Multimodal Integration of Natural Gaze Behavior for Intention Recognition During Object Manipulation

Thomas Bader
Universität Karlsruhe
Institute for Anthropomatics
76131 Karlsruhe, Germany
bader@ies.uni-karlsruhe.de

Matthias Vogelgesang
Fraunhofer IITB
Fraunhoferstrasse 1
76131 Karlsruhe, Germany
vogels@iitb.fraunhofer.de

Edmund Klaus
Fraunhofer IITB
Fraunhoferstrasse 1
76131 Karlsruhe, Germany
klaus@iitb-extern.fraunhofer.de

ABSTRACT

Naturally gaze is used for visual perception of our environment and gaze movements are mainly controlled subconsciously. Forcing the user to consciously diverge from that natural gaze behavior for interaction purposes causes high cognitive workload and destroys information contained in natural gaze movements. Instead of proposing a new gaze-based interaction technique, we analyze natural gaze behavior during an object manipulation task and show ways how it can be used for intention recognition, which provides a universal basis for integrating gaze into multimodal interfaces for different applications. We propose a model for multimodal integration of natural gaze behavior and evaluate it for two different use cases, namely for improvement of robustness of other potentially noisy input cues and for the design of proactive interaction techniques.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Human Factors, Experimentation, Theory

1. 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 goals. However, in most state-of-the-art interfaces gaze is only used as an explicit pointing device, e.g. as a 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, using gaze-based pointing as a general input technique for human computer interaction (HCI) has many limitations. First, forcing the user to stare consciously at a certain location on a display to trigger a desired action of the system is often at odds with his/her natural gaze behavior. Hence, such usage of gaze as input modality causes high cognitive workload. Second, by preventing the user from using natural gaze behavior, valuable information contained in it (e.g. the user's intention) is lost and cannot be used for interaction purposes. Third, using gaze as sole input modality leads to the so-called "Midas-Touch" problem[6]. Everything we look at immediately changes its state, even if we only want to perceive it's current state.

In order to allow for a design of more sophisticated gaze-based interaction techniques in a multimodal context, we analyzed natural gaze behavior during an object manipulation task and propose a model which allows for integration of it as an additional modality. In contrast to other studies on natural gaze behavior we chose a large scale display as experimental platform. We consider gaze as particularly interesting for interaction with large scale screens, since large spatial distances need to be bridged during interaction and in some setups not all regions are within the grasping range of the user but within his/her field of view. Gaze provides promising properties for interaction with distant objects and for covering large spatial distances without major physical fatigue.

Numerous studies of natural gaze behavior and hand-eye coordination during manipulative activities in natural environments like block-copying [12], basic object manipulation [7], driving [8] and playing cricket [9] revealed gaze shifts and fixations to be commonly proactive (eye-movements occured previous to movements of the manipulated object or the manipulator). In addition, a detailed study on hand-eye coordination during an object manipulation task [7] revealed, that subjects almost exclusively fixated landmarks critical for the control of the task and never the moving object or hand. Such landmarks could be obstacles or objects in general that are critical for the completion of the task, like in [9] where batsmen concentrated on the ball, and not on their hands or the bat. These studies show, that natural

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

gaze behavior is complex and determined by many different parameters (e.g. position of obstacles in [7] or previous experience of a person [9]).

Gaze behavior was also studied in various tasks related to HCI. However, in contrast to our experiment these studies were conducted with indirect input devices (e.g. [15]). Other studies from the field of psychology and physiology, e.g. [5, 4] investigated differences in gaze behavior during action execution and observation. They distinguished three different gaze behaviors, namely proactive, reactive and tracking gaze behavior [5].

In all of the above studies natural gaze behavior numerous different gaze patterns were observed during task execution and were described informally. To our knowledge no model exists which describes the causal relations between natural gaze behavior and other input modalities or system states. However, an understanding of the reasons why a person looks at a certain location in a certain situation is necessary to judge the usefulness of natural gaze behavior for HCI and to integrate gaze with other modalities, respectively.

In this paper we develop a model which systematically describes the causal relationship between gaze and other input modalities during an object manipulation task and use it for intention recognition. We chose an object manipulation task for our experiments since this is one of the most important basic tasks in HCI. We first describe a user study we conducted in order to get an impression of natural gaze behavior during direct object manipulation at a large horizontal screen. The observed gaze patterns are categorized systematically and compared with observations made in other interactive environments already described in literature. On this basis a model describing the causal relations between natural gaze behavior and the context of interaction is developed and formally described by means of a Bayesian net. Finally we illustrate how the model can be used for two use cases, namely for improvement of robustness and efficiency of video-based input devices and for designing proactive multimodal user interfaces.

2. NATURAL GAZE BEHAVIOR DURING OBJECT MANIPULATION

To understand and formally characterize natural gaze behavior during object manipulation at large screens we conducted a user study in a tabletop scenario as shown in Figure 1.

2.1 Method

2.1.1 Participants

Eleven subjects (2 female, 9 male) participated in the study. All participants were right-handed and none of them wore glasses or contact lenses. All of them were experienced computer users, although no one had experience on direct-touch interaction with large scale tabletop displays.

2.1.2 Apparatus

During the experiments participants stood in front of the tabletop with a horizontal 90x120 cm sized display (Figure 1). The display is realized as back projection with a resolution of 1024x768 pixels. This resolution leads to a pixel size of 1.17x1.17 mm. The working area was restricted to a rectangle with the size of 824x500 pixels in the center of the horizontal display in order to be easily reachable by the user.

