Estimating the Gaze Point of a Student in a Driving Simulator

Wim Fikkert[1], Dirk Heylen[1], Betsy van Dijk[1], Anton Nijholt[1], Jorrit Kuipers[2], Arnd Brugman[2]
[1] *Human Media Interaction Group, University of Twente, The Netherlands*
{f.w.fikkert, d.k.j.heylen, e.m.a.g.vandijk, a.nijholt}@ewi.utwente.nl
[2] *Green Dino Virtual Realities, The Netherlands*
{jorrit, arnd}@greendino.nl

Abstract

In this paper we discuss an approach towards passively observing students in a driving simulator. The goal is to enhance the learning experience for students taking lessons in this simulator. To this end, a virtual driving instructor is provided with added information consisting of the gaze behavior of its student. The gaze behavior is defined by estimated head locations and orientations. The learning experience for the student is enhanced by providing added feedback to the student based on his observed behavior.

1. Introduction

"De Nederlandse Rijsimulator" (DNR) is a virtual reality driving simulator that enables driving schools to teach their students how to drive a car. Green Dino Virtual Realities developed the DNR for commercial exploitation; providing a cheap alternative for real-life driving lessons. The DNR is not intended as a complete substitute for learning how to drive in a car; instead it focuses on training how to participate in realistic traffic situations and on the techniques to training how to control the car itself.

The Virtual Driving Instructor (VDI) of the DNR provides its student with positive as well as negative auditory feedback before and during driving lessons. The VDI bases the type and form of feedback on input from the car controls, manipulated by the student, in addition to the current learning level of the student, learning rate, traffic situation, and lesson-topic.

Our work focused on observing students in the DNR, described in sections 2 and 3. In this section we first describe the need for a driving simulator. We then validate our work by relating it to previous research approaches. Third, more detailed overviews of the DNR and its student are provided.

1.1. Driving lessons – doing it virtual

On the road, it is impossible to experience every conceivable scenario during a driving course, e.g. specific weather conditions. It is impossible to trigger road user interactions e.g. a car to suddenly appear around the next bend or to arrive simultaneously at the next intersection. Driving on the road the student is confronted solely with these chance occurrences. Driving simulators have been developed to provide a safe and more controlled environment. The DNR adds to this notion via its VDI; eliminating the need for a human operator, thus ensuring commercial viability. The goal of a driving simulator is typically to maximize the number of realistic learning moments during a lesson. This allows a student to experience situations that he would normally be able to experience sporadically, or not at all.

A distinction can be made between research- and tutoring-oriented driving simulators. The former type focuses more on user-behavior analysis. In contrast, tutoring-driving simulators focus more on detecting user-errors that require feedback for the student to learn: users are typically not familiar with operating a car. In both simulator types, student observation is required, although the focus for observation differs greatly. In (driving) simulators a student is confronted with a maximum of generated interaction/learning moments. This maximum is often facilitated by the amount of reduced realism resulting in a more complex (learning) interaction, e.g. more realistic traffic situations but fewer fancy graphics. Especially in the case of the DNR, computational power had to be minimized to ensure a competitive edge.

1.2. Related work

Observing a user in a simulator is typically done by a human operator manually, or automatically via a user-worn sensor device. The added cost for an operator is negligible in more expensive simulators. Commercial exploited simulators therefore lack automated, passive user observation solutions.

Proceedings of the Sixth International Conference on Advanced Learning Technologies (ICALT'06)
0-7695-2632-2/06 $20.00 © 2006 IEEE

FaceLab [1] is a system that applies a set of two specialized cameras to actively track the face of a user. The University of Massachusetts adopts a similar approach. They combine a computer vision system with a head-mounted motion tracker [2]. User-worn sensors are quite common but limit freedom of movement.

Observing users in general is a broad and complex process that has been researched extensively in human computer interaction studies, especially in the field of computer vision. Observing users via a computer vision approach typically provides non-obtrusive information on the actions performed by the user.

1.3. The DNR

The DNR is designed as a learning environment in which no human operator or driving instructor is required, enabling its student to autonomously learn how to participate in traffic situations. In the DNR, the key component is the VDI that replaces its human counterpart. In this respect, the DNR currently is unique in the world. The DNR contains all essential control elements: steering wheel, directional indicator, clutch-, brake- and throttle-pedals, parking-brake, gearshift lever, and the driver's seat (see Figure 1). All elements are fully functional but the rest of the immersive environment is present only by projection.

Figure 1. "De Nederlandse Rijsimulator".

