IEEE Transactions on Intelligent Transportation Systems, Volume 21, Issue 10, October 2020, pp. 4318-4327
DOI:10.1109/TITS.2019.2939676

A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring

Lichao Yang, Kuo Dong, Arkadiusz Jan Dmitruk, James Brighton, and Yifan Zhao, *Senior Member, IEEE*

***Abstract*— Characterisation of the driver's non-driving activities (NDAs) is of great importance to the design of the take-over control strategy in Level 3 automation. Gaze estimation is a typical approach to monitor the driver's behaviour since the eye gaze is normally engaged with the human activities. However, current eye gaze tracking techniques are either costly or intrusive which limits their applicability in vehicles. This paper proposes a low-cost and non-intrusive dual-cameras based gaze mapping system that visualises the driver's gaze using a heat map. The challenges introduced by complex head movement during NDAs and camera distortion are addressed by proposing a nonlinear polynomial model to establish the relationship between the face features and eye gaze on the simulated driver's view. The Root Mean Square Error of this system in the in-vehicle experiment for the X and Y direction is 7.80±5.99 pixel and 4.64±3.47 pixel respectively with the image resolution of 1440 × 1080 pixels. This system is successfully demonstrated to evaluate three NDAs with visual attention. This technique, acting as a generic tool to monitor driver's visual attention, will have wide applications on NDA characterisation for intelligent design of take over strategy and driving environment awareness for current and future automated vehicles.**

***Index Terms*— Driver attention evaluation, Level 3 automation, camera mapping, system identification, heat map.**

I. Introduction

IN RECENT years, the exciting developments of highly automated driving (HAD) vehicle have been made in the field of both academic research and industrial manufacturing [1]. According to SAE (J3016) Automation Levels, all the dynamic driving tasks can be achieved by the automated driving system, but drivers are expected to response appropriately when the intervene is requested by vehicle in Level 3 [2]. Although at present legislation does not allow drivers in a Level 3 autonomous vehicle to engage in non-driving activities (NDAs), HAD may in the future allow drivers to more freely engage in NDAs during much of the time while the automated system monitors and reacts to the driving environment. Based on the research of Sivak and Schoettle [3], when vehicle is in the self-driving mode, 57.1% respondents in UK choose to watch the road. The main NDAs are reading (9.9%), sleeping (9.4%), texting or talking with friends (7.1%), working (6.4%) and watching movie (5.4%). Since driving is considered as a complex activity which requests the people's sensory, cognitive and psychomotor process to be synchronized, for the concern of driving safety, it needs to be ensured that the state of driver is suitable for driving. Therefore, the drivers' behaviour in NDAs needs to be monitored and their attention level needs to be evaluated. This is particularly important when vehicle requests the drivers' intervene when the driving environment is complicated, for example, when approaching a complex or less predictable driving scenario, such as temporary road works. Furthermore, it is crucial to evaluate drivers' awareness of driving environment (e.g. pedestrian, obstacle or neighbouring vehicles) right before the take-over. Developing a drivers' gaze mapping system for evaluating visual attention is therefore hugely demanded for the development of Level 3 automated vehicles. Such a research can also be easily extended for the studies of driver distraction in human driving [4]–[6].

Gaze tracking and head movement estimation is the common way for allocating the driver's visual attention in automated driving [7]–[9]. Several approaches and devices were designed by using scleral search lens and head-mounted eye trackers [10], [11]. They provide the accurate gaze information but are not practical in real applications as they are intrusive and costly. The non-intrusive eye tracking based on the human-computer interaction (HCI) system is well used due to the user friendliness [12], [13]. Mizuno *et al.* [14] used a vehicle-mounted camera in front of the driver to estimate the driver's visual attention area to check if the driver is aware of the driving environment. Some researchers extracts the facial features of drivers and allocates the driver's gaze into different regions with a single camera to provide the information regarding to the driver's attention [7], [15]–[17]. All of these studies made a fixed assumption between the eye gaze direction and the driver's behaviour, which is not applicable for characterisation of NDA due to its high complexity and uncertainty. The main difference of this paper with other researches about eye gaze tracking is how to represent the eye gaze. The existing gaze tracking researches, including pupil and gaze modelling [18]–[20], camera-based

Manuscript received April 17, 2019; revised July 5, 2019; accepted August 28, 2019. The Associate Editor for this article was L. M. Bergasa. *(Corresponding author: Yifan Zhao.)*

L. Yang, A. J. Dmitruk, J. Brighton, and Y. Zhao are with the School of Aerospace, Transport and Manufacturing, Cranfield University, Bedfordshire MK43 0AL, U.K. (e-mail: yifan.zhao@cranfield.ac.uk).

K. Dong is with the Chongqing Automotive Collaborative Innovation Centre, Chongqing University, Chongqing 400044, China.

This article has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Digital Object Identifier 10.1109/TITS.2019.2939676

