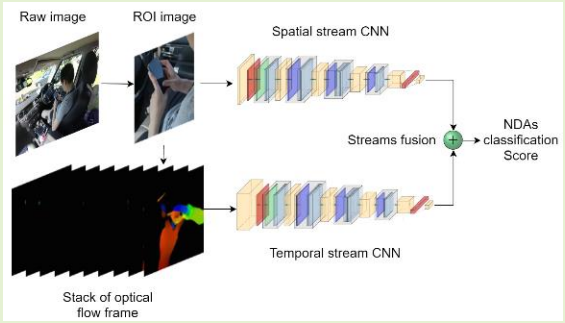
IEEE Sensors Journal, Volume 21, Number 14, pp. 15574-15583, July 2021
DOI: 10.1109/JSEN.2020.3005810

# A refined non-driving activity classification using a two-stream convolutional neural network

Lichao Yang, Tingyu Yang, Haochen Liu, Xiaocai Shan, James Brighton, Lee Skrypchuk, Alexandros Mouzakitis and Yifan Zhao*, *Senior Member, IEEE*

***Abstract*— It is of great importance to monitor the driver's status to achieve an intelligent and safe take-over transition in the level 3 automated driving vehicle. We present a camera-based system to recognise the non-driving activities (NDAs) which may lead to different cognitive capabilities for take-over based on a fusion of spatial and temporal information. The region of interest (ROI) is automatically selected based on the extracted masks of the driver and the object/device interacting with. Then, the RGB image of the ROI (the spatial stream) and its associated current and historical optical flow frames (the temporal stream) are fed into a two-stream convolutional neural network (CNN) for the classification of NDAs. Such an approach is able to identify not only the object/device but also the interaction mode between the object and the driver, which enables a refined NDA classification. In this paper, we evaluated the performance of classifying 10 NDAs with two types of devices (tablet and phone) and 5 types of tasks (emailing, reading, watching videos, web-browsing and gaming) for 10 participants. Results show that the proposed system improves the averaged classification accuracy from 61.0% when using a single spatial stream to 90.5%.**




***Index Terms*— NDA classification, Level 3 automation, optical flow, 2-stream CNN.**

## I. Introduction

FREELY engaging in non-driving activities (NDAs) may be allowed in the future when the driver is driving a level 3 automated driving vehicle [1]. According to the definition of the SAE (J3016) Automation Levels [2], the driver should respond appropriately to the request to intervene. However, the engagement of NDAs could reduce the driver's perceptual and cognitive capability on driving and situation awareness, which could result in a negative impact on the take-over response [3]. Therefore, it is necessary to investigate the implication of NDA engagement on the driver's status and attention level to ensure the driver is in an appropriate condition to take over the vehicle. From the perspective of driving safety, Kim *et al.* [4] suggested when the take-over request is given by the vehicle, the driving performance after the take-over could be affected by the driver's age, gender and experience, but the status before the take-over might be more relevant. Although some approaches [4], [5] have been proposed in recent years to directly evaluate the driver's mental workload, the evaluated accuracy is not satisfactory due to the lack of convincible ground truth. The evaluation of the workload could be subjective and it is hard to be quantified. The further research results show that different types of NDA and driving scenarios could cause different cognitive loads of the driver which affect the performance of the take-over quality and take-over time [6], [7]. For instance, visual related activities tended to take longer reaction time than the auditory related activities [8]. To achieve high-quality take-over and safety enhancement [9], it is therefore crucial to precisely identify, distinguish and track the type of NDA that the driver is engaging with, then to evaluate the status and attention level or workload for the improvement of vehicle safety and operational efficiency. However, there is very limited literature focusing on that.