As input device we used a vision-based hand gesture recognition system [1]. The system is able to detect and distinguish between several hand symbols above the tabletop as well as to detect a touch of the display surface by the user. We only used simple pointing gestures as input during the experiments, which were captured at a rate of 25 Hz. The system has a latency of 1-2 video frames.

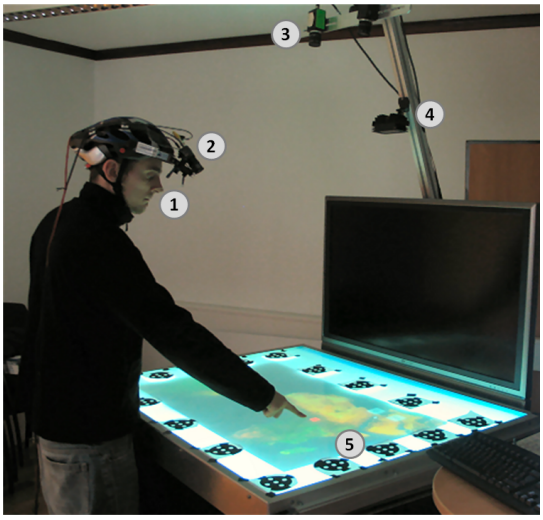

Figure 1: Experiment participant at the tabletop. (1: eye tracker, 2: scene camera, 3: gesture recognition cameras, 4: infrared lighting, 5: coded markers)

Participants' eye movements were captured by a SMI iViewX™ HED [14] head mounted eye tracking system at a sampling rate of 50 Hz. The SMI iViewX™ system delivers a video of the surrounding scene and gaze positions in pixel coordinates of the scene video. To enable a mapping between gaze positions and location of user interface items displayed on the tabletop surface, a marker tracking system was connected to the scene video camera of the eye tracking system. Coded markers were attached to the edges of the tabletop display to determine its position in the scene video (see Figure 1).

2.1.3 Tasks

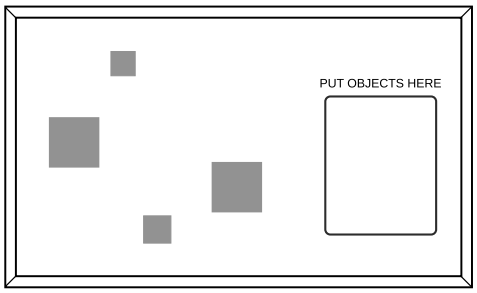

Figure 2: Example of an experimental task.

Every participant had to perform simple object manipulation tasks. At the beginning, squares of different sizes and a rectangular target area were displayed on the tabletop surface (see Figure 2). Squares were placed with different distances to the target area. The goal was to move the squares to the target area and place them there. The squares disappeared as they were placed in the target area. The task was successfully completed if all squares were placed in the target area. The same task with varying initial positions and number of squares was repeated four times by every participant. The position of a square could be manipulated by touching it with the index finger and dragging it to the desired location. As soon as the index finger was lifted, the object was deselected and disappeared if placed within the target area. This interaction cycle, consisting of three consecutive phases, namely *selection*, *manipulation* and *deselection* of a graphical object, is illustrated in Figure 3. Successful termination of each phase is indicated by appropriate system feedback. To indicate that an object is selected a red frame surrounding the object is displayed.

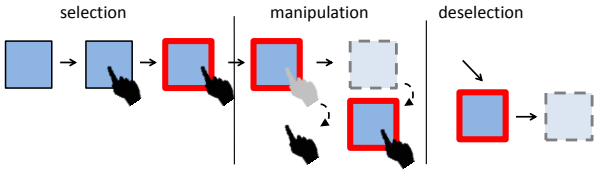

Figure 3: Schematic illustration of a direct hand gesture based interaction cycle.

2.1.4 Procedure

Each participant first was introduced to the use of the tabletop. This was followed by the calibration of the eye tracker and two tests to detect the accuracy of the eye tracker and the mapping of gaze positions to user interface items. The average accuracy of the eye tracker for all participants was between 33.9(29) and 44.0(38) mm(px) in x- and 51.2(44) and 73.2(63) mm(px) in y-direction. The average completion time of all tasks was about 6 minutes per person.

2.1.5 Data Analysis and Classification

In order to understand and systematically characterize the gaze behavior of participants we analyzed the recorded hand and object movements, as well as the gaze data. Gaze movements consist of 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 [13]. The first algorithm clusters gaze points according to their spatial distribution, the second one according to the velocity of gaze movements. An interaction cycle as described before starts with the selection of an object and ends with the deselection of this object at the target area. First examinations of recorded data showed that fixations during an interaction are mainly determined by object movements and not by hand movements: None of the participants ever fixated the hand during object manipulation. We assume that the reason is the proprioceptive feedback of the hand position that mediates an approximate anticipation where the hand is located. In the remainder of this paper we therefore only focus on the relation between object and fixation positions and do not consider hand-object or hand-eye relations any further.