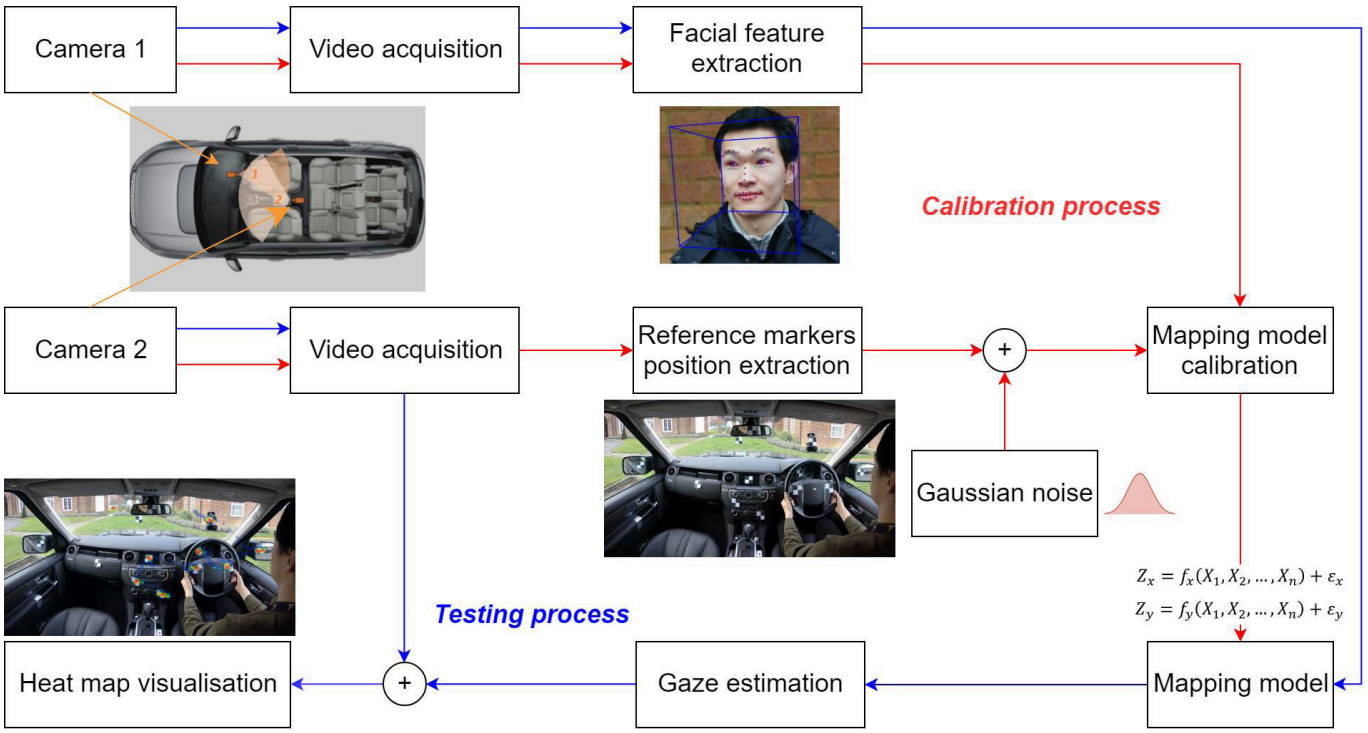


Fig. 1. The proposed system flowchart. There are two processes in this system including calibration in red and testing in blue.

tracking [15], [21], [22] or artificial intelligence implemented system [23]–[25], use the gaze data to validate the system and optimise the accuracy. This research is interested in the gaze mapping on the driver's visual scene, which focuses on how the gaze engages with the driving environment, including both inside and outside the vehicle. The mapping can then be further used for attention analysis of NDAs or evaluation of environment awareness. Xiao and Feng [26] proposed a driver's visual attention system by using a smartphone. The back camera of the smartphone is used to capture the moving object and the frontal camera is used to estimate the driver's gaze. The view of the back camera is divided into 9 zones. The system aims to check if the driver is aware there is a moving object inside zones. However, this solution is not appropriate to monitor the driver's NDAs inside the vehicle. On the other hand, the eye gaze tracking researches in automotive field are generally used for driving behaviour and distraction analysis in human drive [7], [27], [28]. As far as we are concerned, there is very limited literature reported aiming to implement the system into autonomous vehicles for characterisation of NDAs.

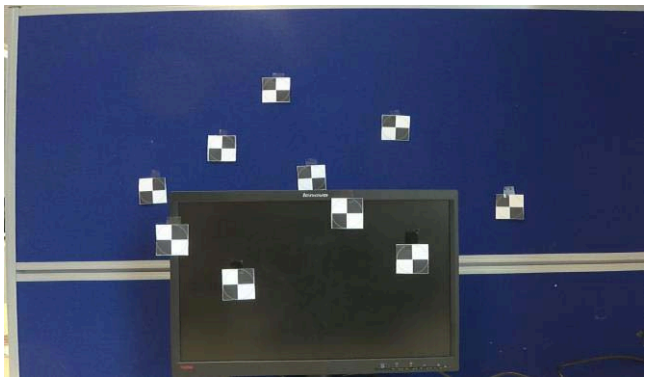
This paper presents a non-intrusive, low-cost and user-friendly driver gaze mapping solution based on two cameras. A nonlinear finite impulse response model powered by Error Reduction Ratio is introduced to estimate and map the gaze location by automatically selecting the face features and corresponding parameter estimation. Heat map is introduced in this system to visualise the trajectory of eye gaze, which could be used to identify the type of NDA and even its attention level.

II. Methodology

A. System Architecture

The framework of the proposed system is divided into four steps including video acquisition, feature extraction, gaze mapping and heat map visualisation. As shown in the flowchart illustrated by Fig. 1, the first feed of video is captured through a camera placed in front of the driver, as indicated Camera-1, to capture the facial features including eye gaze and head movement. The second feed of video is captured through a camera placed on the top of the driver, referred as Camera-2 in Fig. 1, to mimic the driver's view. The driver's gaze directions along with other parameters including face location and orientation are extracted based on videos from Camera-1. These parameters are considered as the inputs of the model for gaze mapping. The proposed method tends to include the driver's face features as more as possible and let the later modelling/mapping process to determine which features should be included for estimating the output, the mapped location of eye gaze in images of Camera-2. The mapping model calibration is to establish a model to represent the relationship between the face features in images of Camera-1 and eye-gaze locations on images of Camera-2. In the calibration (or training) process, the eye-gaze location in images of Camera-2 is known using markers placed on the vehicle. This paper assumes that the gaze is a region with an approximate Gaussian distribution which represents the driver's observation intensity [29]. Gaussian noise with a pre-set sigma is therefore applied on the marker locations on images from Camera-2, as the known outputs of training data.

(a)



(b)

Fig. 2. (a) The spatial distribution of the markers in land rover discovery 4 for the in-vehicle experiment. (b) The spatial distribution of the markers in laboratory for the indoor experiment.

From the system identification point of view, adding noise to the desired output can reduce overfitting and improve model generalisation. A large value of sigma will reduce the accuracy of fitting, but improve the model generalisation. In this paper, the sigma value was chosen as 10 pixels to achieve the optimal balance. Once this relationship is established, this model can be deployed on face features extracted from a testing video of Camera-1 and produce a mapping on the scene captured by Camera-2.

Considering the NDA as a dynamical process, this study focuses on the eye gaze on a certain time window and a form of heat map is proposed for visualisation. The details of each step are presented below.

B. Video Acquisition

The Land Rover Discovery 4 was used as the test vehicle. Camera-1 is located in the wind shield in front of the driver. The location of Camera-2 is set on the top of driver towards the windscreen. The markers shown in Fig. 2(a) were placed on the strategic locations inside the vehicle including dashboard, side mirrors, rear-view mirror, windscreen, multimedia display and steering wheel etc. These locations are fixed and friendly for the driver to look at. In this paper, a total number of 12 markers were used.