Analogous to NDAs, secondary tasks as non-driving related tasks have been widely researched in human-driving in recent years. Li and Busso [10] claimed that secondary tasks can be recognised by evaluating the driver's mirror-checking action. However, when the driver is doing NDAs in an automated driving vehicle, the frequency of the mirror-checking will significantly decline. Therefore, this action is not considered as an appropriate indicator for NDA recognition. Jin *et al.* [11] proposed to recognise 6 secondary tasks (Bluetooth calls, cell phone calls, sending text messages, operating car-mounted players, chatting and singing) by combining both extracted eye movement and vehicle state characteristics. Martin *et al.* [12] presented a 3-stream recurrent neural network (RNN) system based on the driver's upper body pose. This system evaluates the transient skeleton movement, the spatial relationship of

Manuscript received …

L. Yang, Y.Yang, H. Liu, J. Brighton and Y. Zhao are with School of Aerospace, Transport and Manufacturing, Cranfield University, Bedfordshire MK43 0AL, UK.

X. Shan is with Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China.

L. Skrypchuk and A. Mouzakitis are with Research & Technology, Jaguar Land Rover, UK.

*the corresponding author: Y. Zhao (yifan.zhao@cranfield.ac.uk).

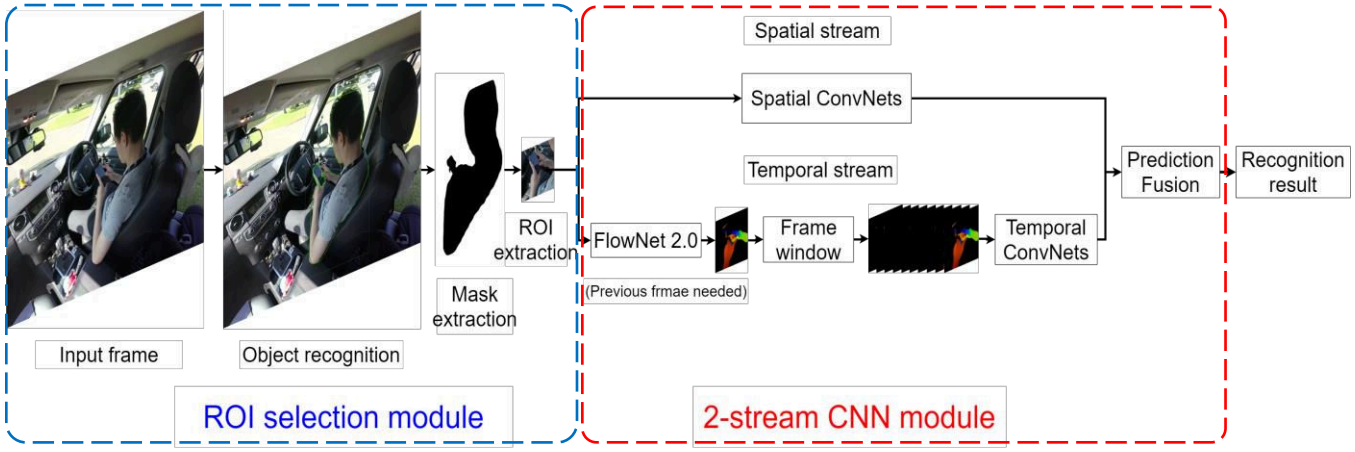



Fig. 1. The proposed framework for NDAs recognition consisting two parts: ROI selection module and 2-stream CNN module.

body parts and the knowledge about the vehicle interior to recognise 6 secondary tasks (drinking from a bottle, eating, using a phone for texting, making a call and reading a book). Xing *et al.* [13] collected both the colour and depth images of the driver's behaviour inside the vehicle cabin. Besides, the Kinect recorded the 3-D head rotation angles and the upper body joint position. A feedforward neural network (FFNN) was established to analyse the collected data and identify the secondary tasks. All these studies can recognise some kinds of secondary tasks like using a phone, operating car-mounted player and chatting while driving manually. They presume that the primary task is driving which limits the diversity and continuity of the secondary tasks. These methods, therefore, cannot be directly applied for recognising NDAs with high complexity and uncertainty.