A fixation normally lasts about 150 to 600 ms [3]. During this time a manipulated object may change its state (selected/unselected) or its position, respectively, while gaze position remains at a fixed location. According to fixation-object relations at the beginning and end of a fixation we assign a fixation to one of the following five categories for analysis purposes:

- O_0: Gaze and object position is similar during the whole fixation. State of object does not change during the fixation.
- O_c: Gaze and object position is similar during the whole fixation. State of object changes during the fixation (e.g. selection).
- P (Proactive): Gaze and object position is different at the beginning of the fixation. At the end of the fixation object and gaze position is similar / closer.
- R (Reactive): Gaze and object position is similar at the beginning of the fixation. At the end of the fixation object position is different from gaze position.
- N: Gaze and object position is different during the whole fixation. State of object does not change during the fixation or moves away from gaze position.

Note that in the above definitions with "object" we denote that object, which is closest to the gaze position. The criteria above are formulated generically to be also applicable to other tasks for categorization of fixations. The notions "similar" and "closer" as well as the categorization criteria are formally defined in 3.3.

2.2 Results of the Study

We evaluated the gaze behavior of participants for two phases of the task, namely selection and manipulation, with respect to the number and duration of fixations as well as the distribution of fixations between the different categories defined in 2.1.5.

In Figure 4 the mean number of fixations over all participants and tasks is shown. The distance between the initial location of an object to its target position seems to have a significant influence on the number of fixations made during the manipulation phase. This dependency was not described in any previous publication. Probably due to significant smaller displays used in earlier studies (e.g.[15, 11]) this effect could not be observed or was significantly smaller.

The duration of fixations varied strongly during task execution. Fixations directly preceding the selection of a certain object lasted significantly longer than fixations during visual exploration of the GUI or during manipulation phase. The end of fixations which comprise an object state change (O_c,R,P) correlates with the point in time when the visual feedback about the changed state of the object was displayed (e.g. new position or red frame for indicating a selection).

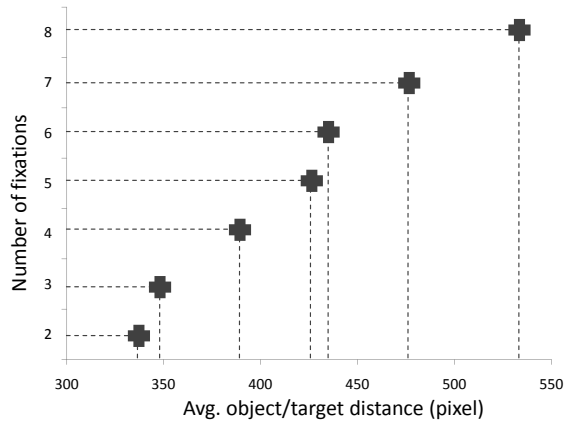

Figure 4: Relation of fixation positions and the distance between initial object position and target area.

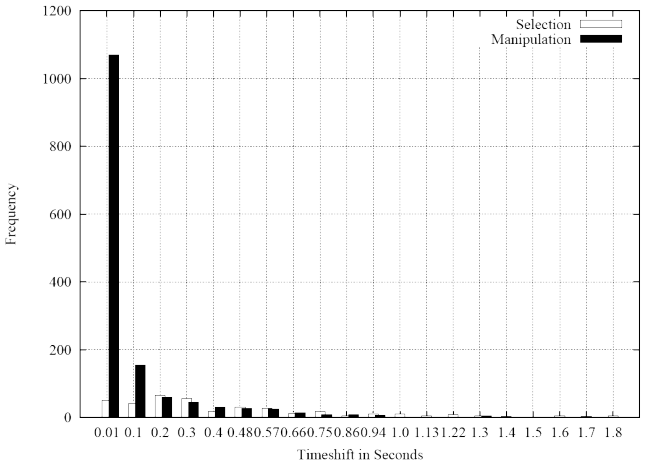

Figure 5: Distribution of time shift between end of fixation and change of object state.

The distribution of the time shift between the end of a fixation and the state change of an object for all fixations of type O_c,R and P is shown in Figure 5.

The frequency of the different fixation types during selection and manipulation phase are shown in Table 1. During selection phase only O_0-, O_c and N fixations were observed, during manipulation mainly P- and R-fixations. However, during manipulation not only one type of fixations was used constantly over the whole phase. For further analysis we denote the different gaze patterns during manipulation phase consisting of one or multiple fixations with $[R]_i$ (only R-fixations), $[P]_i$ (only P-fixations), $[R \rightarrow P]_i$ (switch from R- to P-fixations), $[P \rightarrow R]_i$ (switch from P- to R-fixations) and $[m]_i$ (any other mixed pattern). The index $i \geq 1$ denotes the number of fixations. In Figure 6 the frequencies of the different patterns are shown.

Reactive gaze behavior $[R]_i$ and the proactive pattern $[P]_1$ with only one fixation at the target area have been already observed in other studies [15, 11]. Consistently, we call these two patterns *object following* and *target gaze* behavior. In our study we observed several appearances of patterns with more than one proactive gaze switch during manipulation phase ($[P]_{i>1}$) which we denote with *stepwise proactive* behavior. Such patterns have also been reported in [5], however they were not explicitly distinguished from *target gaze* behavior. In [15, 11] *stepwise proactive* behavior was not observed at all. We think this is mainly due to the limited display size compared to the large scale tabletop we used in our study, which is strongly supported by the results shown in Figure 4, namely the influence of object-target distance on the number of fixations made during an interaction cycle.