Before implementing the system in vehicle, a feasibility study has been conducted in the laboratory to better evaluate the performance of the proposed system. The layout of

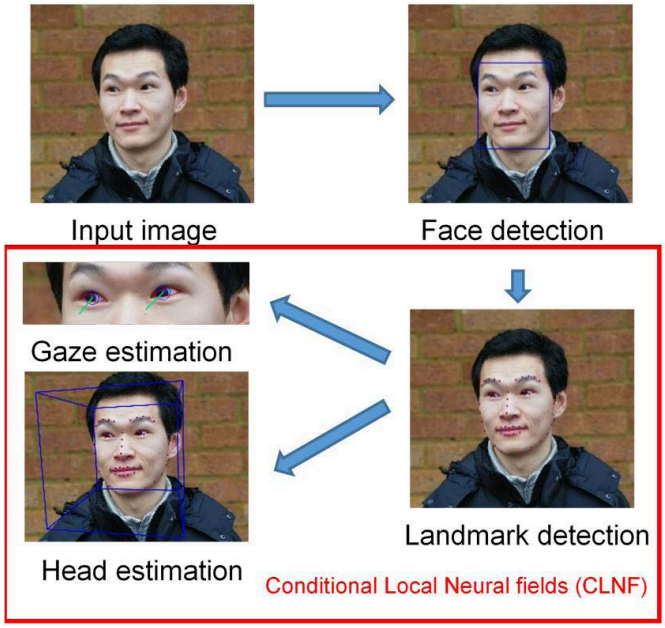


Fig. 3. OpenFace facial behaviour analysis process.

cameras is the same as mentioned above. Ten markers are located randomly and shown in Fig. 2(b). It should be noted that there are 4 markers on the monitor which has shorter distance to Camera-2.

For both experiments, the employed cameras were Garmin Virb Action Camera. Camera-1 provides the video with a resolution of 1024×768 pixel and 24 frames per second. Since a wider field of view is requested for Camera-2, the resolution is set as 1440×1080 pixels and the temporal resolution remains the same value.

C. Feature Extraction

In recent years, several gaze and head tracking methods have been proposed [30], [31]. As one of the most popular open-source facial analysis tools, OpenFace is utilized for the purpose of extracting the features of the driver's gaze and head due to its fine performance and robustness. It is capable of facial landmark detection and action unit recognition, head pose and eye-gaze estimation [32], [33]. The algorithm starts with face detection then is followed by the 68 facial landmarks detection. These landmarks are used to estimate the head pose and track the eye gaze. The process is illustrated in Fig. 3. Conditional Local Neural Fields (CLNF) framework is utilized as a shape registration approach for detecting the facial landmarks [34]. There are two components for CLNF which are Point Distribution Model (PDM) and patch experts. PDM captures variations of the landmark shape and the local appearance variations of each landmark are captured by patch experts. For head pose estimation, the orthographic camera projection is used to project the 3D representation of facial landmarks. The SynthesEyes training dataset [35] is used to train the PDM and CLNF patch experts for the eye-region registration task. Once the eye and the pupil are located, the data are used to calculate the gaze vector for each eye.

The gaze estimation ability of this model is validated by the MPIIGaze dataset [36]. The performance of this approach on driver monitoring has been evaluated in the research of Zhao *et al.* [1].

Considering the complexity and uncertainty of the driver's behaviours during NDAs, this paper proposes to use both head information and gaze information to build up the gaze heat map. The selected parameters are divided into two categories: the head pose related parameters (HRPs) and the gaze related parameters (GRPs). HRPs include the position of detected head with respect to Camera-1, denoted by *pose_Tx*, *pose_Ty* and *pose_Tz*, and head orientation in 3D, denoted by *pose_Rx*, *pose_Ry* and *pose_Rz*. GRPs include the information of the gaze direction in radians, denoted by *gaze_angle_x* and *gaze_angle_y*. It should be noted that, to simplify the model, this study only considers the information of one eye. Experiments show that the information of another eye has no significant contribution to the performance.

D. Feature Mapping

This paper proposes to use the orthogonal least squares (OLS) algorithm to establish the correspondence between the face features based on the coordinate of Camera-1 and the eye gaze mapping based on the coordinate of Camera-2. This is an approach that has been used in nonlinear system identification where OLS searches through all possible candidate model terms to select the most effective ones to build the model. The significance of each selected model term is measured by the Error Reduction Ratio (ERR) index which indicates how much of the change in the system response, in percentage, can be accounted for by including the relevant model terms. This capability is important for this study because the face features have been extracted as more as possible to ensure the proposed system can accommodate the diversity of driver's behaviour, meanwhile we need avoid producing an over-complex model that over-fits the training data and produces relatively poor testing performance. This algorithm allows us to only select the important face features for modelling to reach the balance between model complexity and gaze estimation performance. Furthermore, the capability to accommodate nonlinear modelling is important to cope the distortion of images of Camera-2, which is the by-product where a wide field-of-view is required.

The Volterra Non-linear Regressive with eXogenous inputs (VNRX) model, also known as nonlinear finite impulse response (NFIR) model, is used in this paper to represent a multi-inputs and single-output system, where the inputs are the face features and the output is the eye gaze location on images of Camera-2. It should be noted that the eye gaze location includes two values: x and y, which will be modelled independently. The models can be expressed as:

$$Z_x = f_x\left(X_1, X_2, \ldots, X_n\right) + \varepsilon_x \tag{1}$$

$$Z_y = f_y\left(X_1, X_2, \ldots, X_n\right) + \varepsilon_y \tag{2}$$

where $X_1, X_2, \ldots, X_n$ are the face features; n is the number of collected face features; Z_x and Z_y are the eye gaze location in x and y direction respectively; f_x and f_y are some unknown linear or nonlinear mappings link the inputs and output; ε_x and ε_y are module residual.

Consider a function in a linear form:

$$Y\left(k\right) = \sum\nolimits_{i=0}^{N} \theta_i \, p_i\left(k\right), \quad k = 1, 2, \ldots, M \tag{3}$$

where $Y(k)$ is the system output (eye gaze location in x or y direction), $p_i\left(k\right)$ are regressors constructed by input variables, θ_i is the vector of unknown coefficients of regressions to be estimated, M denotes the number of data points in the training data set, and N denotes the number of terms in the model that is yet to be determined. If the model order is set as q, the candidate term set where $p_i\left(k\right)$ select from, denoted by C, can be expressed

$$C = C_1 \cup C_2 \cup \ldots \cup C_l \cup \ldots \cup C_q, \tag{4}$$

where C_1 is the linear term set, expressed as

$$C_1 = \bigcup\nolimits_{a=1}^{n} X_a, \tag{5}$$

and C_2 is the 2^{nd} order nonlinear term set, expressed as

$$C_2 = \bigcup\nolimits_{a_1=1}^{n} \bigcup\nolimits_{a_2=a_1}^{n} X_{a_1} X_{a_2} \tag{6}$$

and C_l is the l^{th} order nonlinear term set, expressed as

$$C_l = \bigcup\nolimits_{a_1=1}^{n} \bigcup\nolimits_{a_2=a_1}^{n} \cdots \bigcup\nolimits_{a_l=a_{l-1}}^{n} \prod\nolimits_{i=1}^{l} X_{a_i} \tag{7}$$

