Pattern Recognition, Volume 116, August 2021, Article number 107955
DOI: 10.1016/j.patcog.2021.107955

Recognition of visual-related non-driving activities using a dual-camera monitoring system

Lichao Yang[a], Kuo Dong[b], Yan Ding[c], James Brighton[a], Zhenfei Zhan[b] and Yifan Zhao[a, *]

[a]School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield, UK
[b]Chongqing Automotive Collaborative Innovation Centre, Chongqing University, No.174 Shazheng St., Shapingba District, Chongqing, 400044 China
[c]Key Laboratory of Dynamics and Control of Flight Vehicle, Ministry of Education, School of Aerospace Engineering, Beijing Institute of Technology, Beijing 100081 China
*The corresponding author: Y. Zhao (yifan.zhao@cranfield.ac.uk)

Abstract

For a Level 3 automated vehicle, according to the SAE International Automation Levels definition (J3016), the identification of non-driving activities (NDAs) that the driver is engaging with is of great importance in the design of an intelligent take-over interface. Much of the existing literature focuses on the driver take-over strategy with associated Human-Machine Interaction design. This paper proposes a dual-camera based framework to identify and track NDAs that require visual attention. This is achieved by mapping the driver's gaze using a nonlinear system identification approach, on the object scene, recognised by a deep learning algorithm. A novel gaze-based region of interest (ROI) selection module is introduced and contributes about a 30% improvement in average success rate and about a 60% reduction in average processing time compared to the results without this module. This framework has been successfully demonstrated to identify five types of NDA required visual attention with an average success rate of 86.18%. The outcome of this research could be applicable to the identification of other NDAs and the tracking of NDAs within a certain time window could potentially be used to evaluate the driver's attention level for both automated and human-driving vehicles.

Keywords

Driver behaviour, level 3 automation, computer vision, deep learning, activities identification

Declarations of interest: none

1. Introduction

In a Level 3 automated vehicle, according to the SAE (J3016) Automation Levels definition, the driver could engage in some non-driving activities (NDAs) or non-driving related (NDR) activities when the vehicle is under the automated driving mode [1]. However, since the level of full automation has not been reached, the driver is still expected to respond appropriately to a takeover request from the vehicle [2]. NDAs could affect the driver's hazard awareness and a high attention level on the NDA could result in a negative effect on driving quality [3] or even accidents during the transition of control between the vehicle and driver [4,5]. Therefore, the effect of various NDAs on takeover quality needs to be investigated and evaluated. In recent years, some studies have been reported to evaluate the take-over performance (e.g. reaction time and driving quality) after switching from NDAs [6,7] and the effects of Human-Machine Interaction (HMI) design supporting this activity switch and monitoring the driving environment [8]. There is very limited literature investigating the identification and tracking of NDAs automatically. Sivak and Schoettle [9] reported that the main NDAs that drivers engage in the UK are reading (9.9%), sleeping (9.4%), texting or talking with friends (7.1%), working (6.4%) and watching movies (5.4%). Since NDAs in a level 3 and above automated vehicle are diverse and the type and engaged duration of NDAs will lead to different take-over performance [4], it is necessary to classify/identify and track them automatically for designing an intelligent HMI for takeover.

As objects and human poses are feature-rich, human-object interaction has been widely investigated in the early stages of human action recognition [10,11], through the integration of object recognition, pose estimation and action identification [12]. For successful NDA identification, the driver is constrained on the seat, space limitation and body occultation pose a challenge for driver action estimation. Le *et al.* [13] proposed a convolutional neural network (CNN)-based approach to achieve the driver behaviour parsing. It localises some body parts of

the driver like head, hand, etc. by semantic segmentation in still images to achieve the detection of some actions, such as hands on steering wheel and hands on phone. Several deep learning-based approaches have been proposed for video-based human action recognition, with the development of artificial intelligence in multi-object detection [10]. Such approaches extend object detection to action detection through the multi-stream CNN, which combines the spatial and temporal information [14,15]. It recognises the action by using the moving parts of the human body instead of pose estimation. Some CNN-based approaches have been proposed for the NDA or secondary task recognition in recent years. Xing *et al.* [16] used the image of the driver's body by removing the background, as the input of the CNN model to recognise NDAs. Eraqi *et al.* [17] extended inputs by including raw images, skin-segmented images, face images, hands images, and "face+hands" images. Then trained CNN model for each stream is further used to obtain the final prediction using a genetic algorithm based on their outputs. Yang *et al.* [18] proposed a 2-stream CNN based system, which extracts the spatial features from raw images and the movement features from the corresponding optical flow images to achieve the NDA recognition. However, such CNN-based NDA recognition approaches mainly focus on encoding the specific movement of the driver. It lacks the capability of tracking the driver's visual attention. Studying the driver's visual attention can directly determine whether the driver is engaging with NDAs, which is important to develop an intelligent HMI design for a safe take-over. As most NDAs (e.g. reading, texting, working and watching movies) require interaction between objects (e.g. book, tablet, or dashboard) and the human eye (gaze), this paper proposes a novel framework for gaze-related NDA identification and tracking that consists of three parts: object recognition, gaze estimation and an activity classifier. Several object detection frameworks have been proposed such as YOLO [19] and Faster R-CNN [20]. These frameworks can recognise a few general objects in real-time, but they lack semantic segmentation and accurate outliner detection. Since object segmentation is needed in this

proposed framework, the Mask R-CNN [21] is used as the part of the proposed framework. The eye gaze features have been applied in some applications of advanced driver-assistance systems (ADAS) for the purpose of distraction and fatigue detection [22,23] or gaze attention estimation [24]. The developed gaze estimation systems mainly focus on the modelling of the eye-gaze based on the image captured by the camera, which is in front of the human face [25,26]. Since the image used in these systems have no further information about the activity that the driver could engage with. It can not be used directly for driver behaviour recognition. The applications of gaze estimation for ADAS are normally driver gaze zone estimation [27]. Fridman *et al.* [28] allocated the driver's gaze into different regions by extracting their facial features with a single camera. Xiao and Feng [29] proposed a driver's visual attention system by using a smartphone, in this method, the rear camera is used to capture the moving object and the front camera is used to estimate the driver's gaze. The view of the rear camera is divided into 9 zones, and the system aims to check if the driver is aware there is a moving object inside these zones. Both studies made a fixed assumption between the eye gaze direction and the driver's behaviour, which is not applicable for characterisation of NDAs due to its high complexity and uncertainty. Therefore, to further implement the gaze estimation method into NDAs recognition, the estimated gaze need to be mapped into a view, which contains the driver's behaviour in the vehicle cabin.

This paper presents a non-intrusive and cost-effective dual-camera based NDA identification and tracking framework. It maps the driver's eye gaze, achieved by the first/front camera facing the driver, and the object scene is captured by the second/rear camera, using a complex system modelling technique, called Volterra Non-linear Regressive with eXogenous inputs (VNRX) model. The object is automatically recognised and located through the Mask R-CNN algorithm. Based on the mapped gaze and the location of the segmented object, an

activity classier using the sliding time window technique is proposed to identify and track the type of NDA.