Fixation type	O_0	O_c	N	P	R
Selection	111	42	2140	0	0
Manipulation	25	78	0	1623	1035

Table 1: Occurrence of different fixation types during selection and manipulation.

Worth to mention is that the usage of the different patterns is highly user dependent. For example 3 out of 9 users did not use the "target gaze" pattern at all.

The direction of saccades after reactive R-fixations are determined by the new position of the manipulated object at the end of the fixation. Unlike the gaze position during proactive P-fixations, which is chosen freely by the user without having a visual reference to look at. However, as illustrated in Figure 7, gaze movements previous to a proactive fixation are mainly performed in manipulation direction of the object and towards the target area.

These results have the following implications for using gaze as an input modality for the basic tasks selection and manipulation as described above.

- An object is always fixated previous to a selection. Fixations previous to a selection last longer than fixations during visual exploration. Therefore the intention for selection of a certain object could be estimated from gaze data and could be used as basis for implementing proactive interaction techniques.
- Natural gaze behavior during direct manipulation highly varies for different users and situations. To use natural gaze behavior as an additional modality the reasons behind the individual gaze patterns have to be understood and the system has to react according to the anticipation of the user.
- Natural gaze behavior may change during the manipulation phase. Such switches implicitly could convey interesting information to be used for HCI.

3. A MODEL FOR INTEGRATION OF NATURAL GAZE BEHAVIOR

In the following section we propose a model which describes the causal relations between natural gaze behavior and the context of interaction. This model provides the basis for integrating gaze with other input modalities and also provides a possible explanation for each of the different gaze patterns described in the previous section.

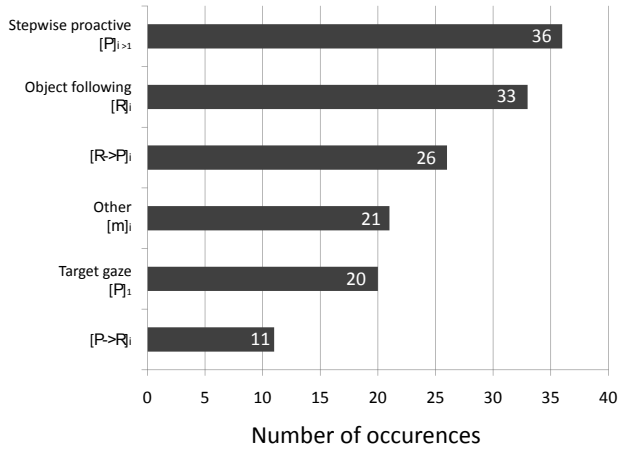

Figure 6: Number of occurrences of gaze behavior patterns in the study.

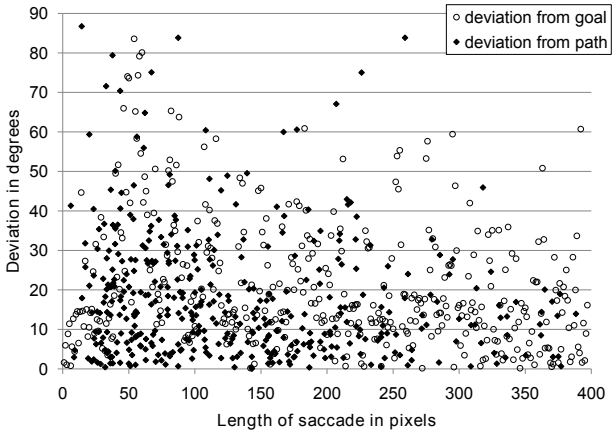

Figure 7: Relation of the length of a proactive saccade and the deviation of gaze position from actual object path and target position.

3.1 Fundamental Causal Relations

When interacting with a system the user usually has a certain goal in mind. In our case this is the selection and positioning of objects displayed on the tabletop as described in section 2.1.3. In order to reach a goal or sub-goal (e.g. selection of the object to be manipulated), the user has to perform certain actions which are captured by input devices, in our case the gesture recognition system. If a certain action was performed the user either verifies whether the reaction of the system to the input conforms with the user's model of the system or whether the goal or sub-goal was reached (e.g. object is at desired position). We call these two different behaviors action- and goal-directed verification (see Figure 8).

The feedback from the system in our experiment was mainly encoded in the visual channel. By measuring gaze movement and position we can, therefore, infer where the user perceived this visual feedback or where she/he expected visual feedback to be displayed.

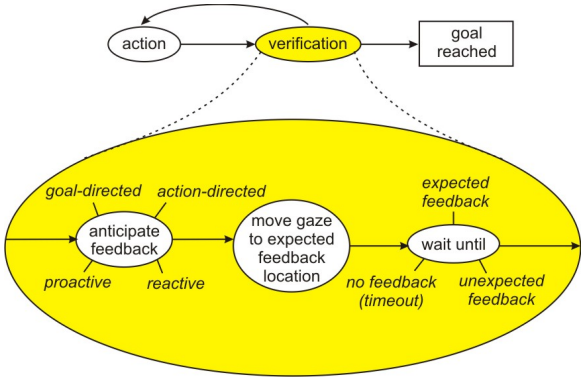

Figure 8: A simple model describing fundamental causal relations between gaze behavior and the context of interaction.