Equation (3) is re-written as

$$Y = P\Theta \tag{8}$$

where

$$Y = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(M) \end{bmatrix}, \quad P = \begin{bmatrix} P^T(1) \\ P^T(2) \\ \vdots \\ P^T(M) \end{bmatrix}, \quad \Theta = \begin{bmatrix} \theta(1) \\ \theta(2) \\ \vdots \\ \theta(M) \end{bmatrix} \tag{9}$$

and $P^T\left(k\right) = \left(p_1\left(k\right), p_2\left(k\right), \ldots, p_N\left(k\right)\right)$. Matrix P can be decomposed as $P = W \times A$ where

$$W = \begin{bmatrix} w_1(1) & w_2(1) & \ldots & w_N(1) \\ w_1(2) & w_2(2) & \ldots & w_N(2) \\ \vdots & \ddots & \ddots & \vdots \\ w_1(M) & w_2(M) & \ldots & w_N(M) \end{bmatrix}, \tag{10}$$

and $A = \{a_{ij}\}$ is an upper triangular matrix with unity diagonal elements. Equation (4) is then rewritten as

$$Y = WG \tag{11}$$

where $G = A\Theta = \begin{bmatrix} g_1 & g_2 & \ldots & g_N \end{bmatrix}^T$. Equation (11) is now ready to represent the relation between Y and G.

We then estimate the importance of each model term to the variation of the system output. Initially, set values $a_{ij} = 0$for $i \neq j$ (A then becomes an identity matrix), so $w_1\left(k\right) = p_1(k)$, and calculate g_1 as

$$g_1 = \frac{\sum_{k=1}^{M} w_1(k) y(k)}{\sum_{k=1}^{M} w_1^2(k)}. \tag{12}$$

For $j = 2, 3, \ldots, M$, set $a_{jj} = 1$ and then calculate

$$a_{ij} = \frac{\sum_{k=1}^{M} w_i(k)\, p_j(k)}{\sum_{k=1}^{M} w_i^2(k)} \tag{13}$$

where $i = 1, 2, \ldots, j-1$. Next, the algorithm calculates

$$w_j(k) = p_j(k) - \sum_{i=1}^{j-1} a_{ij}\, w_i(k), \tag{14}$$

and

$$g_1 = \frac{\sum_{k=1}^{M} w_j(k)\, y(k)}{\sum_{k=1}^{M} w_j^2(k)}. \tag{15}$$

The ERR value for each term p_i is finally defined as

$$ERR_i = \frac{g_1^2 \sum_{k=1}^{M} w_i^2(k)}{\sum_{k=1}^{M} y^2(k)}. \tag{16}$$

Values of ERR range always from 0% to 100%. The larger the ERR the higher dependence between the p_i terms and the output. Therefore, it is an indicator to represent the importance of each term (constructed by the face features as inputs) to the output.

The estimation of the coefficient of each selected term can be computed from

$$\left.\begin{aligned} \hat{\theta_N} &= \hat{g_N} \\ \hat{\theta_i} &= \hat{g_i} - \sum_{k=i+1}^{N} a_{ik}\theta_k, i = N-1, \ldots, 1 \end{aligned}\right\} \tag{17}$$

Through the above algorithm, a polynomial model based on Eq. (3) can be established for each direction of the eye gaze location. The models can then be used for estimation of eye gaze location by given the face features.

E. Heat Map Visualisation

Heat map is a common visualisation approach to represent the spatial distribution of the data [37]. This paper assumes that the eye gaze at a certain time or frame (t) can be represented by a circle which is defined by three parameters: $x_0(t)$ and $y_0(t)$, the location of the centre, and d, the diameter of the circle. The spatial distribution inside the circle follows the Gaussian distribution. The value of d is affected by the image resolution of Camera-2. It was set as 40 pixels for gaze visualisation.

Considering at the frame$t$, the eye gaze centred at $(x_0(t), y_0(t))$, the intensity of the pixel(x, y) in the heat map, where $2 * |x - x_0(t)| \leq d$ and $2 * |y - y_0(t)| \leq d$, can be defined as

$$s(x, y, t) = e^{\left(-\frac{(x-x_0(t))^2 + (y-y_0(t))^2}{2\sigma^2}\right)} * 100\% \tag{18}$$

The intensity of the pixels unsatisfied with the constraints is set as 0.

To represent the trajectory of gaze, this study integrates the gaze spatial distribution within a certain time window $[t-h, t]$, where t is the number of the frame and h is the window length. The accumulated eye gaze can be written as

$$s_a(x, y, t) = \frac{1}{h} \sum_{i=0}^{h-1} s(x, y, t-i) \tag{19}$$

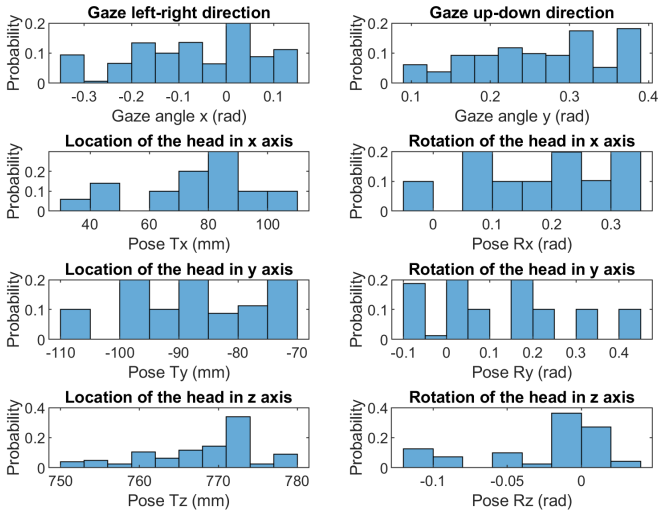


Fig. 4. Histograms of the facial features of the training data for the model calibration of the first test of the indoor experiment.

To better visualise the gaze trajectory in real-time, this paper proposes a weighted accumulation of eye gaze to construct the trajectory, written as

$$s_c(x, y, t) = \frac{1}{h} \sum_{i=0}^{h-1} s(x, y, t-i) * (1 - \frac{i}{h}) \tag{20}$$

The value of $s_a(x, y, t)$ and $s_c(x, y, t)$ is between 0 and 1. The window length h can be adjusted in terms of various applications.