As shown in Table I, Sivak and Schoettle [14] suggested that the common NDAs are reading, texting, working, watching movies and playing games. Yang *et al.* [1] proposed a dual-cameras based drive gaze mapping system which could be used to recognise some NDAs with visual attention by mapping the gaze on the object that the driver is engaging with. However, such an object-based recognition approach can only identify that the driver is interacting with a phone but cannot recognise whether the driver is watching a movie (passive interaction) or playing a game (active interaction). The level of the driver's engagement in these activities in terms of perception and cognition is different according to the interaction mode, which leads to different performance after the take-over. The activities like reading or watching videos are considered as passive-interaction activities since the driver intakes the information passively. But some like texting and playing games request a strong active interaction between the device and the driver. Consequently, the interaction mode could result in a different workload of the driver [4], [8]. A further refinement of NDAs classification in terms of object/device and task is therefore highly essential to design a more intelligent and efficient take-over process. This paper proposes a novel region of interest (ROI) based 2-stream (visual scene and optical flow) convolutional neural network system to achieve this target through identifying both the device that the driver is engaging with and the task (e.g. reading, playing a game, watch a movie, emailing etc.) simultaneously.

TABLE I
THE NDAS THAT DRIVERS WANT TO DO IN AUTOMATED DRIVER VEHICLE [14]

| NDAs | U.S. | China | India | Japan | U.K. | Australia |
|---|---|---|---|---|---|---|
| Read | 14% | 10.8% | 11.1% | 8.4% | 9.9% | 8.3% |
| Text or talk | 12.7% | 21.5% | 16.3% | 11.0% | 7.1% | 10.1% |
| Sleep | 8.8% | 11.2% | 5.1% | 18.9% | 9.4% | 9.0% |
| Watch movies | 7.8% | 1.7% | 13.4% | 9.2% | 5.4% | 7.3% |
| Work | 6.2% | 5.6% | 17.7% | 1.0% | 6.4% | 6.5% |
| Play games | 2.6% | 1.4% | 2.3% | 1.8% | 2.5% | 2.5% |
| Other | 1.8% | 0.7% | 0.8% | 0.3% | 2.2% | 1.3% |

## II. METHODOLOGY

### A. System Architecture

In the early stage of human action recognition, the human-object interaction has been widely researched, through the integration of object recognition, pose estimation and action identification [15], [16]. For the NDAs recognition, the movement restriction and the body occultation enhance the difficulty of human pose estimation since the driver is sitting on the seat. Object detection methods can also be used to recognise some actions inside a vehicle such as hands-on-steering-wheel or using a phone [17]. Such methods recognise the human body parts and the object by sematic instance segmentation. With the development of multi-object detection, several CNN-based approaches have been proposed for action recognition in video. The achievements have been made from the perspective of CNN framework or network design [18]–[20]. The evaluation of such existing researches is based on the representative video datasets, such as HMDB-51 [21], UCF-101 [22], Kinetics [23]. These researches focus on the classification of actions with distinctive features like cutting in kitchen, swing, archery etc. [22]. However, in this paper, we focus on the classification of those phone-using and tablet-using NDAs with high similarities. Such NDAs happen inside of a vehicle and the driver is constrained on the seat. The spatial moving scale and intensity of activities are quite lower and harder to distinguish than the distinctive ones abovementioned. In this paper, we

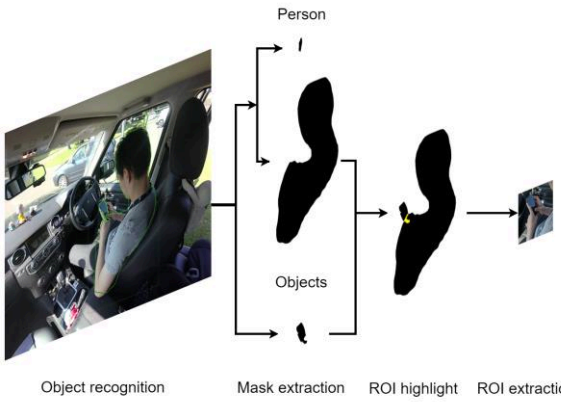



Fig. 2. The flowchart of the ROI selection module.

propose that the classification process can be divided into 3 steps. In the first step, by extracting the ROI of the raw image captured by the camera, the interaction between the driver and the object can be limited to a region, which is helpful to reduce the noise and the processing time. The second step is to classify the object or device the driver is operating on. It relies on the analysis of the object's spatial information. The last step is to indicate how the driver interacts with the object based on pattern recognition. It is achieved by motion estimation. The last 2 steps can be run in parallel. The final result is given by fusing the 2 steps.