During reactive gaze behavior the user is only able to verify, if an object moved away from a certain location on the display, possibly also into which direction in the peripheral field of view. During proactive behavior the user anticipates a certain system reaction and object state (e.g. its position), respectively, and acts proactively for verification. This enables the user to verify, e.g. during a movement phase, if the object has moved at all (like it is also possible with reactive gaze behavior) *and* if the object moved to the expected location according to the user's model of the system. Therefore, gain of information for the user is higher with proactive verification behavior than with reactive verification. On the other hand, more knowledge about the system is necessary in order to verify system reactions proactively.

The results presented in the previous section strongly support this model and interpretation in the following ways. Both, action- and goal-directed verification behavior were observed during the movement phase of the task. While gaze pattern "target gaze" is proactive and goal-directed, "stepwise" pro- and reactive patterns contain action-directed verification steps. In accordance with observations made during task execution in a natural environment [11], goal-directed "target gaze" patterns occured more often during interaction when the initially inexperienced users got used to the system. This supports the proposition above, that more knowledge about the system is needed for that kind of verification behavior. The same holds for switches from proactive to reactive verification behavior during an interaction cycle which are characterized by $[P \rightarrow R]_i$ gaze patterns. While duration of P-fixations over all experiments and participants was about 320 ms, the last P-fixation before the switch to a reactive pattern in $[P \rightarrow R]_i$ in average took about 200 ms longer (520 ms). This indicates that a delayed system feedback, thus unexpected feedback was the reason for switching from proactive to reactive verification behavior. It also indicates that loss of trust into the system and the mental model of the system, due to unexpected or missing feedback, leads to a transition from proactive to reactive verification behavior.

The model of the verification process as illustrated in Figure 8 is also supported by the fact, that the end of fixations highly correlate with changes of object states and the corresponding visual feedback, respectively.

3.2 A Probabilistic Model for Integration

According to the model above, the fundamental link between natural gaze behavior and other input modalities is that gaze is used for verification of goals and actions, either proactively or reactively. In order to allow for integration of natural gaze with other input modalities in this section we describe a formal probabilistic model which contains all of the involved components and their interrelationships. Figure 9 gives an overview of the different random variables of the model and their inter-dependencies.

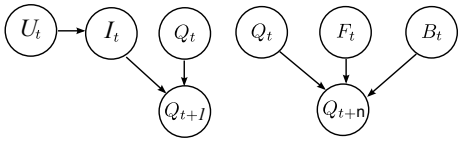

Figure 9: Conditional dependencies of user input, system states and gaze.

Q_t and Q_{t+n} are discrete random variables describing object states prior and after n manipulation steps. In our case an object's state is represented as $\boldsymbol{q} = (\boldsymbol{p}, \alpha)$ with $\boldsymbol{p} \in \mathbb{N} \times \mathbb{N}$ and $\alpha \in \{0, 1\}$, where $\boldsymbol{p}$ denotes the current position in display coordinates and α indicates whether the object is selected ($\alpha = 1$) or not ($\alpha = 0$).

We denote $\mathcal{Q}$ as the set of all possible object states and $\mathcal{G} \subseteq \mathcal{Q}$ as the set of all states which represent a target state. In our case we have separate sub-goals for each of the three phases, namely selection, manipulation and deselection, which can be defined as

$$\mathcal{G}_{\text{sel}} = \{\boldsymbol{q} | \boldsymbol{p} = \boldsymbol{p}_0 \wedge \alpha = 1\} \quad (1)$$

$$\mathcal{G}_{\text{move}} = \{\boldsymbol{q} | \boldsymbol{p} \in \mathcal{T} \wedge \alpha = 1\} \quad (2)$$

$$\mathcal{G}_{\text{desel}} = \{\boldsymbol{q} | \boldsymbol{p} \in \mathcal{T} \wedge \alpha = 0\} \quad (3)$$

where $\mathcal{T}$ describes all points in the target area and $\boldsymbol{p}_0$ is the initial position of an object. Note that the sets $\mathcal{G}_{\text{move}}$ and $\mathcal{G}_{\text{desel}}$ are the same for all objects. The subgoal $\mathcal{G}_{\text{sel}}$ is different for each object, due to different initial positions.

F_t is a random variable with values in $\mathbb{N} \times \mathbb{N}$ describing a fixation position in display coordinates at time $t \in [t_s, t_e]$, where t_s is the start and t_e the end of the fixation. As described in previous section 3.1 by combining the fixation position F_t with the underlying verification behavior B_t with values in $\{goal, action\} \times \{proactive, reactive\}$ we can estimate the visual feedback anticipated by the user and the corresponding object state in a future time step $t + n$, respectively. The calculation of B_t, namely the classification of gaze behavior is described in section 3.3 and the calculation of Q_{t+n} in section 3.4.

According to Card [2] an input device can be defined as a translation mechanism which maps physical properties of the real world to logical values in an application. However, the transformed signal in some cases is affected by noise, like wrongly classified hand gestures in our video-based recognition system. Therefore we describe the interrelationship between the input expressed by the user U and the input recognized by the system I with the conditional probability distribution $P(I \mid U = \boldsymbol{u})$. This distribution describes the characteristic of an input device and can be determined empirically.

