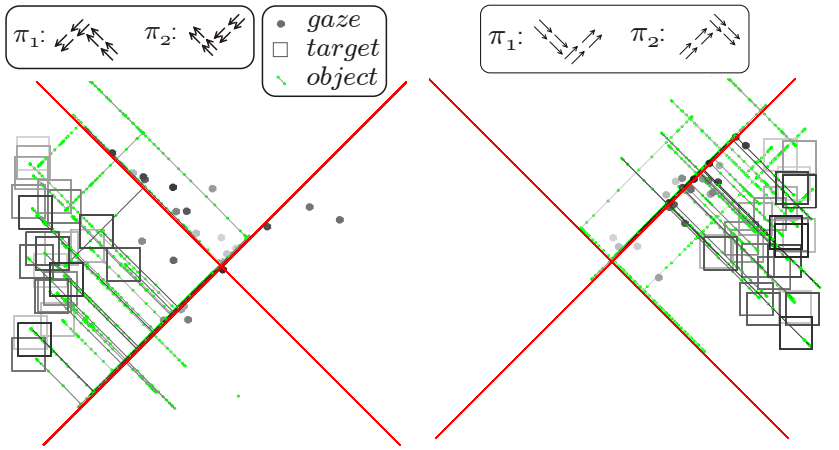
The policy $\pi^{\mathcal{G}}_m$ ($m = \underset{i}{\text{argmax}}\, s_i$) with the highest score is considered to be the policy, which was chosen by the user for executing the task.

We assume a mismatch between the mental model of the user and the real system if the estimated user intention differs from the policy actually taken:

$$\text{model mismatch} := (\hat{\Pi}^{\mathcal{G}} \neq \{\}) \wedge (\pi^{\mathcal{G}}_m \notin \hat{\Pi}^{\mathcal{G}}) \tag{17}$$

Results

Figures 4 and 5 show target areas, object movements, and the last pre-object gaze positions in each task for two different participants. Data from earlier tasks in the experiments are drawn in light gray, later tasks in dark gray and black, respectively. In the left figures only data from tasks of type B and in the right figures only from tasks of type D are shown.



(a) Task of type B (b) Task of type D

Figure 4. Object paths and last pre-object gaze positions for "*User 1*" showing predominantly *proactive gaze behavior* for tasks of type B (a) and D (b)

For both of the two different tasks the policies used by the user converged towards a certain policy. However, gaze behavior differed significantly for the different users and for the tasks.

Figure 4 shows data captured from a user ("*User 1*") who used *proactive gaze behavior*. In the right Figure 4 (b) proactive gaze positions are distributed along policy π_2, which is the policy most frequently chosen by the user. In the left Figure 4 (a) gaze positions are also proactive, but most of them are distributed along π_1, while the most frequently used policy is π_2. Hence, the figure for task B indicates a model mismatch, since pre-object gaze movements indicate that a wrong system reaction was anticipated by the user.

The user who's data is shown in Figure 5 ("*User 2*") almost exclusively used *reactive gaze behavior* at the beginning of the task sequence. For both tasks last pre-object gaze positions are concentrated around the initial object position. Additionally, object paths in Figure 5 show much more deviation from one of the two "optimal paths" compared to those in Figure 4. This indicates that the user with proactive gaze behavior has built a more accurate mental model than the user with reactive gaze behavior or, the other way round, the user with the accurate model uses proactive gaze behavior while the user with the inaccurate model uses a reactive one. At the end of the task sequence for task D "*User 2*" switched from reactive to proactive gaze behavior (see Figure 5(b)). This also supports the proposition that a well established mental model build over time leads to more proactive gaze behavior.

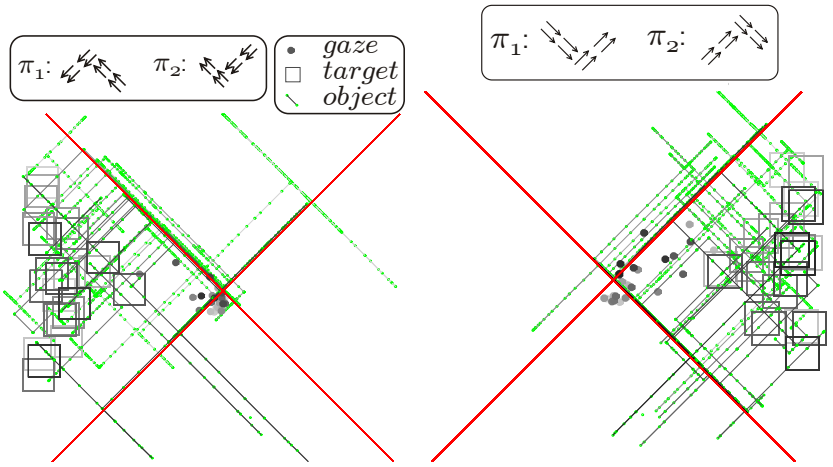
In order to evaluate the proposed model more formally all pre-object fixations and saccades are classified as described in previous sections. Further, user's intention is estimated as the set of most probable policies $\hat{\Pi}^{\mathcal{G}}$ based on the gaze data and is compared with the true policy $\pi_m^{\mathcal{G}}$ chosen by the user. Note that for calculating $\hat{\Pi}^{\mathcal{G}}$ not only last pre-object fixation as shown in Figure 5 and 4 is considered, but all pre-object fixations and saccades are used. The results of this evaluation for data collected from four participants are shown in Figure 6.