A student takes lessons that are predefined in a curriculum. Each lesson is set around a specific topic and is made up out of multiple lesson blocks, confronting the student with specific procedures to perform. The difficulty level increases as the student progresses through the curriculum. While performing driving procedures, the student is rated by the number of times a procedure is (in)correctly executed. A procedure is sufficiently mastered when the student has reached the highest of three rating levels. Procedure levels can change during any lesson and are not necessarily obtained in a single lesson.

A human instructor is still responsible for a student; he assesses when the student has reached a sufficient level to continue his education in a real car. To that end after each lesson in the DNR, the student and a human instructor can go over the obtained lesson results; e.g. repeating mistakes are identified in this manner.

1.4. The student – what to observe

The DNR was designed as a tool for any person capable to take driving lessons in real-life. The student is taught two types of procedures. Car control procedures require physical handling to operate the vehicle itself e.g. steering and shifting between gears. In contrast, insight procedures are needed to safely traverse encountered traffic situations e.g. crossing an intersection that requires a specific gaze procedure combined with a steering procedure.

Currently, the information available to the VDI consists solely of the information obtained via the control elements. It bases its feedback to the student on that information. However, insight procedures cannot be observed nor deducted using solely the currently available information. In the following sections we discuss our approach towards passive observation of the student in the DNR.

2. Observing students in the DNR

In this section we describe the current and desired situations in which the student is observed. This then leads to a short roadmap of required developments.

2.1. Current situation

The student is modeled by his current progress in the curriculum. The student interacts with road users that are part of a multi-agent system (MAS), as described in [3]. The MAS design was inspired by a film set in which a director, location scouts, actors, etc. participate to provide a script-like scenario at each adequate location, e.g. an intersection. Actors fill the role of road users, the scouts detect suitable locations, and the director orchestrates the scenario. The role performed by the student, from a system's point of view, can be seen as identical to an agent-actor. The VDI is also a part of this MAS structure in such that it observes student input and the subsequent results on the environment. The feedback towards the student is based on this input in addition to his current learning level, learning rate, traffic situation, and lesson-topic.

2.2. Desired situation

The information available to the VDI does not suffice for complete feedback to the student on his actions when compared with taking a driving lesson in a real car, i.e. feedback from a human instructor. Lacking information includes but is not limited to body pose and gaze behavior. The VDI requires information

on the current student body pose in order to provide feedback on incorrect execution of a procedure style, e.g. *how* a steering procedure is performed. In addition, information on the gaze point – the point in the DNR the student is gazing towards – provides clues on what was seen. The gaze point is not necessarily found on the projection screens. The goal of this research is to provide the VDI with sufficient information with which it can indirectly improve upon the learning experience of the student via additional feedback e.g. providing added audiovisual cues.

2.3. Roadmap

The required developments to reach this desired situation are subjected to some requirements: student physical characteristics are not taken into account, markerless non-obtrusive passive observation, and an off-the-shelf sensor solution. An approach towards passive observation of the student should be designed that can capture its observations in a model usable by the VDI to draw conclusions. The approach should be dynamic and expandable so that new elements can be added. In addition, the proposed approach is meant as an extension on the current version of the DNR.

3. Modeling student observations

In this section our generic approach towards automated student observing is described. We focused on estimating the gaze direction of the student whilst designing the approach. The estimated gaze direction leads, combined with the gaze origin, to the gaze point. The gaze direction is derived from head pose and the relative eye orientation. In addition, the gaze origin is defined as the point centered between the eyes. Eye orientations are omitted and we focus solely on the head pose for proof of concept. Moreover, in practice driving instructors require their students to indicate their gaze direction with the head.

Required improvements concerning the observation of users are described first. Second, the process of estimating a student gaze point during lessons is described, leading to a more general approach. Third, the possibilities for the analysis of gaze behavior, based on sequences of estimated gaze points, follow. Fourth, we discuss how the estimated student behavior can be recognized. These last subsections describe how other user observation topics can benefit from our approach.

3.1. Sensors for observation

One restriction for the observation approach is that the student should perceive no changes in his interactions with the simulator during a lesson when compared with the current version of the DNR, i.e. without the added feedback. Our choice for computer vision is a direct consequence of this restriction. Due to added commercially inspired requirements, a low-cost, minimal, yet scalable set of off-the-shelf cameras is selected. The information currently lacking to the VDI consists of the body pose and gaze behavior of the student. To be able to fully observe a student in the DNR, given its size and shape, a minimum of two cameras is required: a top down camera that provides an optimal view of student body pose and an *en face* camera that provides a maximal amount of facial detail that can be obtained in the DNR (see Figure 2).