The mapping between a measured input $\boldsymbol{i}$ and the system reaction should be deterministic in most systems. For example, pressing the left mouse button is always mapped to a "mouse button pressed" event at the current position by the underlying operating system, which in turn changes some internal application state. Therefore a state transition induced by input $\boldsymbol{i}_t$ at time t can be described as

$$P(Q_{t+1} = \boldsymbol{q}^i_{t+1} \mid I_t = \boldsymbol{i}_t, Q_t = \boldsymbol{q}_t) = 1 \text{ and} \quad (4)$$

$$P(Q_{t+1} \neq \boldsymbol{q}^i_{t+1} \mid I_t = \boldsymbol{i}_t, Q_t = \boldsymbol{q}_t) = 0, \quad (5)$$

where $\boldsymbol{q}_t$ denotes a certain initial state and $\boldsymbol{q}^i_{t+1}$ the state after the state transition.

3.3 Classification of Gaze Behaviour

The gaze behavior B_t at time t is estimated on the basis of a fixation $\boldsymbol{f}_t$. For every fixation we decide whether it is pro- or reactive and action- or goal-directed. First, we classify a fixation $\boldsymbol{f}_t$ according to its spatial relation to the closest object into the five categories described in 2.1.5. While the categories are described in 2.1.5 we define them formally in this section.

First we define the spatial relation $\boldsymbol{v}_{t_s} = \boldsymbol{f}_{t_s} - \boldsymbol{p}_{t_s}$ between the gaze position and the position of the closest object at the beginning of the fixation (see Figure 10). The same relation $\boldsymbol{v}_{t_r} = \boldsymbol{f}_{t_r} - \boldsymbol{p}_{t_r}$ is calculated at a reference point $t_r \in (t_s, t_e]$ during or at the end of the fixation. For off-line analysis of gaze behavior as described in 2.1.5 we used the end of the fixation as reference point ($t_r = t_e$). However, for on-line intention recognition we want to classify a fixation as early as possible. Therefore, for on-line fixation classification the reference point could be either the time when the object state changed or a value bound by an empirically determined threshold. The following list shows the different formal conditions to be fulfilled for a fixation $\boldsymbol{f}_t$ to be assigned to the respective category:

$$O_0 : (\|\boldsymbol{v}_{t_s}\| \leq v_0) \wedge (\|\boldsymbol{v}_{t_r}\| \leq v_0) \wedge (\boldsymbol{q}_{t_s} = \boldsymbol{q}_{t_r}) \quad (6)$$

$$O_c : (\|\boldsymbol{v}_{t_s}\| \leq v_0) \wedge (\|\boldsymbol{v}_{t_r}\| \leq v_0) \wedge (\boldsymbol{q}_{t_s} \neq \boldsymbol{q}_{t_r}) \quad (7)$$

$$P : (\|\boldsymbol{v}_{t_s}\| > v_0) \wedge (\langle \boldsymbol{v}_{t_s}, \boldsymbol{p}_{t_r} - \boldsymbol{p}_{t_s} \rangle > 0) \quad (8)$$

$$R : (\|\boldsymbol{v}_{t_s}\| \leq v_0) \wedge (\|\boldsymbol{v}_{t_r}\| > v_0) \quad (9)$$

$$N : (\|\boldsymbol{v}_{t_s}\| > v_0) \wedge (\langle \boldsymbol{v}_{t_s}, \boldsymbol{p}_{t_r} - \boldsymbol{p}_{t_s} \rangle \leq 0) \quad (10)$$

In equations for P- and N-fixations $\langle \cdot \rangle$ denotes the standard scalar product. A threshold value v_0 specifies when a fixation is considered as being on an object. In our study v_0 was set to half the object's dimension plus an additional constant of two pixels in order to accommodate bad eye tracking and gesture recognition results.

If the object state changes multiple times during one fixation the most recent object state is considered for classification. Thus, a O_0-fixation can change into a O_c- or a R-fixation and a N-fixation into a P-fixation.

We classify a fixation as proactive if it belongs to category P. If $\boldsymbol{f}_t$ is a N-fixation and $(\boldsymbol{q}_{t_s} = \boldsymbol{q}_{t_r})$ it is also classified as a proactive fixation, since the fixation will eventually be categorized as a P-fixation. Due to the same reason O_0- and O_c-fixations are classified as proactive during selection and as reactive during manipulation phase. R-fixations are always classified as reactive.

A fixation is considered as being goal-directed if the gaze position $\boldsymbol{f}_t$ lies within one of the target areas, hence

$$(\boldsymbol{f}_t, \{0, 1\}) \cap \mathcal{G}_t \neq \emptyset \quad (11)$$

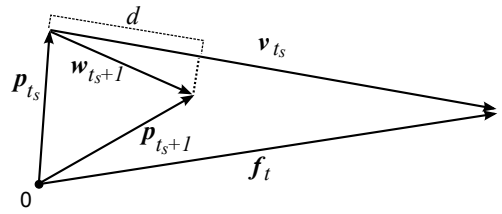

Figure 10: Relations between fixations and object states.

where $\mathcal{G}_t$ is the set of all target states at time t which are relevant for successfully completing the task. If no objects are selected at time t then $\mathcal{G}_t = \mathcal{G}_{\text{sel}} \cup \mathcal{G}_{\text{move}} \cup \mathcal{G}_{\text{desel}}$. After object selection the set of relevant target states reduces to $\mathcal{G}_{t+1} = \mathcal{G}_{\text{move}} \cup \mathcal{G}_{\text{desel}}$. If condition (11) is not fulfilled $\boldsymbol{f}_t$ is action-directed.