2. Methodology

2.1. Framework architecture

The proposed framework has 3 components: gaze mapping, object recognition and an NDA classifier. As shown in Fig. 1, the driver's gaze is estimated by a dual-camera system. The front camera is used to capture and extract the driver's facial and gaze features which are used to estimate the gaze which is then mapped into the scene of the rear camera and visualised by a heat map. The estimated gaze location in the scene of the rear camera helps define a region of interest (ROI) for the object recognition, which will significantly reduce the recognition time and increase the accuracy and success rate. The recognition result shows the object label, confidence score and location of each object represented by an object-mask list. The sliding time window technique is used to construct a novel NDA classifier for decision-making through considering the historic information of eye gaze location and recognised object-masks. The details of each part are introduced below.

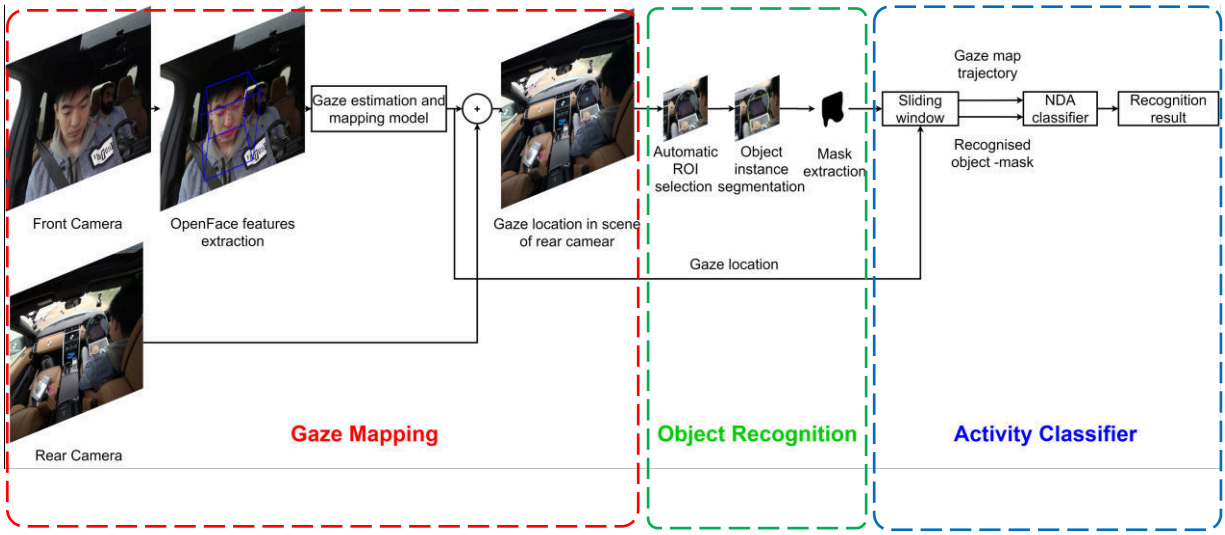

Fig. 1. The proposed framework for NDA identification that consists of three parts: gaze mapping, object recognition and activity classifier.

2.2. Gaze estimation

Two Garmin Virb Action Cameras were used due to the fine performance on image stabilisation. Video images were captured by the front camera using a resolution of 1024×768 pixels and 25 frames per second (fps). The rear camera was set at a resolution of 1920×1440 pixels with a wide field of view at 25 fps.

This gaze estimation system has four elements: video acquisition, feature extraction, gaze mapping and heat map visualisation. The facial features captured from the front camera can be extracted through OpenFace [30] which is an open-source facial analysis tool. The selected face features include the driver's head related features, such as head position and orientation both in 3D dimension, and gaze related features, such as gaze directions in x and y directions. It should be noted that the gaze direction is in the unit of radian in the world coordinate, which is averaged for both eyes. If the driver looks from left to right, the gaze angle in x direction will change from positive to negative; if the driver looks from up to down, this will result in the gaze angle in y direction changing from negative to positive; both of the gaze angles will be close to 0, if the driver looks straight ahead.

To establish the corresponding between two image planes, additional information is required, such as location, intrinsic and extrinsic parameters of the two cameras. To simplify the calibration process, we propose a system identification approach based on orthogonal least squares (OLS) algorithm to directly map the inputs (facial information in the image plane of the front camera) into the output (the gaze location in the image plane of the rear camera) without the aforementioned additionally information. Furthermore, there are many facial and eye features extracted from the front camera. If all features are used the model could be overfitting, which will lead to the poor performance of prediction. Filtering of inputs must be taken place to only select the significant facial information for eye gaze mapping. The below-proposed method 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. Moreover, the capability to accommodate nonlinear modelling is important to cope with 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 the rear camera. 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:

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

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

where $X_1, X_2, \dots, X_n$ 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:

$$Y(k) = \sum_{i=0}^{N} \theta_i p_i(k), k = 1,2, \dots, M \tag{3}$$

where $Y(k)$ is the system output (eye gaze location in x or y direction), $p_i(k)$ 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 $p_i(k)$ select from, denoted by C, can be expressed

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

where C_1 is the linear term set, expressed as

$$C_1 = \cup_{a=1}^{n} X_a \tag{5}$$

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

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

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

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

Eq. (3) is re-written as

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

where

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

and $P^T(k) = \left(p_1(k), p_2(k), \ldots, p_N(k)\right)$. 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. Eq. (4) is then rewritten as

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

where $G = A\Theta = [g_1 \quad g_2 \quad \cdots \quad g_N]^T$. Eq. (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 $w_1(k) = p_1(k)$, 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 $j = 2,3, \dots, M$, 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 $i = 1,2, \dots, j-1$. 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 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

$$\left.\begin{aligned} \hat{\theta}_N &= \hat{g}_N \\ \hat{\theta}_i &= \hat{g}_i - \sum_{k=i+1}^{N} a_{ik} \theta_k, i = N-1, \dots, 1 \end{aligned}\right\} \tag{17}$$

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. The details of estimating the VNRX model from the calibration data can be found in [31]. The estimated gaze can be calculated using the extracted features and gaze-mapping model, and the estimated gaze location can be calculated and mapped into the scene of the rear camera for further heat map visualisation.

There are two processes in this system: calibration process and testing process. In the calibration process, the driver needs to gaze on a few markers placed on certain locations

strategically. Since the locations of the markers in the rear camera are known, a model can then be established between these locations and the corresponding extracted features. The trained model is used to estimate the gaze purely based on extracted features in the testing process.

2.3. Object recognition

CNN-based algorithms have achieved critical advances for the object recognition problem. The object recognition models based on CNNs can be categorised into two different types: one-stage and two-stage. Two-stage models such as Faster R-CNN [20] and Mask R-CNN [21] usually produce higher accuracy than one-stage models such as YOLO [19] and SSD [32] but they perform under lower detection speed. To match the requirement of this framework, this paper selects the Mask R-CNN model, which is an extension of the Faster R-CNN model for pixel-to-pixel instance segmentation task. There are two reasons for this selection: 1) compared with the recent frameworks training the COCO dataset, Mask R-CNN outperforms Faster R-CNN, YOLO and SSD in terms of accuracy and the speed is acceptable; 2) Mask R-CNN extends previous frameworks and locates exact pixels of each object instead of only bounding boxes, which is important for this study because the region of the object must be accurate to determine if the eye gaze is located in this region.