III. Results

A. Indoor Experiment

Two tests were conducted in this experiment based on the level of freedom of head movement. In the first test, the participant was asked to gaze at the 10 markers one by one avoiding moving head forward or backward, so the shift of eye gaze was primarily achieved by head rotation. In the second test, the participant was given more freedom of head movement and both rotation and translation were allowed, aiming to simulate the increased complexity and uncertainty of head movement during NDAs. The participant was required to gaze at each marker for at least 5 seconds. The data of transition period when moving from one marker to another was removed. A total number of 1000 frames (100 frames per marker) were selected for training and testing. For each marker, 70% data were randomly selected for training and the remaining 30% data were for testing.

Fig. 4 presents the histograms of eight facial features of the training data in the first test. It can be observed that the head rotation movement is within 0.5 rad in pitch (*pose_Rx*) and yaw (*pose_Ry*). The roll movement of head (*pose_Rz*) is relatively small, within 0.15 rad, which is expected because there is not much rolling movement required to scan all markers. The variation of head position in z axis (*pose_Tz*), indicating the distance from the head to Camera-1, is within 30 mm. Although the translation of head was limited in this test, the head rotation caused a small variation of head depth.

TABLE I

AN EXAMPLE OF THE ESTIMATED 2nd ORDER NONLINEAR MODEL FOR THE FIRST TEST OF THE INDOOR EXPERIMENT

Model	X		Y	
Priority	Model term	Coefficient	Priority	Model term
1	constant	811.25	constant	528.27
2	gaze_angle_x	-1582.93	gaze_angle_y	474.31
3	pose_Tz* pose_Rx	53.43	pose_Tx	-2.03
4	gaze_angle_y	-917.28	pose_Tx* pose_Rx	-35.57
5	pose_Tx	-0.58	gaze_angle_x	465.13
6	pose_Tx* pose_Rx	-46.94	gaze_angle_y* gaze_angle_y	821.15
7	pose_Rx	853.26	pose_Ty* pose_Ty	0.21
8	pose_Rx* pose_Ry	-4419.04	gaze_angle_y* pose_Rz	16140.28
9	pose_Ry	450.32	gaze_angle_x* gaze_angle_y	-2771.10
10	gaze_angle_y* pose_Tz	-62.47	pose_Rx	1248.56

TABLE II

MODEL PERFORMANCE COMPARISON OF THE INDOOR EXPERIMENT

Term	Root Mean Square Error	
	Pixel	Millimetre
Test1_X	11.89 ± 9.00	9.25 ± 7.00
Test1_Y	9.22 ± 6.55	7.17 ± 5.10
Test2_X	27.33±14.29	21.26 ± 11.12
Test2_Y	20.71±11.51	16.11 ± 8.95

TABLE III

THE DATA RANGE COMPARISON OF THE EXTRACTED FACIAL FEATURES OF THE TRAINING DATA OF THE INDOOR EXPERIMENT

Features	The first test	The second test
Gaze_angle_x	[-0.35, 0.15] rad	[-0.35, 0.25] rad
Gaze_angle_y	[0.08, 0.38] rad	[0.1, 0.4] rad
Pose_Tx	[30, 110] mm	[25, 115] mm
Pose_Ty	[-110, -70] mm	[-108, -70] mm
Pose_Tz	[750, 780] mm	[660, 870] mm
Pose_Rx	[-0.05, 0.35] rad	[-0.05, 0.38] rad
Pose_Ry	[-0.1, 0.45] rad	[-0.25, 0.45] rad
Pose_Rz	[-0.14, 0.04] rad	[-0.14, 0.07] rad

Table I shows an example of the estimated 2nd-order nonlinear models of gaze in X and Y directions. The number of model term is limited to 10. The model term is ranked based on the ERR value which represents the importance of each model term to the variation of gaze. It can been observed from Table I that the most important term is 'constant' for both X and Y which refers to the base line of the head movement and relates to the initial state of the participant. As expected, the second important term is the gaze angle for the considered direction. It is interesting to observe that HRPs also make significant contribution to the model, which suggests that both HRPs and GRPs must be considered due to the complexity of human behaviour and distortion of cameras.

To quantify the performance of the proposed system in the first test, the produced models were applied in the testing data and the Root Mean Square Error (RMSE) of the estimated gaze location and the centre of marker (without adding Gaussian noise) for all markers was computed to represent the model accuracy. Since the testing data were randomly selected, this process was repeated 1000 times to ensure the statistical significance. Table II provides the mean (overall accuracy of model) and the standard deviation (precision of model) of the 1000 calculated accuracies for two tests. The size of the markers in the mapping frame is 37 pixels (28.8 mm). From the results of the first test in the Table II, it can be observed that the error of gaze estimation in X direction is 11.89±9.00 pixel (9.25±7.00 mm), while for the Y direction the error is smaller with a value of 9.22±6.55 pixel (7.17±5.10 mm). The errors of both directions are well smaller than the marker size, which indicates a fine performance of the proposed system when the head translation is limited. The performance of Y direction is better than X direction. This observation is reasonable because the makers cover larger range of X direction which leads to a higher level of distortion. The accumulated eye gaze map, calculated by Eq. (19), is presented in Fig. 5(a), where all estimated gaze points well fall into the markers, although there are some slight shifts between the centre of the gaze circle and the centre of the markers.

In the second test, the head movement was more complex by introducing both translation and rotation of head. It has been observed from Table III that the head position in z axis (*pose_Tz*) has a variation of 210 mm, which is 7 times higher than the first test. The ranges of other features are similar with the ones of the first test. The second test aims to test the flexibility of the proposed mapping algorithm against the diverse head movement of NDAs. Tables IV presents an example of the estimated 2nd-order nonlinear models of gaze in X and Y directions. The number of model term is limited to 10. It can be observed that the top 2 terms are the same as the ones of the first test, however, HRPs make more contribution to the model evident by more appearance in the selected model terms, particularly *pose_Tz*. As shown in Table SI and SII in Supplementary Materials, the proportion of ERR of HRPs in x direction is increased from 0.55% in the first test to 5.88% in the second test. The proposed algorithm successfully demonstrated the flexibility by selecting terms including *pose_Tz* to reflect the increased variation of head translation.