3.4 Inference of user intention

Having F_t and B_t, we can estimate the user's intention and the probability of future system states. In this section we describe how estimates for the next state Q_{t_s+1} and for the next goal $\boldsymbol{q}^g \in \mathcal{G}_{t_s}$ can be calculated from a proactive fixation. Reactive fixations are not considered for intention estimation, since they do not convey information about potential future system states. However, they might be useful for detecting unexperienced users as denoted in 3.1.

During selection phase the next state and the next goal can be estimated from goal-directed fixations, namely O_0- or O_c-fixations. The next state or goal is $\hat{\boldsymbol{q}}_{t_s+1} = \hat{\boldsymbol{q}}^g = (\boldsymbol{p}_{t_s}, 1)$ of the fixated object. During the manipulation phase, goal- and action-directed fixations must be treated separately.

goal-directed proactive fixation: $\boldsymbol{b}_t = (goal, proactive)$
Estimating the next goal is trivial, since by definition a goal-directed proactive fixation directly indicates a target state. In order to estimate the next state we need to calculate $P(Q_{t_s+1} \mid Q_{t_s}, F_{t_s}, B_{t_s})$. However, using a state transition matrix is in most cases impossible due to the large state spaces of Q_{t_s+1}, Q_{t_s} and F_{t_s}. Therefore we either could approximate the probability distribution by a parametric distribution or, as it is sufficient for many applications, only calculate an estimate $\hat{\boldsymbol{q}}_{t_s+1}$ for the condition expectation of Q_{t_s+1}. For goal-directed proactive fixations during the manipulation phase we first calculate the expected direction of object movement

$$\hat{\boldsymbol{v}}_{t_s} = \boldsymbol{f}_{t_s} - \boldsymbol{p}_{t_s}. \tag{12}$$

Assuming a linear movement, the next position of the object should lie somewhere close to the line defined by

$$\boldsymbol{p}_{t_s+1} = \boldsymbol{p}_{t_s} + D \cdot \frac{\hat{\boldsymbol{v}}_{t_s}}{\|\hat{\boldsymbol{v}}_{t_s}\|}. \tag{13}$$

In order to use this equation for calculating $\hat{\boldsymbol{p}}_{t_s+1}$, D can be considered as a random variable with values in $\mathbb{R}$. For goal-directed proactive fixations we assume $D \propto \mathcal{N}(\mu, \sigma)$. The parameters of the normal distribution can be estimated by using

$$d = \left\langle \boldsymbol{w}_{t_s+1}, \frac{\boldsymbol{v}_{t_s}}{\|\boldsymbol{v}_{t_s}\|} \right\rangle \tag{14}$$

for calculating realizations d of D from gaze and object data, where $\boldsymbol{w}_{t_s+1} = \boldsymbol{p}_{t_s+1} - \boldsymbol{p}_{t_s}$. On our data from the user study described in 2 we obtained by means of maximum-likelihood estimation $\mu = 16.3$ and $\sigma = 17.49$ for all proactive fixations.

action-directed proactive fixation: $\boldsymbol{b}_t = (action, proactive)$
For estimating the next state from an action-directed proactive fixation we use the same method and distribution for D as for goal-directed fixations.

The next goal can be estimated by first calculating the expected movement direction $\hat{\boldsymbol{v}}$ as in (12). Assuming a linear movement, the goal should lie somewhere close to or on the line defined in (13). The next goal can therefore be estimated by calculating

$$\mathcal{G}_g = \underset{\boldsymbol{q}^g \in \mathcal{G}_{t_s}}{\operatorname{argmin}} \measuredangle(\boldsymbol{p}^g - \boldsymbol{p}_{t_s}, \hat{\boldsymbol{v}}_{t_s}) \tag{15}$$

obtaining $\mathcal{G}_g$ as the set of potential next target states, where $\boldsymbol{q}^g = (\boldsymbol{p}^q, \alpha^q)$. If $\mathcal{G}_g$ contains more than one state, multiple potential target states in $\mathcal{G}_{t_s}$ have the same deviation from the estimated movement direction (e.g. lie on the line defined by (13)).

4. RESULTS AND DISCUSSION

We evaluated our model for two different use cases for the object manipulation task described in section 2. In the first case the estimated user intention is to be used for improving the robustness of an input device. In the second case the estimated intention is to be used for realizing a proactive user interface.

In the first case we want to calculate the probability distribution $P(I_{t_s} | Q_{t_s+1}, Q_{t_s} = \boldsymbol{q}_{t_s})$. It can be used as a-priori knowledge for I_{t_s} in order to improve the estimation obtained from noisy user input by $P(I_{t_s} | U_{t_s})$. Especially video-based input devices like the gesture recognition system used in this paper can benefit from such a-priori knowledge for the sake of detection, classification and tracking robustness as well as efficiency.

Since $\boldsymbol{q}_{t_s}$ is known, we only need to calculate $\hat{\boldsymbol{q}}_{t_s+1}$. During a selection phase $\hat{\boldsymbol{q}}_{t_s+1}$ determines when an object will be selected and which of the objects will be selected next. For all fixations in the selection phase we calculated $\hat{\boldsymbol{q}}_{t_s+1}$ as described in 3.4. 80.7% of the selections were predicted correctly, for 5 out of 9 participants even 100%. The percentage of false positive predictions was 12.3% and was mainly caused by noise in gaze position measurement. This number could be reduced by incorporating the duration of a fixation as an additional feature into F_t, because fixations directly preceding a selection last significantly longer than other fixations as we described in 2.2.