Instance segmentation is a challenging task that combines two independent processes: object detection and semantic segmentation. The multi-task scheme could create spurious edges and produce systematic errors in overlapping instances [33]. To solve this problem, Mask R-CNN extends Faster R-CNN by adding a branch for predicting segmentation masks in a pixel-to-pixel manner, in parallel with the existing branch for classification and bounding box regression. The core operation in Faster R-CNN for attending to instances, RoIPool, performs coarse spatial quantisation for feature extraction [34]. To fix the misalignment, Mask R-CNN replaces the RoIPool layer with a simple and quantisation-free layer which is called RoIAlign and faithfully preserves exact spatial locations. The RoIAlign layer uses bilinear interpolation

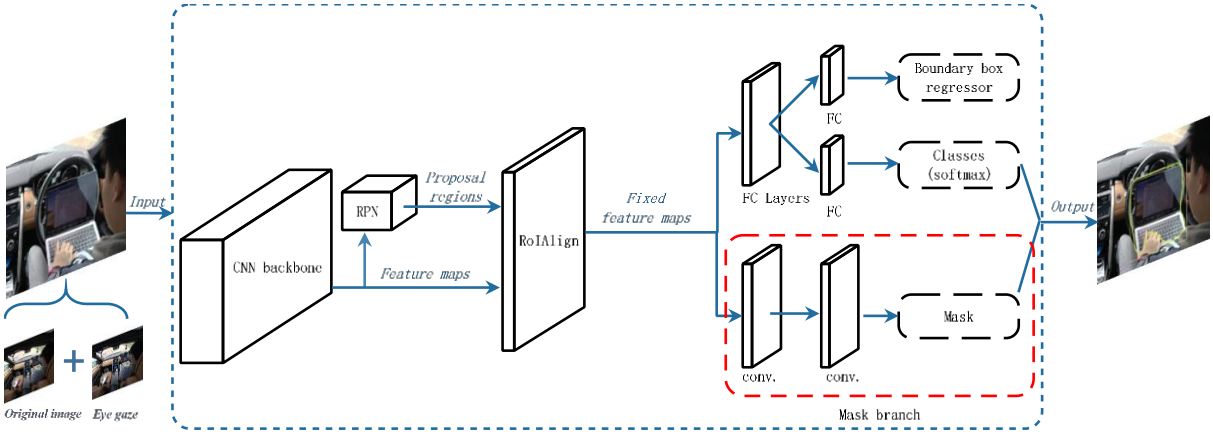

Fig. 2. The proposed Mask R-CNN architecture for the object recognition.

to compute the exact values of the input features at four regularly sampled locations in each ROI bin and then performs max or average pooling on features. In spite of being a seemingly minor change, RoIAlign improves mask accuracy significantly. The proposed Mask R-CNN architecture is illustrated by Fig. 2. It should be noted that the input image is a cropped image considering the eye gaze. The size of this ROI is a parameter to set considering the size of the targeted objects.

There are several implementations of Mask R-CNN so far. This paper selected the *maskrcnn_benchmark* for the proposed system due to its best performance in training and inference. Maskrcnn_benchmark is up to twice as fast as a *Detectron* while matching and exceeding *Detectron* accuracy [35]. There are 5 NDAs (involving 5 types of object) considered in this paper where phone, laptop, and book can be detected with the COCO-pre-trained model, but a tablet and car interior cannot be detected with this pre-trained model. A dedicated database was then created for the latter cases. The dataset consists of 200 images acquired from the rear camera for each object including tablet, control console, wing mirror, windscreen, rear-view mirror, and dashboard. We selected the ResNet-101-FPN as the backbone network and trained these data starting with a learning rate of 0.001 and ending with a learning rate of 0.0001. All training works were implemented on an NVIDIA Quadro P6000 graphics card machine which has 24 GB DDR5X memory.

2.4 Activity classifier

The hypothesis of this study is that activities required visual attention can be identified by estimating the driver's gaze and recognising the object that is gazed on. Therefore, the inputs of the classifier are the representation of the driver's gaze map and the recognised object-masks (there could be multiple objects in the ROI). Considering that the driver's behaviour during the engagement of NDA is continuous, the historical temporal information is crucial for activity recognition. The proposed classifier employs the sliding time window technique to enhance the resilience to noise.

Consider the time-series $Z(t) = \{Z_x(t), Z_y(t)\}$ where $Z(t)$ represents the gaze location at the time t, estimated from Eqs. (1)-(2). In this study, the driver's gaze is assumed as a form of the circle. The centre of the circle is the estimated gaze location $Z(t)$ and the diameter is denoted by d. The spatial distribution inside the circle follows the Gaussian distribution to represent the driver's observation intensity [36]. At the time t, the intensity of the pixel (x, y) in the gaze map, denoted by $S(x, y, t)$, where $2 * |x - Z_x(t)| \leq d$ and $2 * |y - Z_y(t)| \leq d$, can be expressed as:

$$S(x, y, t) = \mathrm{e}^{\left(-\frac{(x-Z_x(t))^2+(y-Z_y(t))^2}{2\sigma^2}\right)} * 100\% \tag{18}$$

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

For the current sliding time window with a size of H, assuming that the weight of gaze at each time follows the linear relationship to the time, the trajectory of gaze for this time window can be computed as:

$$S_c(x, y, t) = \frac{1}{H}\sum_{i=0}^{H-1} S(x, y, t - i) * (1 - \frac{i}{H}) \tag{19}$$

For further decision making, the trajectory of gaze is binarized as:

$$\bar{S_c}(x,y,t) = \begin{cases} 1 & if\ S_c(x,y,t) \geq T_g \\ 0 & otherwise \end{cases} \quad (20)$$

where T_g is the threshold.

The object could be easily occulted by the driver's hand during the NDA engagement, which could lead to the poor performance of object recognition. To increase the robustness of overall performance, the recognised object-masks are estimated based on the historical information within the time window. The list of the segmented mask of objects at the time t, expressed as $N(x,y,t) = \{N_1(x,y,t), N_2(x,y,t), \dots, N_k(x,y,t)\}$, is a set of binary images, where k is the total number of recognised objects. Since the ROI to produce $N(x,y,t)$ is selected based on the gaze location $Z(t)$, to create an object-mask within a time window H, an offset needs to be considered as the selected ROI could be different for each time step. The revised mask for the i^{th} object by removing the offset can be expressed as:

$$N_{\text{offset},i}(x,y,t) = N_i\big(x - Z_x(t), y - Z_y(t), t\big) \quad (21)$$

The final recognised object-mask for the i^{th} object is achieved by calculating the union of all masks for this object within this widow. It can be expressed as

$$ON_i(x,y,t) = \bigcup_{j=0}^{H-1} N_{\text{offset},i}(x,y,t-j) \quad (22)$$

Finally, the intersection of the binarized trajectory of gaze $\bar{S_c}(x,y,t)$ and each recognised object-mask $ON_i(x,y,t)$ is calculated. The one that has the maximal area of intersection is selected as the recognised class l, which can be written as

$$l(t) = \arg\max_{1 \leq i \leq k} \|\bar{S_c}(x,y,t) \cap ON_i(x,y,t)\| \quad (23)$$

where $\|\cdot\|$ indicates the operation to calculate the area.

2.5 Experiment design

A Land Rover Discovery 5 was used as the test vehicle, this remained stationary in the experiment. Two cameras and eight markers were mounted for the gaze mapping process. The front camera was mounted on the windscreen and faced towards the driver to capture the driver's facial and gaze features. The rear camera was located on the roof of the vehicle between two frontal seats and towards the windscreen. The markers were allocated strategically on the windscreen, dashboard, centre of the steering wheel, wing mirrors, and centre of the centre console.