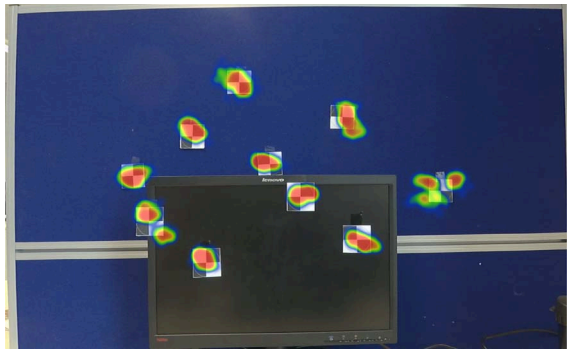
Table II also shows the quantified performance of the second test, using the same approach as the first test. It is shown that the RMSE in X direction is 27.33±14.29 pixel (21.26 ± 11.12 mm) and 20.71±11.51 pixel (16.11 ± 8.95 mm) for Y direction. As expected, the overall

TABLE IV

AN EXAMPLE OF THE ESTIMATED 2 nd ORDER NONLINEAR MODEL FOR THE SECOND TEST OF THE INDOOR EXPERIMENT

Model	X		Y	
Priority	Model term	Coefficient	Model term	Coefficient
1	constant	776.76	constant	545.12
2	gaze_angle_x	-2253.33	gaze_angle_y	1612.76
3	pose_Tz* pose_Ry	1.92	gaze_angle_x	-207.97
4	pose_Tx* pose_Tx	0.04	pose_Tz* pose_Tz	-0.01
5	pose_Ry	-863.78	pose_Tx* pose_Rx	-27.94
6	pose_Tx	-5.23	gaze_angle_y* gaze_angle_y	3018.13
7	gaze_angle_x* pose_Tz	-2.32	pose_Tx* pose_Ty	0.23
8	pose_Rz	-62.47	pose_Ry* pose_Rz	-2826.91
9	gaze_angle_x* pose_Rz	987.56	pose_Ry	-432.25
10	pose_Tz	0.46	pose_Ty	2.89

(a)



(b)

Fig. 5. (a) The accumulated eye gaze mapping for the first test of the indoor experiment. (b) The accumulated eye gaze mapping for the second test of the indoor experiment.

performance is not as good as the first test due to the increased complexity of head behaviour, but the error is still smaller than the marker size (28.8 mm). It is interesting to observe that the performance in X and Y directions are similar for this case which suggests that the interference caused by camera distortion is overtaken by the interference caused by severe head movement. Fig. 5(b) illustrates the accumulated eye gaze map for the second test. In comparison with Fig. 5(a), the regions of the gaze estimation are larger and more irregular but still well cover the majority markers. It can be observed from Fig. 5 that the visualised results of the 4 markers on the monitor which have shorter distance to Camera-2 than other markers also show a similar performance, which demontrates the robustness of the proposed system in terms of the depths of object.

TABLE V

MODEL PERFORMANCE OF THE IN-VEHICLE EXPERIMENT

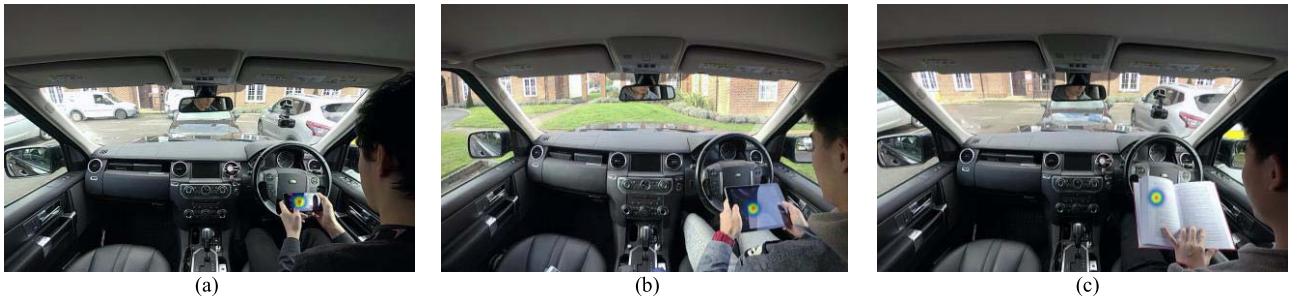
Term	Root Mean Square Error	
	Pixel	Millimetre
X	7.80 ± 5.99	12.00 ± 9.22
Y	4.64 ± 3.47	7.14 ± 5.34

Fig. 6. The accumulated eye gaze mapping for the vehicle experiment.

B. In-Vehicle Experiment

In this experiment, a wider field of view of Camera-2 in comparison to the in-door experiment was used due to limited space in vehicle, which inevitably introduced more distortion on images. Furthermore, 12 markers were laid out on the regions of interest, which have more diverse distances to the plane of Camera-2 in comparison with the indoor testes. Due to these factors, a more sophisticated model is required to cope the increased complexity. Therefore, a 3rd-order nonlinear model was estimated with the number of model term of 25. It should be noted that in this experiment the participant was asked to scan the markers with limited translation of head, as the first indoor test. The approach to select the training and testing data was the same as the indoor experiment.

As shown in Table V, the RMSE in X and Y direction is 12.00 ± 9.22 mm and 7.14±5.34 mm respectively, which is well smaller than the marker size (28.8 mm). The performance is better than the first indoor test with a cost of increased



(a) (b) (c)

Fig. 7. (a) Gaze mapping when the driver is playing game by using a cell phone. (b) Gaze mapping when the driver is watching movie by using a tablet. (c) Gaze mapping when the driver is reading a book.

TABLE VI
MODEL PERFORMANCE BASED ON DIFFERENT MODEL ORDER FOR THE IN-VEHICLE EXPERIMENT

Term	Root Mean Square Error (pixel)	
	X	Y
Linear	56.49 ± 11.09	20.21 ± 7.91
2nd order	10.39 ± 7.50	5.94 ± 4.49
3rd order	9.04 ± 6.77	4.87 ± 3.66

mode complexity. The error in Y direction is almost half of that of X direction which is due to the head movement range in X direction is much larger than the range in Y direction. The interference of distortion is therefore more significant in X direction. The accumulated eye gaze mapping is visualised in Fig. 6, which clearly demonstrates the fine performance of the proposed system.

C. Use Case

After established the model from the vehicle experiment, the participant was requested to conduct three NDAs including playing games on cell phone, watching movie on tablet and reading a book. This use case aims to demonstrate how to use this system to identify the NDAs with the support of object recognition. As illustrated by Fig. 7, the gazes are successfully estimated and mapped on the regions of cell phone, tablet and book. Two videos are available for download from the link provided in the Acknowledgment.