The flowchart of the system is illustrated in Fig. 1, where the proposed system contains two modules: the ROI selection module and the 2-stream CNN module. The input frames are collected by a camera which is mounted on the roof of the vehicle to ensure that the object and hands are captured. The ROI module provides a region of human-object interacting (highlighted in Fig.1), which aims to significantly reduce the processing time and background noise for the 2-stream CNN module, and furtherly improve the classification accuracy. Then the detected ROI is fed into the 2-stream CNN module. The input of the spatial stream is from the RGB images and the input of the temporal stream is from a stack of optical flow frames which represent the motion between two adjacent frames within a certain time window. Then the prediction scores of the spatial and temporal streams will be fused to promote the final NDA classification result.

## B. ROI Selection

The raw RGB frames captured by the camera carry abundant information from both inside and outside of the vehicle. When we attempt to characterise and identify NDAs, the most important parts are the object operated by the driver and the pattern of the driver's behaviour, especially the figures and hands. This module aims to extract a region covering these parts from the raw frame due to two reasons. The first benefit is to help achieve real-time or near real-time performance. The size of the images fed into CNN should be small and informative. To keep the details of useful information, cropping the useless background is better than downsizing. The second benefit is to eliminate background noise. The scene change on the window during driving could introduce interference to pattern recognition. To achieve these aims, the raw frame is initially analysed by an object recognition algorithm, Mask R-CNN. It

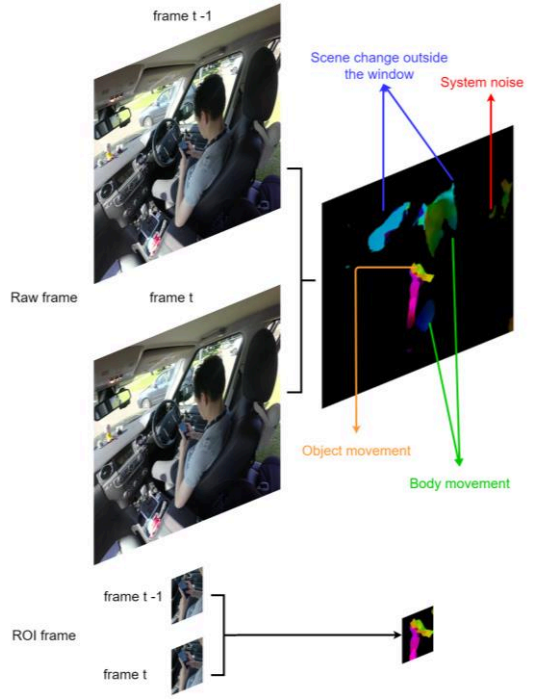



Fig. 3. The comparison of the optical flow frame performance between raw frames and ROT frames.

is a state-of-the-art object instance segmentation algorithm which could classify objects and localise them in pixels [24]. Comparing with the methods which can only provide a bounding box to localise the object, this algorithm offers a more accurate boundary as a mask on the recognised object, which is crucial to determine whether the driver is engaging with the object. The details are presented in Fig. 2. In this module, Mask R-CNN is applied to recognise the driver and potential objects which could be involved in NDAs, along with the masks. Then the ROI is selected based on the centre of the overlapping or connected area between human and object. The cropped frame will then be used as an input of the 2-stream CNN module. If there is no ROI detected, the following module will not be activated, which suggests there is no related object or person in the scene, or the person and the object are recognised but the person is not interacting with it. For the estimation of optical flow, it is assumed that the location of the ROI within the time window does not change over time. If the object or driver is not detected or the ROI location difference between the last frame and current frame is smaller than a pre-set threshold, the current ROI will be the same with the ROI in the last frame. The ROI will only be updated if the location change exceeds the threshold. The threshold was set as 40 pixels in this study. The size of the ROI is customisable. In this case, the size was set as 320 × 320 pixels, where the raw image size is 1920 × 1440 pixels.