At the start of the experiment, the 6 participants were requested to gaze at the markers one by one lasting for at least 6 seconds for each marker. After the calibration process, the participants were asked to conduct 5 selected NDAs which were reading books, watching movies with a cell phone, sending an E-mail by using a laptop, playing games with a tablet and interacting with the centre console to select a radio channel. Each NDA lasted for at least 1

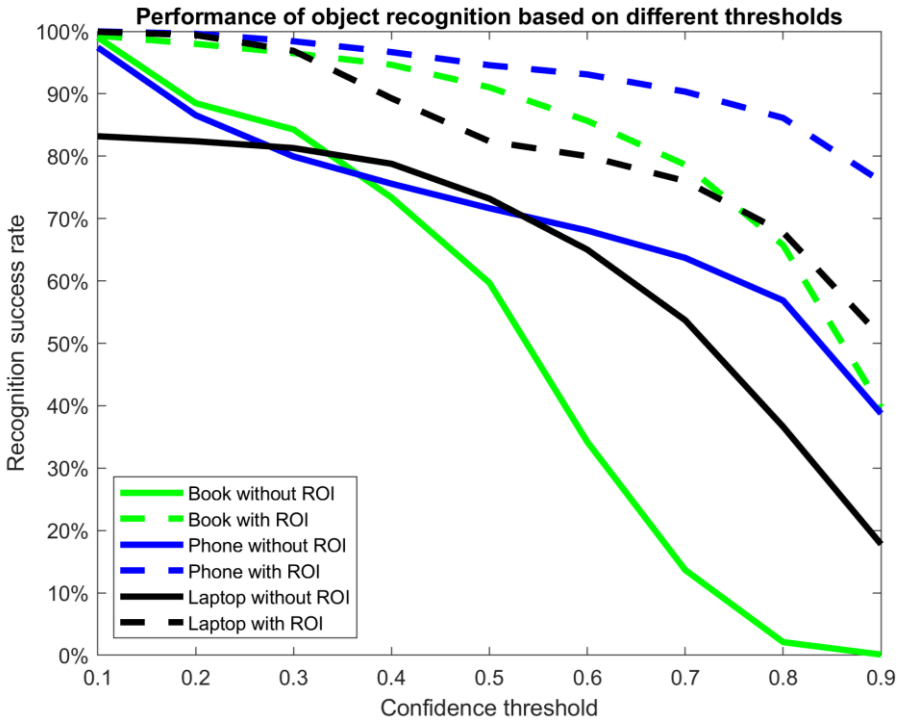

Fig. 3. Comparison of object recognition performance based on the different confidence threshold for book, phone and laptop.

minute. An audio guide was provided to ensure consistency of the experimental process across all participants.

3. Results

As the participants needed some time for NDAs transition, only the middle part of the period (40 seconds or 1000 frames) for each NDA was used for recognition and tracking. The parameters used to create the gaze map were set as d = 40 pixels for estimated gaze region, the variance σ = 0.8 for gaze distribution and T_g = 128 for the threshold of the gaze binarization.

3.1 Object recognition

A pre-trained Mask R-CNN model provides 81 categories based on the COCO dataset, which was used to recognise book, cell phone and laptop. The automatic ROI (with a size of 640×480 pixels) selection module based on the mapped gaze location was applied in this framework. In Fig. 3, the dotted lines represent the object recognition performance with the automatic selection of ROI against different values of the confidence threshold while the solid lines represent those without ROI. It can be clearly seen that the recognition success rate with ROI is consistently higher than that without ROI for all range of the confidence threshold across three types of object. This is probably due to the reduced interference of other objects in the raw image. It is expected that following the increase of the confidence threshold, the recognition success rate decreases. It is shown in Fig. 3 that when the threshold is 0.6, the recognition success rates with ROI for all 3 kinds of object are above 80%. A low confidence threshold leads to a high risk of misrecognition of object, which could result in the accuracy decrease of NDA identification, and it will also increase the system computational cost. A high confidence threshold leads to a high risk of missing the targetted object, which results in the failure of NDA identification. To balance this trade-off, the confidence threshold for the below results was set as 0.6.

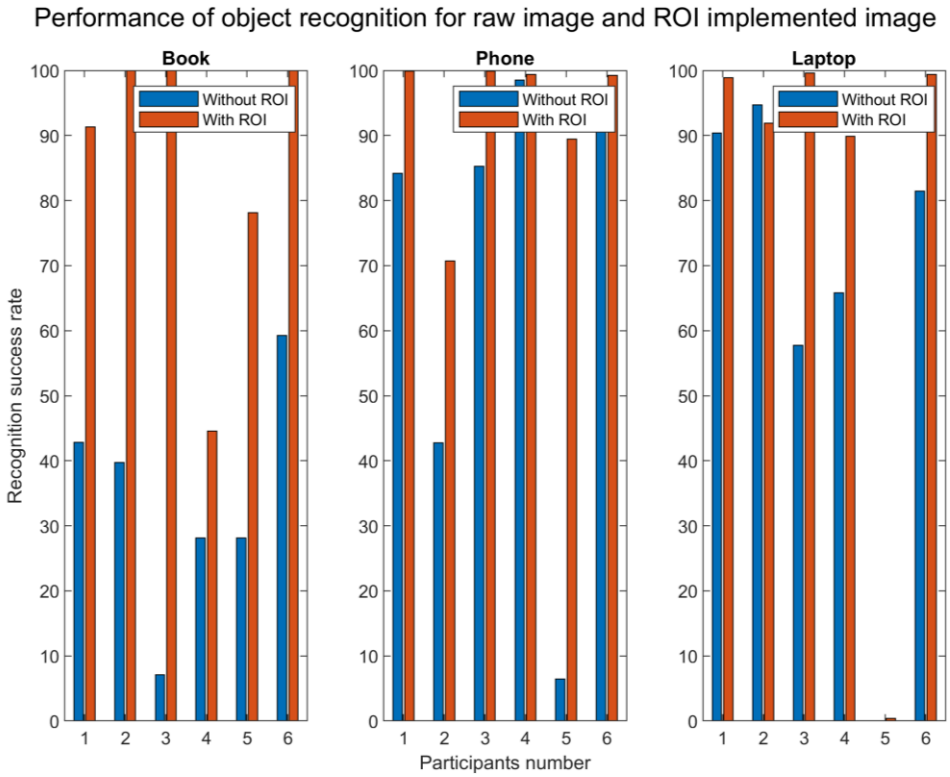

Fig. 4. Object recognition performance comparison for raw image and ROI implemented image based on all participants

Fig. 4 plots the performance comparison of object recognition for images with and without ROI for each participant. An increase in the success rate is shown and is significant for all participants and objects. It should be noticed that the success rate of participant 4 with ROI for book recognition is around 44%, the reason is addressed below. The success rates of participant 5 for laptop recognition are almost 0% due to a heavy occlusion caused by cloth.

Table 1 Comparison of performance of object recognition for data with and without ROI.