Figure 2. DNR camera positioning: the top down and *en face* views.

3.2. Data pre-processing

The DNR uses three projection screens to immerse the student in the virtual driving environment. The student sits at roughly a meter from these screens. A reduced domain in which the student can be found during lessons is defined based on the DNR size. For the bodily features (head, hands, and feet) this domain can be reduced further, as they can be found only at the car controls. By focusing on these specific locations in the DNR, the camera images can be cropped; reducing data as a start in the image analysis process.

Illumination reflecting from the projection screens is homogenous and substantial. In addition, it is dynamic and influences the observed colors greatly. To effectively use color information, e.g. skin colors, color correction is required. A linear form of static gamut compression is used to enhance an image [4].

The reduced image space includes a relatively large quantity of the redundant information: the student is in a single pose at a time. Using temporal tracking an accurate prediction on the student (feature) location(s) is made. By focusing on the predicted regions, an even larger part of the search space is reduced.

3.3. Detecting gaze direction

The *en face* camera perspective provides the most optimal facial view but still is positioned slightly elevated due to its placement on top of the projection

screens. This available facial information is slightly reduced hereby so not all conventional facial analysis algorithms are suited for appliance here [5, 6].

As a first step in the process of gaze direction estimation, the face is located in the camera images using skin detection. Facial features are then detected by focusing on the detected facial area. Multiple skin models are used in a combined effort to robustly classify pixels as skin. A dynamic set of features is chosen that is detected in the images, including: eyes, ears, and mouth. The latter feature is found always by focusing on its characteristics, i.e. its specific shape and location in the face. Eyes and ears are then sought for by locating the two largest eye/ear-like features visible above the mouth [7]. Note that the detected facial feature set varies with these features' characteristics. These three features are modeled in a feature constellation. Taking into account the spanned angles, the shape of these constellations provides clues on the gaze direction. The process of feature constellation detection is illustrated below.

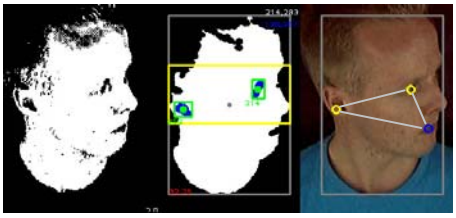

Figure 3. Process of detecting feature constellation: skin and feature detection, and the spanned constellation results.

The feature constellations are matched with a predefined constellation pool that describes the visible facial features of a typical student. A grid is spanned to describe the DNR in unambiguous distinguishable feature constellations as seen from the *student*, i.e. the constellation per cell describes the gaze direction for that cell. Figure 4 depicts this grid division.

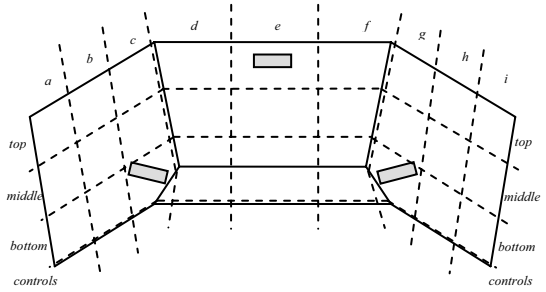

Figure 4. Grid division of the DNR.

Multiple users wore an inertia tracker while gazing along a predefined pattern. The automatically annotated results were averaged and used to create the constellation pool describing the gaze directions of a typical student. Matching the currently observed and predefined constellations is based on a similarity measure and the type of observed features (ear or eye).

The designed approach consists of a number of processing steps, most of which have been introduced above. Logically, distinct goals cannot be obtained using a single implementation; the last step(s) will take a different form. Moeslund and Granum [8] categorize image processing techniques in four subsequent phases: initializing, tracking, pose estimation, and recognition. Our approach fills the first three phases while leaving the last phase open: the VDI is responsible for recognizing specific behavior of the student. The approach consists of the sequence of steps [9]:

1. Initialization
2. Crop image to region of interest
3. Correct image colors
4. Feature detection
5. Constellation matching

The tracking phase described by Moeslund and Granum includes the image cropping (2) and color correction (3) steps. The phase in which the pose is estimated consists of the feature detection (4) step, and is concluded with the constellation matching step (5).