## C. Optical Flow Estimation

Optical flow information has wide applications on studying vision-related tasks such as human pose estimation [25], video classification [26] and action recognition [27]. The rich motion information can be used to characterise the driver's behaviour between two adjacent frames. Compared to other optical flow estimation tools like DeepFlow [28] and Flow Fields [29], FlowNet 2.0 achieves the finest estimation performance. It

provides the end-to-end optical flow estimation with convolutional networks [30]. The motion vector of each pixel is visualised by the colour coding. The detail can be found in [31].

The processed optical flow frames for both raw and ROI frames are presented in Fig. 3. The optical flow frame extracted from two adjacent raw frames includes the pixel motion from various moving sources, e.g., human, device, outside scene. We assume that the driver's behaviour associated with the device trajectory is the most important factor, particularly, the hand movement, to determine the task as detailed as possible. The obtained information from the optical flow frame can be categorised into 4 parts: scene change outside the window, body movement, device movement, and system noise, as marked in Fig. 3. From the optical flow frame, a moving vehicle and a pedestrian outside the window can be observed and regarded as outside noise. There is also some system noise on the right side of the frame. All this information has no strong relevance to the pattern of the driver's behaviour. It can be considered as noises which could result in a negative effect on the performance of the temporal stream. It should be noted that although the driver's head and arm movement could be related to NDAs it is relatively subjective and ignored in this paper. In contrast, the optical flow of the ROI frames provides clear features related to the driver's hand and object movement. It is therefore used as one of the inputs for the 2-stream CNN module.

### D. 2-stream CNN

The challenge of action recognition in a still RGB image is that it cannot provide the spatiotemporal features [19]. Particularly for the NDA recognition, the common methods like pose estimation and scene recognition are not applicable. The driver is constrained on the seat and the only moving parts of the driver are the hands or head. The features extracted from the still image are not enough to differentiate most of NDAs. In recent years, several CNN-based action recognition architectures have been proposed to improve the ability to capture the spatiotemporal features and increase the accuracy of the action recognition in videos, such as CNN with long short-term memory (LSTM) [32], 3D-CNN [18], 2-stream CNN [20] and 2-stream 3D-CNN [23]. The temporal stream of the 2-stream architecture offers the features of movement in the time domain and helps to identify the driver's behaviour. However, the state-of-the-art algorithm provided by the 3D-CNN model in the 2-stream architecture requests large-scale datasets due to the complexity of the network [33]. Unlike the representative datasets mentioned above, the dataset used in this study is relatively small. One of the differences in data is that the features of the driver's behaviour are constrained in a small region. A complex network could increase the training burden and easily lead to an overfitting problem. Hence, a 2-stream architecture with 2D-CNN model is proposed in this paper. To achieve a better recognition performance, the CNN model in the 2-stream architecture is built based on the Residual Network (ResNet) due to its strong capability of training deeper network [34].

The architecture of the CNN module is presented in Fig. 4. The input of the spatial stream is a single ROI RGB frame at the current time and the input of the temporal stream is a stack of 10 optical flow frames (equals to 0.42s with a sample rate of 24 fps) on ROI calculated from 11 adjacent frames including the current frame. Traditionally, the input of the temporal stream is a stack of two-channel frames (two vectors). For an arbitrary pixel $(u, v)$ in a single frame at the time $t$, the motion vector of this pixel is denoted as $(\overrightarrow{p_t^x(u,v)}, \overrightarrow{p_t^y(u,v)})$. The input for the temporal stream is denoted as $S_t(u, v, c)$, where $c$ indicates the channel index. The corresponding input stack can be expressed as follow:

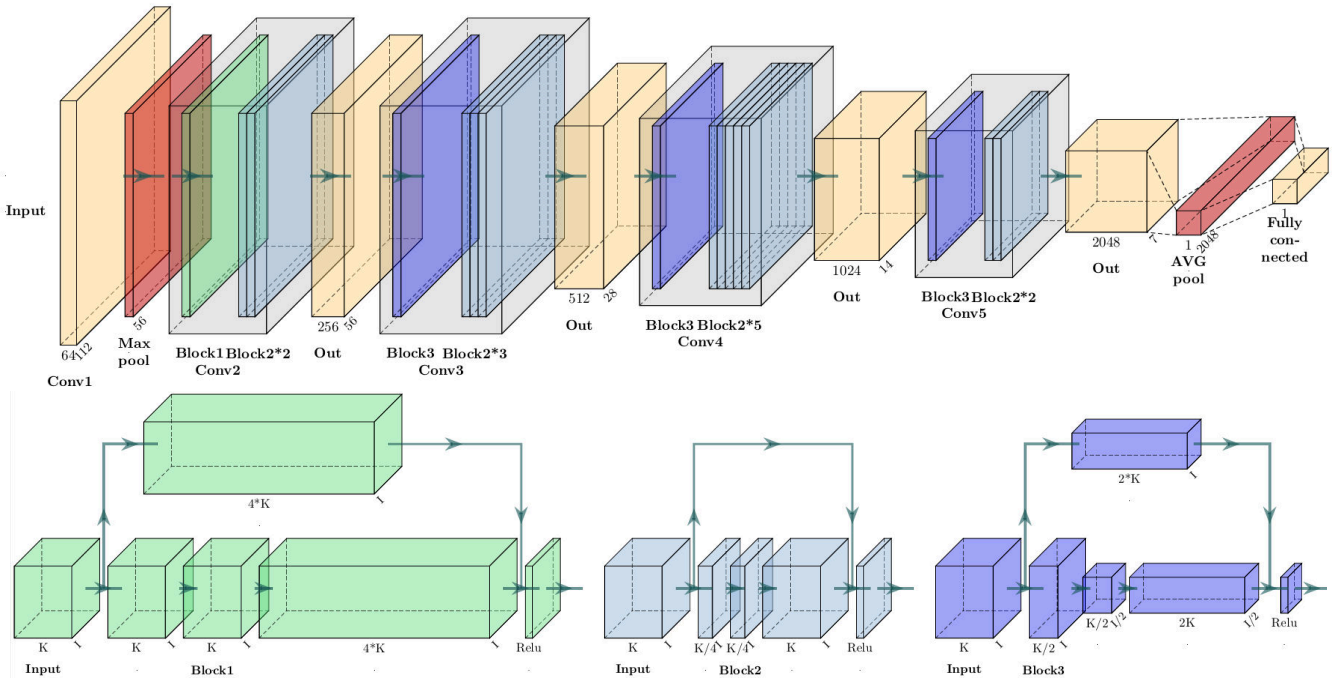



Fig. 4. The architecture of ResNet 50 CNN. There are three types of convolutional blocks in this network, which are detailed in the bottom graph and indicated as different colours.

$$\begin{cases} S_t(u,v,2k-1) = \overrightarrow{p^x_{t-k+1}}(u,v) \\ S_t(u,v,2k) = \overrightarrow{p^y_{t-k+1}}(u,v) \end{cases}, \quad (1)$$

where $u = [1,w], v = [1,h], k = [1,N]$, $w$ and $h$ are the width and height of the frame respectively, $N$ denotes the number of the frame inside the stack.

In this paper, we visualise the optical flow with colour coding. The vector field is then converted from two channels into three RGB channels. The input stack for the current frame $t$ can then be expressed as follow:

$$\begin{cases} S_t(u,v,3k-2) = \overrightarrow{p^R_{t-k+1}}(u,v) \\ S_t(u,v,3k-1) = \overrightarrow{p^G_{t-k+1}}(u,v) \\ S_t(u,v,3k) = \overrightarrow{p^B_{t-k+1}}(u,v) \end{cases}. \quad (2)$$

The number of optical flow frames in the stack, $N$, is configurable. It depends on how much historical information is required. Its performance will be addressed and discussed below.

ResNet-50 models are then built for both streams independently. There are 5 groups of convolution layer shown in Fig.4. In the convolutional layer 1, both models extract 64 feature maps from the input. The difference of these 2 streams is the input, which is a 3-channel RGB image for the spatial stream or 30-channel optical flow stack for the temporal stream. The last 4 convolution layer groups are made up of 3 types of residual block, which are shown in the bottom of the Fig.4. The design of the shortcut structure in the block can be expressed as:

$$x_{l+1} = \mathrm{F}(x_l, \{W_i\}) + x_l, \quad (3)$$

where $x_l$ is the input of the layer $l$. $\mathrm{F}(x_l, \{W_i\})$ represents the function where the residual mapping is learned. Such residual structure alleviates the problem of exploding and vanishing gradient and usually achieves good performance in a deeper network [34].