Object	**Success rate**		**Processing time per image (s)**	
	Without ROI	**With ROI**	**Without ROI**	**With ROI**
Book	34.20%	85.66%	0.462	0.191
Phone	68.07%	93.10%	0.443	0.181
Laptop	65.01%	80.00%	0.476	0.183
Average	55.76%	86.25%	0.460	0.185

Table 1 shows the overall performance of both accuracy and processing time by averaging all participants. Phone recognition shows the highest success rate (93.10%), probably due to the smallest size of the object. From phone, book, to laptop, the object size becomes bigger, and it is observed that the success rate becomes lower. This is because a large object has a high possibility of being covered by the human body, which leads to high dissimilarity with the training data. An increment of more than 30% has been observed in terms of the average success rate with the gaze-based ROI detection implemented. The average processing time for all objects decreases from 0.460 s to 0.185 s with a time reduction percentage of 60%. There is no significant difference in processing time in terms of the type of object.

For a tablet and car interior, since a dedicated training database was developed for a specific tablet and vehicle only, the success rates are almost 100%. To extend its application on other types of tablet and vehicle, much more training data are required, which is not the focus of this research.

3.2 NDAs identification and analysis

The recognised object mask and the gaze trajectory map were used to identify the type of NDA that the driver is engaging with. Table 2 presents the identification accuracy of 5 tested NDAs for all participants. For the NDA of reading a book, the average accuracy is 85.04%

Table 2 NDAs identification accuracy for all participants

NDAs	Participants						Average
	1	2	3	4	5	6	
Reading a book	99.80%	100.00%	88.74%	43.07%	84.67%	93.94%	85.04% ± 19.56%
Playing a phone	100.00%	72.78%	92.26%	91.14%	87.40%	97.08%	90.11% ± 8.75%
Working on a laptop	96.35%	77.98%	81.51%	79.96%	6.26%	96.56%	73.10% ± 30.82%
Playing a tablet	100.00%	100.00%	98.68%	100.00%	99.17%	100.00%	99.64% ± 0.53%
Interacting with centre console	96.66%	86.17%	80.09%	70.84%	76.54%	87.71%	83.00% ± 8.34%

with a standard deviation of 19.56%. The high value of standard deviation is caused by the result of participant 4 which is only 43.07%. High identification accuracy has been achieved for the NDAs of playing phone and playing tablet, with an averaged value of 90.11% and 99.64% respectively. The standard deviations across participants are less than 10%. The performance of the NDA of working on a laptop is more than 85% except participant 5 caused by the failure of object recognition due to cloth occlusion. The average accuracy of the NDA of interacting with the centre console is 83.00%, which is relatively lower than others and the reason that will be explained below.

Table 3 Analysis for the failed cases, where ER and EM refer to the error caused by recognition failure and unmatched gaze map and object mask, respectively.

NDAs	Participants											
	1		2		3		4		5		6	
	ER	EM	ER	EM	ER	EM	ER	EM	ER	EM	ER	EM
Reading a book	0%	0.20%	0%	0%	8.44%	2.82%	53.49%	3.44%	10.12%	5.21%	3.45%	2.61%
Playing a phone	0%	0%	22.53%	4.69%	1.15%	6.59%	1.98%	6.88%	5.11%	8.49%	0.94%	1.98%
Working on a laptop	0.66%	2.99%	10.20%	11.82%	8.15%	10.34%	9.62%	10.42%	88.89%	4.85%	1.19%	2.25%
Playing a tablet	0%	0%	0%	0%	0%	1.32%	0%	0%	0%	0.83%	0%	0%
Interacting with centre console	0%	3.34%	0%	13.83%	0%	19.92%	3.54%	25.62%	0%	23.46%	0%	12.29%

The cases of failed NDA identification can be caused by either the failure to recognise the engaged object (referred to ER) or the unmatched gaze map and recognised object mask (referred to EM). Table 3 presents the percentage of ER and EM over all instances, which aims to provide a deeper insight into how the object recognition and gaze trajectory estimation affect the identification results. For the NDA of reading a book, ER is larger than EM for most of the participants. Specifically, the low accuracy of participant 4 is mainly caused by the failure to recognise the object (53.49%). For the NDA of playing a phone, except participant 2, EM is larger than ER, which suggests that a small object can be recognised more easily but has a higher risk of unmatched gaze map and object mask. The main reason for the low accuracy for

the NDA of playing a laptop is ER, which is caused mainly by body occultation, especially for participant 5. It should be also noted that EM is relatively large in comparison with other NDAs with small objects, which is because even the partially blocked laptop can be recognised, the eye gaze could be on the blocked area. For the NDA of interacting with centre console, the recognition error is mainly caused by EM, the reason for which will be explained below.

Fig. 5 presents some snapshots of eye gaze mapping and object recognition, which suggests that in most of the cases the eye gaze well locates inside the recognised object. For participant

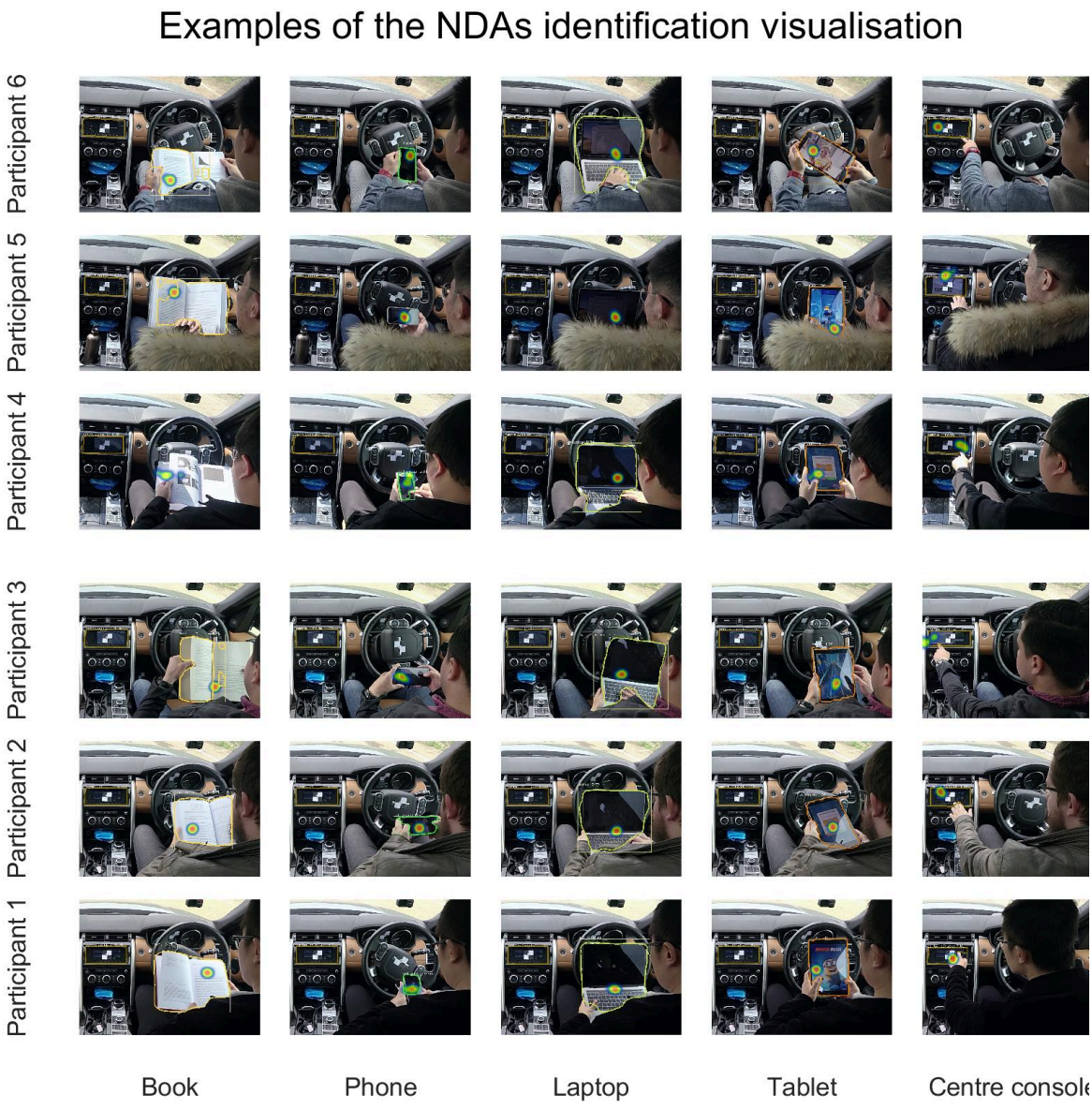

Fig. 5. NDAs identification visualisation examples. These images are cropped from raw images for appropriate visualisation.