Pre-object gaze movements of *User 1* and *User 4* deliver good estimates of the policy actually chosen by the user for task D. The amount of tasks of type D with $\hat{\Pi}^{\mathcal{G}} \neq \{\}$ is 95.83% for *User 1* and 66.67% for *User 4*. For task B this percentage is lower for *User 1* (52.12%) and similar for *User 4* (69.57%). However, especially *User 4* shows many proactive fixations and saccades indicating a different policy than the one actually chosen (see Figure 6(a)). This indicates the existence of a model mismatch for task B. For *User 2* the number of tasks with $\hat{\Pi}^{\mathcal{G}} \neq \{\}$ increased with the time from 50% for task D (34.78% task B) in the first half of the task sequence to 75% (60.87% task B) in the second half (complete task sequence: 62.5% task D, 47.83% task B). However, as we can see in Figure 6(b), *User 2* produced a lot of mismatches between gaze and actual object movement at the end of the task sequence. *User 3* shows a similar development of gaze behavior for task B as *User 2*, however, with a lower amount of tasks with proactive gaze behavior.

Discussion

The above results show that the proposed model provides a good basis for analyzing gaze data in the context of a task. The preliminary study for evaluating our framework already revealed some interesting findings which need to be verified in further, more extensive studies.

Building the correct mental model for task B seems to be more difficult than for task D. All of the four participants showed more mismatches for task B. Also more proactive fixations and saccades could be observed for task D than for task B. The difficulties users had with task B could be explained by a wrong mental model about the functionality of the key, namely that the key determines whether the point moves upwards or downwards instead of along one of the two diagonal axes as shown in Figure 2. However, when



(a) Task of type B (b) Task of type D

Figure 5. Object paths and last gaze position previous to first object movement for *User 2* showing predominantly *reactive gaze behavior* for tasks of type B (a) and a switch from *reactive to proactive behavior* for tasks of type D (b).

asked after the experiments, all users explained the functionality of the key correctly. Gaze behavior seems to be a more reliable measure for the quality of the mental model and for how well it is established than subjective statements.

Gaze behavior strongly varied among different users. However, also similarities between *User 1* and *User 4* as well as *User 2* and *User 3* could be observed. The differences between the two groups of users seem to be correlated with the quality and certainty of the user's mental model. This dependency, if verified in further studies, could be used for automatically adapting the human-computer interface according to the quality of the user's mental model. For users with a correct mental model proactive gaze behavior could be used for intention recognition and for realizing proactive interaction techniques. Users with an uncertain or wrong mental model could be supported by adequate hints, which help to improve or correct their mental representation of the system.

Based on the considerations from the beginning of this paper we would expect the user to look at a certain location on the screen for one or both of the two reasons: "control of input" and/or "verification of system reactions or states". The results of our study can be interpreted in the following way: the proactive gaze behavior of "experienced" users is caused by a switch from verification of system reactions for learning purposes to indirect control of absolute positioning of the object at key points of the task.

The results also show, considering only few of all of the possible policies for the task leads already to a good alignment of gaze data with the task. This is particularly important for generalization of the proposed framework to more complex tasks with larger state or action space or number of goals.

CONCLUSION

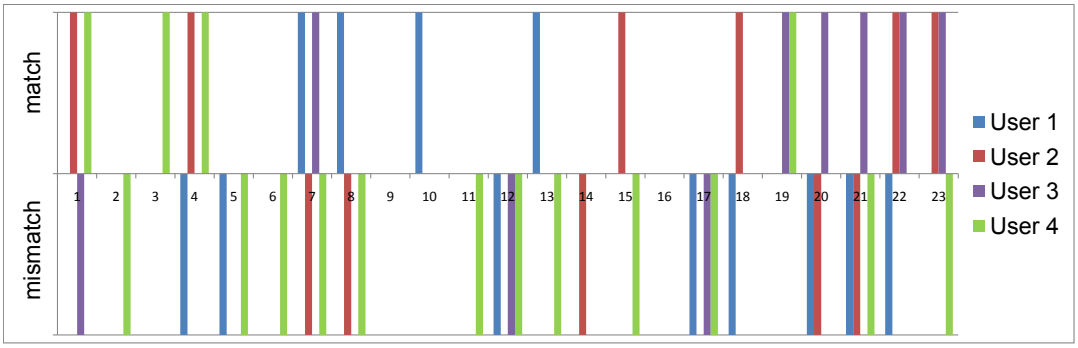
The proposed model allows for analysis of gaze data in the context of a task currently executed by the user. Uncertainty of the mental model of the user can be modeled and therefore it also provides a basis for studying the influence of learning processes on natural gaze behavior. We have demonstrated how it can be used to provide an explanation for "why we look where we look" and as a framework for analyzing gaze data in dynamic contexts.