For evaluation of $\hat{\boldsymbol{p}}_{t_s+1}$ during the manipulation phase we interpreted the euclidean distance to the actual following state $\boldsymbol{p}_{t_s+1}$ as error. Depending on the participant we measured averaged errors of 9 pixels ($\sigma = 10$ pixels) up to 16 pixels ($\sigma = 12$ pixels) over all proactive fixations. The direction of movement was estimated with a mean deviation of 37.86 degrees from the true movement direction. Although we do not get perfect estimates, we can take advantage of them e.g. for improving the efficiency of detection or tracking processes in vision-based input devices by restricting the search space.

In the second use case, we estimated the next goal the user wants the system to take. This information also could be used for improving the robustness of an input device as in the first use case, but it also allows for the design of new

gaze-based interaction techniques. For the task considered here, the estimated target state $\hat{\boldsymbol{q}}^g \in \mathcal{G}_g$ could for example automatically be taken by the system without requiring the user to perform the whole manipulative movement. Especially when working at large screens such techniques could reduce physical fatigue induced by wide ranging hand movements.

Estimating the next goal during the selection phase is actually the same as in the first use case, since the next target state at the same time is the next state, namely the selection. During manipulation, however, many intermediate steps can be taken by the object before reaching its target state. Additionally action- and goal-directed gaze behavior needs to be treated differently as described in 3.4. With goal-directed fixations we obtained as expected 100% correct estimations due to the trivial task with only one target area and the definition of goal-directed fixations. For evaluating the accuracy of the estimated movement direction $\hat{\boldsymbol{v}}_{t_s}$, we calculated the reference vector $\boldsymbol{w}^g = \boldsymbol{p}^g - \boldsymbol{p}_{t_s}$ which is the difference between the actual target state $\boldsymbol{q}^g = (\boldsymbol{p}^g, 0)$ of an object in the target area and its state at the start of a fixation t_s. The mean absolute difference between $\hat{\boldsymbol{v}}_{t_s}$ and $\boldsymbol{w}^g$ was 21.77 degrees over all proactive fixations. By incorporating additional information, e.g. the movement direction of the hand or multiple fixations, the estimation could be further improved.

The promising results presented above show, that the proposed model provides a good basis for integrating natural gaze behavior in a multimodal context, at least for the simple but very common task considered here. The formulation of the model and the methods for classification of gaze behavior in section 3, however, are kept general and therefore can be applied easily to other tasks.

5. CONCLUSION

The presented user study revealed, as expected, that natural gaze behavior during an object manipulation task is highly complex and is determined by different parameters. Many influences can be explained by the proposed model, which describes fundamental causal relations between natural gaze behavior and the current task. By categorizing fixations into proactive and reactive as well as goal- and action-directed ones, the gaze data is interpreted in the context of interaction and can be handled on a higher level of abstraction for estimating the user's intention.

Since most information about future system states is conveyed by proactive fixations, multimodal interaction techniques using gaze should be designed in a way, which encourages usage of proactive gaze behavior. Our results indicate, that conformity of system reactions with the user's mental model seems to have a significant influence on the proactivity of gaze movements. Therefore using metaphors well known by the user for interaction techniques could be beneficial not only for reducing the training time, but also for increasing the amount of proactive gaze behavior. This however is subject to future work, as well as improvements to the accuracy of eye-tracking systems which has a significant influence on the performance of intention recognition. Another interesting question for future research is to identify what happens to natural gaze behavior if gaze is not only used for visual perception but also for interaction as illustrated in the second use case above. Finally the proposed model needs to be further evaluated for more complex tasks and influence of uncertainty and latency of video-based input devices on natural gaze behavior should be investigated systematically.

6. REFERENCES

[1] T. Bader, R. Räpple, and J. Beyerer. Fast invariant contour-based classification of hand symbols for hci. In *CAIP 2009*, volume 5702 of *LNCS*, pages 689–696. Springer, 2009.

[2] S. K. Card, J. D. Mackinlay, and G. G. Robertson. The design space of input devices. In *CHI '90: Proceedings of the conference on Human factors in computing systems*, pages 117–124, New York, NY, USA, 1990. ACM Press.

[3] A. T. Duchowski. *Eye Tracking Methodology, Theory and Practice*. Springer, 2003.

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

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

[6] R. J. K. Jacob. What you look at is what you get: eye movement-based interaction techniques. In *CHI '90: Proceedings of the conference on Human factors in computing systems*, pages 11–18. ACM, 1990.

[7] 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.

[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] B. Law, M. S. Atkins, A. E. Kirkpatrick, and A. J. Lomax. Eye gaze patterns differentiate novice and experts in a virtual laparoscopic surgery training environment. In *ETRA '04: Proceedings of the symposium on Eye tracking research & applications*, pages 41–48. ACM, 2004.

[12] 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.

[13] 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.

[14] SensoMotoric Instruments GmbH (SMI), Warthestrassee 21 D-14513 Teltow/Berlin. *iView X Manual Version 1.03.09*, September 2003.

[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.