5 who is working on a laptop, it can be seen that the estimated gaze is inside the laptop while the laptop is failed to be recognised, which could be solved by adjusting the position of the rear camera. For the NDA of reading a book, the book cannot be consistently recognised, for example with participant 4, which is suffering from the shadow and strong illumination caused by the sun. This interference reduces the accuracy of the book recognition and further affects the performance of NDA identification. This is the main reason that the identification result of reading a book for participant 4 (43.07%) is significantly lower than the results of others. Comparing with other NDAs, the behaviour during interacting with the centre console shows a different pattern. The eye gaze is not always well located in the centre of the centre console although the object is always well recognised, which leads to a relatively low NDA identification accuracy. There is a relatively large head rotation and body movement towards the left side which lead to a relatively large error of gaze mapping.

To show the performance of NDA tracking, the blue bar in Fig. 6 represents the successful identification of NDAs. It should be noted that only the middle 40 s for each NDA was analysed which justifies the large blank areas at the beginning and ending stages of each NDA. It can be observed that there are some discrete failures of identification due to the failed object recognition or inaccurate gaze estimation. An additional reason could be that the participants are looking away from the object, which is highly possible in real applications. To improve the accuracy of tracking, a large time window size in the classifier is suggested. However, it will sacrifice the performance of tracking the rapid change of NDAs.

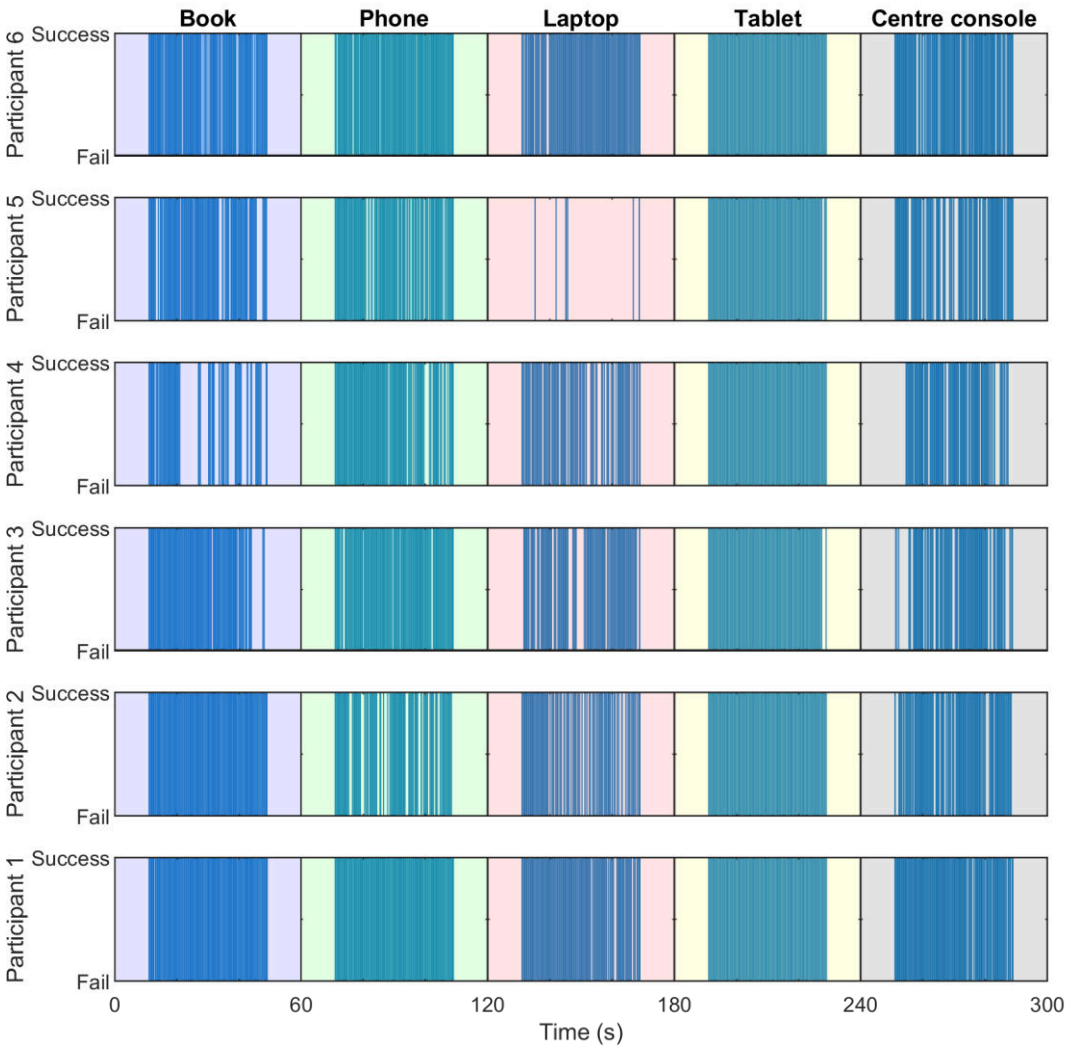

Fig. 6. NDAs identification and tracking for all participants. Five activities are distinguished by different colours of the background

3.3 Comparision with the state-of-the-art

The proposed approach has been compared with some state-of-the-art methods from the perspective of both action recognition and specific NDAs recognition, which are

(1) ResNet-50 [37]. It has a 50 layer 2D CNN architecture and achieves the action classification in the spatial domain. A pre-trained model on the ImageNet dataset is employed in this study.

(2) Two-stream CNN-based approach (2-stream) [18] for NDAs recognition. This method uses the information of both the RGB image stream and its associated current and historical optical flow frames to achieve the classification of the NDAs.

(3) 3D ResNets-18 (R3D18) [38]. It is based on ResNes-18 architecture that mainly utilises the 3D residual block in the whole network to encode the spatial-temporal information for action recognition.

(4) (2+1)D ResNets (R(2+1)D) [39]. It factorises the 3D residual block in R3D18 into a 2D spatial residual block and a 1D temporal residual block. Comparing to the 3D convolution with the same number of parameters, such a structure doubles the number of nonlinearities, which improves the model's capability on representing complex functions.