The training process started with a pre-trained ResNet-50 model. The loss function used in training can be described as:

$$Loss(x, label) = -x[label] + \log\left(\sum_j e^{(x[j])}\right) \quad (4)$$

where $x$ is the output which has been one hot encoded. *Label* is the true class. $j$ is the index of the classes. The stochastic gradient descent (SGD) algorithm is used as optimizer [35], which can be expressed as:

$$w_{n+1} = w_n - \gamma \nabla_w \mathrm{L}(z_n, w_n), \quad (5)$$

where $n$ is the number of iteration. The gradient descent method focuses on the randomly picked mini-batch $z_n$. The loss L is minimised bases on the gradient of the weight vector $w$ and the chosen gain $\gamma$. Furthermore, the learning rate is controlled in the training process. It starts with a high learning rate to accelerate the process and then reduces when the loss of the validation dataset stops improving.

After the training process, the trained model assesses the prediction scores of both streams. Finally, both scores are fused through a model expressed as:

$$S_i = \frac{R_i}{\sum_{i=0}^{n-1}|R_i|} + \frac{O_i}{\sum_{i=0}^{n-1}|O_i|}, \quad (6)$$

where $S$ is the fusion score, $R$ is the prediction score from the spatial CNN module, $O$ is the prediction score from the temporal stream, $i$ is the class index, and $n$ is the number of NDAs class.

TABLE II
CATEGORIES FOR NDAS RECOGNITION

| Term | Browsing websites | Sending emails | Playing games | Reading | Watching videos |
|---|---|---|---|---|---|
| Phone | PB | PE | PG | PR | PV |
| Tablet | TB | TE | TG | TR | TV |

### E. Experiment Setup and Performance Validation

A Land Rover Discovery 4 was used as the test vehicle. The employed camera was Garmin Virb Action Camera which was mounted on the roof of the vehicle between two front seats. The resolution of the camera was set as 1920 × 1440 pixels and images were sampled at 24 frames per second (fps). A PC with an Intel i7 9700k CPU, 32GB memory and an NVIDIA GeForce RTX 2080 GPU was employed for all deep learning related work.

During the experiment, the vehicle stayed stationary. A total of 10 participants (6 male and 4 female) were recruited for this experiment. The participants' age is in a range from 22 to 26. They were requested to sit on the driving seat with the fastened seat belt and engage with the same phone and tablet to conduct the selected activities one by one. Each activity lasted 1 minute. A total of 10 types of NDA were evaluated in this experiment, as presented in TABLE II. The class of each activity is presented by 2 capital letters for the convenience of result presentation. The first letter refers to the object (P and T stand for phone and tablet respectively), and the second letter refers to the task. For instance, PE refers to sending emails using a phone. Auditory guidance using Google Cloud Text-to-Speech was provided in this experiment to ensure consistency across all participants.

In this experiment, the participants need some time to follow the auditory guide for the NDAs transition. Therefore, only the middle 40 seconds video was used for training, validation and testing. Each video has been split into 20 segments with a length of 2 seconds for each segment. There are 2000 segments in total for all participants and all NDAs. From these segments, 64% of them was randomly selected for the training process, 16% of them was used validation process and 20% of them was used for the testing process. In the train process, 1 instance was randomly picked from each segment for both streams. The validation process was activated after each training epoch to adjust some hyperparameters like learning rate. 3 instances were randomly picked from each segment for both streams in this process. The testing process happened after the training process to evaluate the performance of the system. The following analysis is based on the results of the testing process.