IV. DISCUSSION

In the proposed system, the order of model and the number of model term determine the model complexity which affects the model performance. Table VI presents the model performance based on the different model orders, where the number of model term was set as 20. It can be observed that the RMSE in X direction has been reduced from 56.49 pixel to 10.39 pixel, equal to 81.6% improvement of accuracy, when the 2nd order model is used instead of the linear model. When the model complexity is increased from 2nd order to 3rd order, the increment of mode performance is much less significant (13% improvement). A similar pattern has been observed in Y direction. On the one hand, the model is preferred

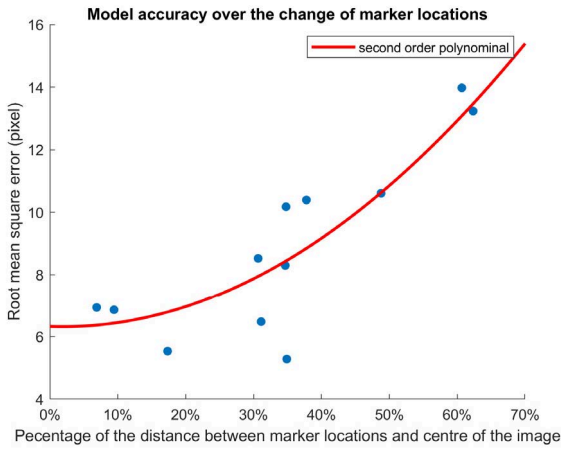


Fig. 8. The model accuracy change of the location of the markers, which suggests the influence of t distortion.

to be as simple as possible to a) ensure low computational time for real-time applications, and b) avoid the over-fitting problem. On the other hand, the model should be sophisticated enough to cope the interference of camera distortion and head movement. For all tests conducted in this study, the 2nd order nonlinear is appropriate. However, the optimal model order can be different if a different camera is used. Generally speaking, a camera with high distortion requires a high order of model and more number of model terms. All these observations can be applied to the number of selected model term. It is suggested to select the model as simple as possible as long as the error of estimation is smaller than the markers. If a high resolution of eye gaze mapping is required, smaller markers should be used.

The selection of the markers' location affects the system performance. In the vehicle experiment, some strategic locations were chosen such as windscreen, wing mirrors, steer wheel and dashboard aiming to cover popular areas which the driver is often gaze on. Fig.8 plots the RMSE of estimated gaze on markers against the percentage of the distance from makers to the centre of image to the image size. It can be seen that there is an average error around 6 pixels for the markers around the center, and the error increases following the increment of distance with an approximately quadratic relationship. This observation is a clear evidence that the model performance is affected by distortion of lens. Apart from the distortion,

another reason of relatively poor performance on the edge of the image is caused by the OpenFace algorithm. When the driver gazes on the area around the edge of the image from Camera-2, the head rotation is usually large. The accuracy of facial features extracted by Camera-1 is compromised because some landmarks are hidden or partly visible. Using multiple cameras to capture the driver's facial features can address the problem but will increase the complexity and cost of the system.

V. Conclusion

This paper proposed a low-cost and non-intrusive eye gaze mapping system based on two cameras, which could act as a powerful tool to reasoning the driver's visual attention. A nonlinear polynomial model was proposed to establish the relationship between the driver's facial features from Camera-1 and the eye gaze on images from Camera-2. Both indoor experiment and in-vehicle experiment qualitatively and quantitatively demonstrated the efficiency of the proposed system. The system was then successfully applied to characterise three common NDAs including playing phone, reading book and playing tablet. There are a few key findings from this study.

1. The error of this system in the in-vehicle experiment for the X and Y direction is 7.80±5.99 pixel and 4.64±3.47 pixel respectively with the image resolution of 1440×1080 pixels.
2. A high order nonlinear model can reduce the interference of distortion caused by Camera-2.
3. Apart from the gaze related parameters, the head pose related parameters must be considered in the model due to the complexity and diversity of eye gaze shifting.
4. Increasing the complexity of head movement will reduce the model performance. Including the object depth into the model may improve the performance, which requires a further study.
4. The model could be subjective. A calibration process therefore is suggested for each driver before testing. Achieving a generic model to remove the calibration process requires further studies.

It should be noted that this paper focuses on the development of the eye gaze mapping system not its applications. Potential problems to apply it in a driving vehicle include that (a) facial feature extraction will be compromised due to the potential heavy movement of driver body and camera movement caused by poor road condition; (b) it will be more difficult for the driver to gaze on an object. A comprehensive and systemic evaluation of its performance in driving vehicles requires a further study. The developed system could impact studies on NDA characterisation for intelligent design of take over strategy, driving environment awareness for current and future automated vehicles.

Acknowledgment

For access to the use case videos underlying this paper, please see the Cranfield University repository, CORD, at DOI: https://doi.org/10.17862/cranfield.rd.8506451.v1

References

[1] Y. Zhao *et al.*, "An orientation sensor-based head tracking system for driver behaviour monitoring," *Sensors*, vol. 17, no. 11, p. 2692, Nov. 2017.

[2] *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE International Standard J3016_201806, 2018.

[3] M. Sivak and B. Schoettle, "Motion sickness in self-driving vehicles," Transp. Res. Inst., Ann Arbor, Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep. UMTRI-2015-12, Apr. 2015.

[4] C. Ahlstrom, K. Kircher, and A. Kircher, "A gaze-based driver distraction warning system and its effect on visual behavior," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 965–973, Jun. 2013.

[5] M. Lundgren, L. Hammarstrand, and T. McKelvey, "Driver-gaze zone estimation using Bayesian filtering and Gaussian processes," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 10, pp. 2739–2750, Oct. 2016.

[6] P. Jiménez, L. M. Bergasa, J. Nuevo, N. Hernández, and I. G. Daza, "Gaze fixation system for the evaluation of driver distractions induced by IVIS," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1167–1178, Sep. 2012.

[7] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intell. Syst.*, vol. 31, no. 3, pp. 49–56, May/Jun. 2016.

[8] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[9] A. Doshi and M. M. Trivedi, "On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes," *IEEE Trans. Intell. Transp. Syst.*, vol. 10, no. 3, pp. 453–462, Sep. 2009.

[10] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, 2017.

[11] S. H. Kwon and M. Y. Kim, "Selective attentional point-tracking through a head-mounted stereo gaze tracker based on trinocular epipolar geometry," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2015, pp. 1617–1621.

[12] X. Yuan, Q. Zhao, D. Tu, and H. Shao, "A novel approach to estimate gaze direction in eye gaze HCI system," in *Proc. 5th Int. Conf. Intell. Hum.-Mach. Syst. Cybern.*, vol. 1, 2013, pp. 588–591.

[13] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 3, pp. 478–500, Mar. 2010.

[14] N. Mizuno, A. Yoshizawa, A. Hayashi, and T. Ishikawa, "Detecting driver's visual attention area by using vehicle-mounted device," in *Proc. IEEE 16th Int. Conf. Cogn. Inform. Cogn. Comput. (ICCI*CC)*, Jul. 2017, pp. 346–352.

[15] F. Vicente, Z. Huang, X. Xiong, F. D. L. Torre, W. Zhang, and D. Levi, "Driver gaze tracking and eyes off the road detection system," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2014–2027, Aug. 2015.