All methods were tested on the collected NDA data. As mentioned before, 40s video data for each activity and each participant was used for NDA recognition, which was split into 40 instances. All training and testing data were extracted from the rear camera by cropping a region which covers the human-object interaction. There are a total of 1200 instances in the dataset. k-fold cross-validation is employed to evaluate the models' performance based on the participants, where k is set as 3 in this study. For each k, data of 4 participants were used for training and the remaining 2 participants for testing.

Table 4 Comparison of the proposed method with 4 state-of-the-art methods on the NDA dataset

Method	**ResNet-50**[37]	**2-stream**[18]	**R3D18**[38]	**R2+1D**[39]	**Ours**
Accuracy	81.8%	86.3%	86.5%	85.3%	86.2%

The results are presented in Table 4. It can be observed that the proposed method achieves similar performance with other state-of-the-art methods, where ResNet-50 has relatively low accuracy. It should be noted that, firstly, most of the existing NDAs recognition methods, including the selected 4 methods, focus on the hand interaction between the driver and the object. Such methods lack the investigation of the driver's visual-attention, which could lead a misdetection of NDA engagement since the driver could check the road with their hand holding the object. The awareness of the driving environment is also important for the take-over strategy. Secondly, although the proposed method uses deep learning methods, it is fundamentally different from other compared methods. This method tends to be transparent and the type of activity is determined by considering the location of eye gaze and the type of object with the gaze. There is no further training process required to include a new type of activity. However, other compared methods will have to be trained again to include more activities. Thirdly, compared with the proposed method, the deep learning-based methods normally take a longer time to provide a prediction as information within a certain time window is used, which could slow down the system response.

4. Discussion

The performance of the proposed framework is largely affected by two factors: the success rate of object recognition and accuracy of the driver's gaze estimation. For object recognition, there are a few challenges:

1. To extend the universality on various models of a certain type of object, a large dataset such as COCO should be used. The dataset also should consider the diverse range of NDAs.

2. The main difference of object recognition between this study and other studies is that the driver is usually holding the object which inevitably leads to occlusion by hands or body if the camera position is not appropriate. The confidence level of recognition will be reduced. The confidence threshold, therefore, must be selected carefully. This problem is especially significant for small objects.
3. The location of the rear camera must consider two factors: avoiding the occlusion of the human body on the object and reducing the noise caused by illumination. This problem is especially significant for large objects.
4. Sunlight will cause the reflection of glass-surface objects such as phone, laptop and tablet. It could decrease the recognisability or confidence score of object recognition.
5. Other developed action recognition algorithms, which has the potential application for NDA recognition, usually focus on the driver's hand location or the interaction between hand and object/device. However, in the real driving scenario, the driver is engaging with NDA while observing surrounding situations. The proposed approach directly measures the driver's visual attention, which is crucial for further evaluation of the driver's awareness of the driving environment for a safe take-over transition.

The accuracy of the driver's gaze estimation is affected by the severe head rotation and body movement, as the facial features could be detected inaccurately and sometimes the face could be failed to be detected. Multiple cameras facing the driver could be one of the solutions, but it will increase the system complexity and cost. Furthermore, the variation of working distance between the driver and the front camera is the main factor to affect the performance. Nevertheless, it has been proven in [31] that the proposed system can accommodate a variety of working distance with a more complex nonlinear polynomial model as long as the diverse working distances are tested in the calibration process. In addition, the distortion of the rear

camera will affect the accuracy of eye gaze mapping. A high-order nonlinear model should be used to mitigate this influence, but the computational speed will be compromised.

5. Conclusions

This paper proposes a dual-camera based NDAs identification framework that benefits from computer vision, nonlinear system modelling and deep learning. It has been successfully demonstrated that this framework can identify the NDAs which require visual attention. The main strengths of this technique are:

1. NDAs required visual attention can be identified by inferring the object that the driver is looking at. The average success rate of this proposed framework is 86.18%. The performance is affected by both object recognition and gaze estimation, which could be further improved through creating the specific dataset for training and better locating the rear camera.
2. The proposed gazed-based ROI module embedded in this framework contributes about 30% improvement of average success rate and about 60% decrease of processing time. The size of this ROI can be customised according to the resolution of the rear camera.
3. The proposed active classifier improves the resilience to noise, such as the object can not be recognised suddenly due to occultation, by using a sliding time window.
4. The research of driver distraction in a human-driving vehicle can benefit from this study. The proposed system can be used to detect the driver's distraction behaviour by extending the types of objects to recognise, such as side mirror checking behaviour, dashboard checking behaviour, etc.

The main limitations are:

1. The proposed framework is not applicable to the NDAs without visual attention, such as listening to music.

2. As a camera-based approach, the performance of object recognition suffers from noise caused by harsh illumination, surface reflection and object occultation.
3. It should be noted that the proposed solution is only based on the object that the driver is engaging with. It cannot show whether the driver is watching a video or texting a message when a phone is used. To refine these NDAs, further studies are required.

For future work, with the increasing computational capability of portable devices like mobile, some existing light-weight models for object recognition can be used to a portable device-based real-time NDA recognition system. Furthermore, the tracking of NDAs could determine the duration of the engagement. The impact of different durations for various NDAs on driver's state and take-over performance needs to be evaluated, which is crucial for further design of take-over strategy to achieve the smooth and safe take-over transition.

References

[1] B. Wandtner, N. Schömig, G. Schmidt, Effects of Non-Driving Related Task Modalities on Takeover Performance in Highly Automated Driving, Hum. Factors J. Hum. Factors Ergon. Soc. 60 (2018) 870–881. https://doi.org/10.1177/0018720818768199.

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

[3] S.H. Yoon, Y.G. Ji, Non-driving-related tasks, workload, and takeover performance in highly automated driving contexts, Transp. Res. Part F Traffic Psychol. Behav. 60 (2019) 620–631. https://doi.org/10.1016/j.trf.2018.11.015.

[4] S.H. Yoon, Y.W. Kim, Y.G. Ji, The effects of takeover request modalities on highly automated car control transitions, Accid. Anal. Prev. 123 (2019) 150–158. https://doi.org/10.1016/j.aap.2018.11.018.

[5] I. JEGHAM, A. BEN KHALIFA, I. ALOUANI, M.A. MAHJOUB, Safe Driving :

Driver Action Recognition using SURF Keypoints, in: 2018 30th Int. Conf. Microelectron., IEEE, 2018: pp. 60–63. https://doi.org/10.1109/ICM.2018.8704009.

[6] B. Wandtner, N. Schömig, G. Schmidt, Secondary task engagement and disengagement in the context of highly automated driving, Transp. Res. Part F Traffic Psychol. Behav. 58 (2018) 253–263. https://doi.org/10.1016/j.trf.2018.06.001.

[7] F. Naujoks, S. Höfling, C. Purucker, K. Zeeb, From partial and high automation to manual driving: Relationship between non-driving related tasks, drowsiness and take-over performance, Accid. Anal. Prev. 121 (2018) 28–42. https://doi.org/10.1016/j.aap.2018.08.018.

[8] A. Eriksson, S.M. Petermeijer, M. Zimmermann, J.C.F. de Winter, K.J. Bengler, N.A. Stanton, Rolling Out the Red (and Green) Carpet: Supporting Driver Decision Making in Automation-to-Manual Transitions, IEEE Trans. Human-Machine Syst. 49 (2019) 20–31. https://doi.org/10.1109/THMS.2018.2883862.

[9] M. Sivak, B. Schoettle, Motion Sickness in Self-Driving Vehicles, (2015). https://deepblue.lib.umich.edu/handle/2027.42/111747.