The evaluation of the framework in a preliminary user study has shown that various kinds of visual control strategies of the user can be captured by the model and thus can be made available for further analysis. In our study gaze behavior was both, task- and user-specific. The proposed framework allows for formal description of the different behaviors and enables model-based explanations of the differences between them. The study also revealed that not only a correct but also a wrong mental model leads to proactive gaze behavior. We have demonstrated that such a model mismatch can be detected with the proposed framework at an early stage, which allows for designing *adaptive systems* providing help to unexperienced users automatically.

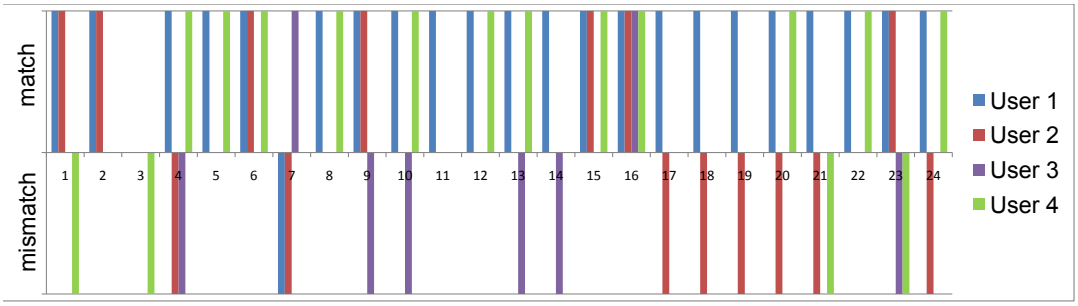
In future work the proposed model needs to be evaluated on larger data sets and with more complex tasks in order to validate its general value. In this paper we have only focused on pre-object gaze data. However, the model also can be used for analyzing gaze data during object movement phases which contain further valuable information about user's cognitive processes. Finally, new user interfaces based on the proposed model need to be implemented and evaluated. In this context the proposed model also provides a good basis for real-time recognition of user's intention and cognitive states.

REFERENCES

1. T. Bader, M. Vogelgesang, and E. Klaus. Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In *Proceedings of ICMI-MLMI*, pages 199–206. ACM, Nov. 2009.



(a) Task of type B



(b) Task of type D

Figure 6. Comparison of estimated policy and true policy chosen by the users for all tasks of type D. Bars above the horizontal axis indicate a match of estimated intention and true policy, bars below indicate a mismatch. Missing bars indicate that no intention estimation was possible.

2. L. E. Brown, D. A. Rosenbaum, and R. L. Sainburg. Limb position drift: Implications for control of posture and movement. *Journal of Neurophysiology*, 90:3105–3118, 2003.

3. R. Flanagan and R. S. Johansson. Action plans used in action observation. *Nature*, 424:769–771, 2003.

4. B. Gesierich, A. Bruzzo, G. Ottoboni, and L. Finos. Human gaze behaviour during action execution and observation. *Acta Psychologica*, 128:324–330, 2008.

5. A. Hyrskykari, P. Majaranta, and K. Räihä. Proactive response to eye movements. In M. Rauterberg, editor, *Human Computer Interaction INTERACT 2003*, pages 129–136. IOS Press, September 2003.

6. R. S. Johansson, G. Westling, A. Bäckström, and J. R. Flanagan. Eye-hand coordination in object manipulation. *The Journal of Neuroscience*, 21(17):6917–6932, 2001.

7. B. E. John. Extensions of goms analyses to expert performance requiring perception of dynamic visual and auditory information. In *CHI '90: Proceedings of the conference on Human factors in computing systems*, pages 107–116, New York, NY, USA, 1990. ACM.

8. M. F. Land and D. N. Lee. Where we look when we steer. *Nature*, 369:742–744, 1994.

9. M. F. Land and P. McLeod. From eye movements to actions: how batsmen hit the ball. *Nature Neuroscience*, 3:1340–1345, 2000.

10. C. Lankford. Effective eye-gaze input into windows. In *ETRA '00: Proceedings of the symposium on Eye tracking research & applications*, pages 23–27, New York, NY, USA, 2000. ACM.

11. J. Pelz, M. M. Hayhoe, and R. Loeber. The coordination of eye, head, and hand movements in a natural task. *Exp Brain Res*, pages 266–277, 2001.

12. D. D. Salvucci and J. R. Anderson. Intelligent gaze-added interfaces. In *CHI '00: Proceedings of the conference on Human factors in computing systems*, pages 273–280. ACM Press, 2000.

13. D. D. Salvucci and J. R. Anderson. Automated eye-movement protocol analysis. *Hum.-Comput. Interact.*, 16(1):39–86, 2001.

14. D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *ETRA '00: Proceedings of the symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

15. B. A. Smith, J. Ho, W. Ark, and S. Zhai. Hand eye coordination patterns in target selection. In *ETRA '00: Proceedings of the symposium on Eye tracking research & applications*, pages 117–122. ACM, 2000.

16. S. Zhai, C. Morimoto, and S. Ihde. Manual and gaze input cascaded (magic) pointing. In *In CHI99*, pages 246–253. ACM Press, 1999.