[16] J. Jo, S. J. Lee, J. Kim, H. G. Jung, and K. R. Park, "Vision-based method for detecting driver drowsiness and distraction in driver monitoring system," *Opt. Eng.*, vol. 50, no. 12, 2011, Art. no. 127202.

[17] R. A. Naqvi, M. Arsalan, G. Batchuluun, H. S. Yoon, and K. R. Park, "Deep learning-based gaze detection system for automobile drivers using a NIR camera sensor," *Sensors*, vol. 18, no. 2, p. 456, Feb. 2018.

[18] H. Mohsin and S. H. Abdullah, "Pupil detection algorithm based on feature extraction for eye gaze," in *Proc. 6th Int. Conf. Inf. Commun. Technol. Accessibility (ICTA)*, 2017, pp. 1–4.

[19] G. Andrienko, N. Andrienko, M. Burch, and D. Weiskopf, "Visual analytics methodology for eye movement studies," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2889–2898, Dec. 2012.

[20] X. Jin, Z. Li, J. Zhang, and X. Yang, "Research on pupil center localization in eye gaze tracking system," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 3211–3215.

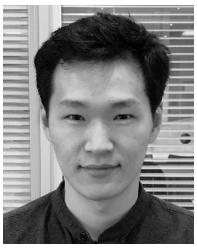
[21] S. M. Kim, M. Sked, and Q. Ji, "Non-intrusive eye gaze tracking under natural head movements," in *Proc. 26th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 3, 2004, pp. 2271–2274.

[22] L. Fridman, J. Lee, B. Reimer, and B. Mehler, "A framework for robust driver gaze classification," in *Proc. SAE World Congr. Exhib.*, 2016.

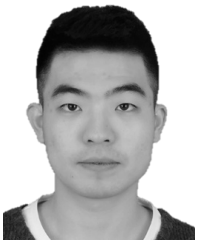
[23] X. Wu, J. Li, Q. Wu, and J. Sun, "Appearance-based gaze block estimation via CNN classification," in *Proc. IEEE 19th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2017, vol. 38, no. 4, pp. 1–5.

[24] A. M. Soccini, "Gaze estimation based on head movements in virtual reality applications using deep learning," in *Proc. IEEE Virtual Reality (VR)*, Mar. 2017, pp. 413–414.

[25] W. Cui, J. Cui, and H. Zha, "Specialized gaze estimation for children by convolutional neural network and domain adaptation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3305–3309.
[26] D. Xiao and C. Feng, "Detection of drivers visual attention using smartphone," in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, 2016, pp. 630–635.
[27] S. Vora, A. Rangesh, and M. M. Trivedi, "Driver gaze zone estimation using convolutional neural networks: A general framework and ablative analysis," *IEEE Trans. Intell. Veh.*, vol. 3, no. 3, pp. 254–265, Sep. 2018.
[28] C. Hennessey, B. Noureddin, and P. Lawrence, "A single camera eye-gaze tracking system with free head motion," in *Proc. Symp. Eye Tracking Res. Appl. (ETRA)*, vol. 1, Mar. 2006, p. 87.
[29] A. T. Duchowski, "A breadth-first survey of eye-tracking applications," *Behav. Res. Methods, Instrum., Comput.*, vol. 34, no. 4, pp. 455–470, Nov. 2002.
[30] M. L. Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 4, pp. 322–336, Apr. 2000.
[31] E. Skodras and N. Fakotakis, "Precise localization of eye centers in low resolution color images," *Image Vis. Comput.*, vol. 36, pp. 51–60, Apr. 2015.
[32] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[33] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial behavior analysis toolkit," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 59–66.
[34] T. Baltrusaitis, P. Robinson, and L.-P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Jun. 2013, pp. 354–361.
[35] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling, "Rendering of Eyes for Eye-Shape Registration and Gaze Estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3756–3764.
[36] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4511–4520.
[37] R. Netzel and D. Weiskopf, "Hilbert attention maps for visualizing spatiotemporal gaze data," in *Proc. IEEE Second Workshop Eye Tracking Vis. (ETVIS)*, Oct. 2016, pp. 21–25.



Lichao Yang was born in Shanxi, China. He received the M.Sc. degree in automotive mechatronics from Cranfield University, Cranfield, U.K., in 2018, where he is currently pursuing the Ph.D. degree in driver non-driving activities analysis with the Through-Life Engineering Services Centre.



Kuo Dong was born in Chengde, China. He received the B.Eng. degree from the School of Automotive Engineering, Chongqing University, China, in 2017, where he is currently pursuing the M.E. degree with the Chongqing Automotive Collaborative Innovation Centre. He is also a Visiting Research Student with the School of Aerospace, Transport and Manufacturing, Cranfield University.

Arkadiusz Jan Dmitruk was born in Ostrowiec Świętokrzyski, Poland. He received the B.S.E. degree in mechatronics from the Warsaw University of Technology, Poland, in 2016, and the M.Sc. degree in computational and software techniques in engineering from Cranfield University, U.K., in 2018.

He is currently a Research Assistant with Cranfield University. His interests include computer vision, automotive mechatronics, and artificial intelligence with engineering applications.

James Brighton is currently a Professor in automotive engineering with the School of Aerospace, Transport and Manufacturing, Cranfield University. He heads the Advanced Vehicle Engineering Centre and has more than 22 years' experience relating to on and off road vehicle dynamics, terra-mechanics, tyre and track system modeling, vehicle instrumentation, and lightweight material structures and his current clients span the globe. He teaches motorsport vehicle dynamics and currently leads research programs in the field of autonomous vehicle development, vehicle light-weighting, and off road vehicle dynamics and his team is able to offer a wide range of vehicle related technical solutions from fundamental research through product design and prototype vehicle subsystem manufacture, supply, evaluation, and testing across a wide range of ground vehicle applications.

Yifan Zhao was born in Zhejiang, China. He received the Ph.D. degree in automatic control and system engineering from The University of Sheffield, U.K., in 2007.

He is currently a Senior Lecturer in data science with Cranfield University. His research interests include computer vision for automated vehicles, super resolution, active thermography, and nonlinear system identification.

A dual-cameras-based driver gaze mapping system with an application on non-driving activities monitoring

Yang, Lichao

2019-09-13

Yang L, Dong K, Dmitruk AJ, et al., (2020) A dual-cameras-based driver gaze mapping system with an application on non-driving activities monitoring. IEEE Transactions on Intelligent Transportation Systems, Volume 21, October 2020, pp. 4318-4327.

https://doi.org/10.1109/TITS.2019.2939676