[10] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, D.-S. Chen, A Comprehensive Survey of Vision-Based Human Action Recognition Methods, Sensors. 19 (2019) 1005. https://doi.org/10.3390/s19051005.

[11] M. Ziaeefard, R. Bergevin, Semantic human activity recognition: A literature review, Pattern Recognit. 48 (2015) 2329–2345. https://doi.org/10.1016/j.patcog.2015.03.006.

[12] Bangpeng Yao, Li Fei-Fei, Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 1691–1703. https://doi.org/10.1109/TPAMI.2012.67.

[13] T.H.N. Le, C. Zhu, Y. Zheng, K. Luu, M. Savvides, DeepSafeDrive: A grammar-aware

driver parsing approach to Driver Behavioral Situational Awareness (DB-SAW), Pattern Recognit. 66 (2017) 229–238. https://doi.org/10.1016/j.patcog.2016.11.028.

[14] A. Ullah, K. Muhammad, I.U. Haq, S.W. Baik, Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments, Futur. Gener. Comput. Syst. 96 (2019) 386–397. https://doi.org/10.1016/j.future.2019.01.029.

[15] J. Zhang, H. Hu, Domain learning joint with semantic adaptation for human action recognition, Pattern Recognit. 90 (2019) 196–209. https://doi.org/10.1016/j.patcog.2019.01.027.

[16] Y. Xing, C. Lv, H. Wang, D. Cao, E. Velenis, F.-Y. Wang, Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach, IEEE Trans. Veh. Technol. 68 (2019) 5379–5390. https://doi.org/10.1109/TVT.2019.2908425.

[17] H.M. Eraqi, Y. Abouelnaga, M.H. Saad, M.N. Moustafa, Driver Distraction Identification with an Ensemble of Convolutional Neural Networks, J. Adv. Transp. 2019 (2019). https://doi.org/10.1155/2019/4125865.

[18] L. Yang, T. Yang, H. Liu, X. Shan, J. Brighton, L. Skrypchuk, A. Mouzakitis, Y. Zhao, A refined non-driving activity classification using a two-stream convolutional neural network, IEEE Sens. J. XX (2020) 1–1. https://doi.org/10.1109/JSEN.2020.3005810.

[19] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: 2016 IEEE Conf. Comput. Vis. Pattern Recognit., IEEE, 2016: pp. 779–788. https://doi.org/10.1109/CVPR.2016.91.

[20] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031.

[21] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask R-CNN, in: 2017 IEEE Int. Conf. Comput. Vis., IEEE, 2017: pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322.

[22] T. D'Orazio, M. Leo, C. Guaragnella, A. Distante, A visual approach for driver inattention detection, Pattern Recognit. 40 (2007) 2341–2355. https://doi.org/10.1016/j.patcog.2007.01.018.

[23] T. D'Orazio, M. Leo, A. Distante, Eye detection in face images for a driver vigilance system, in: IEEE Intell. Veh. Symp. 2004, IEEE, 2004: pp. 95–98. https://doi.org/10.1109/IVS.2004.1336362.

[24] F. Vicente, Z. Huang, X. Xiong, F. De la Torre, W. Zhang, D. Levi, Driver Gaze Tracking and Eyes Off the Road Detection System, IEEE Trans. Intell. Transp. Syst. 16 (2015) 2014–2027. https://doi.org/10.1109/TITS.2015.2396031.

[25] T. Baltrusaitis, P. Robinson, L.-P. Morency, OpenFace: An open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conf. Appl. Comput. Vis., IEEE, 2016: pp. 1–10. https://doi.org/10.1109/WACV.2016.7477553.

[26] A. Kar, P. Corcoran, A Review and Analysis of Eye-Gaze Estimation Systems, Algorithms and Performance Evaluation Methods in Consumer Platforms, IEEE Access. 5 (2017) 16495–16519. https://doi.org/10.1109/ACCESS.2017.2735633.

[27] S. Vora, A. Rangesh, M.M. Trivedi, Driver Gaze Zone Estimation Using Convolutional Neural Networks: A General Framework and Ablative Analysis, IEEE Trans. Intell. Veh. 3 (2018) 254–265. https://doi.org/10.1109/TIV.2018.2843120.

[28] L. Fridman, P. Langhans, J. Lee, B. Reimer, Driver Gaze Region Estimation without Use of Eye Movement, IEEE Intell. Syst. 31 (2016) 49–56. https://doi.org/10.1109/MIS.2016.47.

[29] D. Xiao, C. Feng, Detection of drivers visual attention using smartphone, in: 2016 12th

Int. Conf. Nat. Comput. Fuzzy Syst. Knowl. Discov., IEEE, 2016: pp. 630–635. https://doi.org/10.1109/FSKD.2016.7603247.

[30] T. Baltrusaitis, A. Zadeh, Y.C. Lim, L.-P. Morency, OpenFace 2.0: Facial Behavior Analysis Toolkit, in: 2018 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2018), IEEE, 2018: pp. 59–66. https://doi.org/10.1109/FG.2018.00019.

[31] L. Yang, K. Dong, A.J. Dmitruk, J. Brighton, Y. Zhao, A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring, IEEE Trans. Intell. Transp. Syst. 21 (2020) 4318–4327. https://doi.org/10.1109/TITS.2019.2939676.

[32] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single Shot MultiBox Detector, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2016: pp. 21–37. https://doi.org/10.1007/978-3-319-46448-0_2.

[33] Y. Li, H. Qi, J. Dai, X. Ji, Y. Wei, Fully Convolutional Instance-Aware Semantic Segmentation, in: 2017 IEEE Conf. Comput. Vis. Pattern Recognit., IEEE, 2017: pp. 4438–4446. https://doi.org/10.1109/CVPR.2017.472.

[34] Z.-Q. Zhao, P. Zheng, S.-T. Xu, X. Wu, Object Detection With Deep Learning: A Review, IEEE Trans. Neural Networks Learn. Syst. (2019) 1–21. https://doi.org/10.1109/TNNLS.2018.2876865.

[35] Faster R-CNN and Mask R-CNN in PyTorch 1.0, (n.d.). https://github.com/facebookresearch/maskrcnn-benchmark (accessed April 20, 2019).

[36] A.T. Duchowski, A breadth-first survey of eye-tracking applications, Behav. Res. Methods, Instruments, Comput. 34 (2002) 455–470. https://doi.org/10.3758/BF03195475.

[37] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conf. Comput. Vis. Pattern Recognit., IEEE, 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[38] K. Hara, H. Kataoka, Y. Satoh, Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition, in: 2017 IEEE Int. Conf. Comput. Vis. Work., IEEE, 2017: pp. 3154–3160. https://doi.org/10.1109/ICCVW.2017.373.

[39] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, M. Paluri, A Closer Look at Spatiotemporal Convolutions for Action Recognition, in: 2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit., IEEE, 2018: pp. 6450–6459. https://doi.org/10.1109/CVPR.2018.00675.

2021-03-25

Recognition of visual-related non-driving activities using a dual-camera monitoring system

Yang, Lichao

Elsevier

Yang L, Dong K, Ding Y, et al., (2021) Recognition of visual-related non-driving activities using a dual-camera monitoring system. Pattern Recognition, Volume 116, August 2021, Article number 107955 https://doi.org/10.1016/j.patcog.2021.107955