3.4. Towards body pose estimation

The described approach to gaze point estimation is designed in such a manner that it can be reused to obtain additional student behaviors. One explicit application for the designed approach is the ability to estimate the body pose of a student. The processing steps for estimating a body pose are identical, albeit that the implementation differs slightly. For body pose estimation we are interested in the position of the bodily end-effectors: hands, left-foot, and head. The latter is found during the gaze point estimation process and is assumed known. The right-foot is not interesting in a left-hand-drive car as the student can do nothing wrong with it, in contrast to the left-foot; i.e. hovering above the clutch pedal. The hands can be found in a limited domain in the DNR: either on, in-between, and off the control elements. The constellation that is constructed with these features will have different properties, i.e. shape and contents. Based on a human modeling standard, e.g. H|Anim [10], a simple skeleton can be constructed that takes on the role of the feature constellation. Body pose key frames are then used to match with a currently detected pose.

3.5. Towards behavior analysis

During a lesson, estimated gaze points of the student are communicated to the VDI. The VDI needs to analyze these points, deducting what the student has and, more importantly, has *not* seen. Expanding this notion towards body pose estimates, (in)correct body poses can be recognized. In addition, smaller bodily motions can be recognized, e.g. a steering technique.

Sequences of gaze point estimates and body poses can be recognized by matching them with predefined descriptions of known procedures. When a sufficiently small difference between the descriptions of an estimated and a predefined procedure exist the VDI can conclude it has observed that specific procedure. In addition, partially completed procedures can be recognized using this approach, e.g. the VDI will be able to discern when a student has not looked far enough over his shoulder when performing the mirroring sub procedure during an overtake procedure.

4. Discussion

Automated, non-obtrusive approaches to student observation do not yet exist for appliance in virtual learning environments. A passive sensor solution was chosen with which to observe the student from multiple yet minimal number of perspectives. The image data that was thus obtained was analyzed using a generic approach of processing steps, applicable for user observation topics in learning environments.

Our approach towards student behavior observation focuses on gaze point estimations of the student. This focus can be expanded towards other student observation topics, e.g. body pose estimation; thus providing a full description on the student behavior during driving lessons in the DNR. The VDI can use the provided information to conclude what behavior has been observed. This information can then be used by the VDI to provide additional feedback to the student during the driving lesson on correct and on incorrect execution of specific procedures. Depending on the requirements of the traffic situation, the student can thus be corrected or complemented when needed while performing a procedure.

The approach we propose tries to minimize the amount of data that needs to be analyzed by making explicit use of the *a priori* information on the environment. In most learning environments such a limited domain can be defined. Eliminating a maximum of redundant information early on in the analysis process, computational power is freed that can be used for better purposes. A prototype was constructed to test our proposed approach. This prototype is able to accurately estimate the gaze point of a student in near real-time. However the prototype did have difficulty detecting facial features under dynamic environmental settings, e.g. ambient illumination changes.

Future work should include research on modeling a learning environment and its student for passive observation. In addition, a virtual tutor should be described in more generic terms. Furthermore, the amount of obtrusiveness that is perceived by a student, caused by other sensor solutions, should be studied. Such sensor solutions are typically more mature and provide fewer difficulties and noise in data analysis.

References

[1] Seeing Machines - Creators of Visionary Technology, "FaceLab, online at http://www.seeingmachines.com," n.d.

[2] University of Massachusetts, "Human Performance Laboratory, online at http://www.ecs.umass.edu/hpl/," n.d.

[3] I. Wassink, E. M. A. G. van Dijk, J. Zwiers, and A. Nijholt, "Bringing Hollywood to the Driving School: Dynamic Scenario Generation in Simulations and Games," *Entertainment: First International Conference - INETAIN 2005*, LNAI vol. 3814 / 2005, pp. 288-292, 2005.

[4] G. Sharma, *Digital Colour Imaging Handbook*: CRC Press, 2003.

[5] A. Gee and R. Cipolla, "Determining the gaze of faces in images," *Image and Vision Computing*, vol. 12, pp. 639--647, 1994.

[6] J. Wang and E. Sung, "Study on eye gaze estimation," *Systems, Man and Cybernetics, Part B, IEEE Transactions on*, vol. 32, pp. 332--350, 2002.

[7] D. Hansen, "Committing Eye Tracking." PhD. Thesis. Copenhagen: University of Copenhagen, 2003.

[8] T. Moeslund and E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," *Computer Vision and Image Understanding*, vol. 81, 2001.

[9] W. Fikkert, "Estimating Student Focus of Attention in a Driving Simulator," University of Twente, Enschede, Internal Report 2005.

[10] H|Anim, "VRML Humanoid Animation Working Group, online at http://www.h-anim.org/," n.d.