## III. RESULTS

### A. Two Streams

An example of input frames for the 2-stream CNN module for each NDA is presented in Fig. 5, where the first column is the raw image with a full resolution, the second column is RGB images of the selected ROI as the spatial stream, and the remaining columns are the optical flow frames as the temporal stream. From the RGB images of ROI, the difference can be observed between the phone-related activities and the tablet-

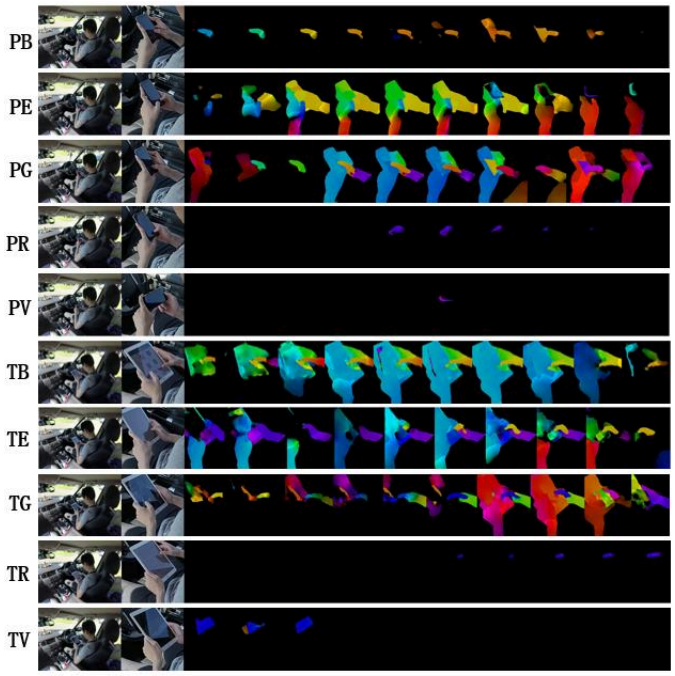



Fig. 5. Examples of raw frame and input frames of 2-stream CNN module. There is some overlap between optical flow frames to fit the figure size.

related activities. The difference includes (a) the size of the object, (b) the distance between the object and the driver's body, and (c) the hand gesture. Therefore, the spatial stream should be able to differentiate the first 5 NDAs and the last 5 NDAs. However, this difference between the first 4 phone-related activities is dramatically dropped. It can be predicted that the classification accuracy for these 4 activities will be relatively low if only the spatial stream is applied. Furthermore, it can be seen that there is some reflection on the screen of the phone and the tablet. The change of illumination could affect the spatial information of the object while the driver is doing the same NDA, which could furtherly bring negative impact on the classification performance.

The optical flow frames contain more information on the driver's moving behaviour. It can be seen that activities like PB, PR, and TR involve one hand most of the time. Meanwhile, some activities like PE, PG, TB, TE, and TG need two hands for interaction. Another dimension of the difference between the two-hand related activities is the hands and fingers movement. For example, the different colour pattern between PE and PG suggests a different interaction mode with the device. The driver's behaviour on these NDAs can be differentiated by the movement vectors of the hands and fingers which are represented by colours and its accumulation in the time domain. It also should be noticed that the optical flow stream is sensitive to the relatively high-frequency interaction for NDAs like playing games, sending emails. For some other NDAs like watching videos or reading, particularly with the tablet, the driver may stay with the same pose for a long time without any movement, as shown in TR and TV.

## B. Classification Performance

The classification performance of the spatial stream only is

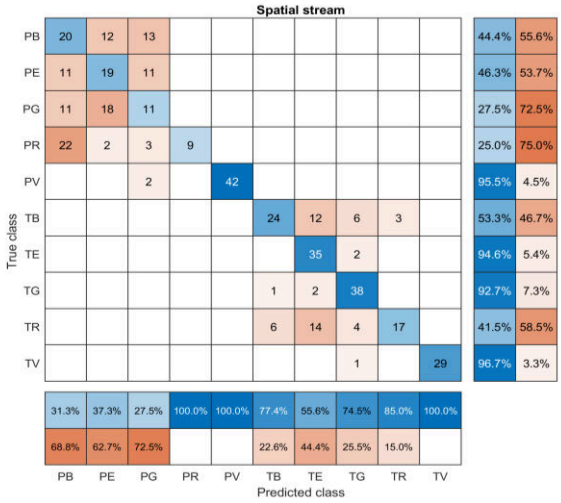



Fig. 6. Confusion matrix of NDAs recognition for the spatial stream. The precision and recall for each class are presented in the bottom and right of the figure, respectively, where the blue colour indicates the true value and the orange colour indicates the false value.

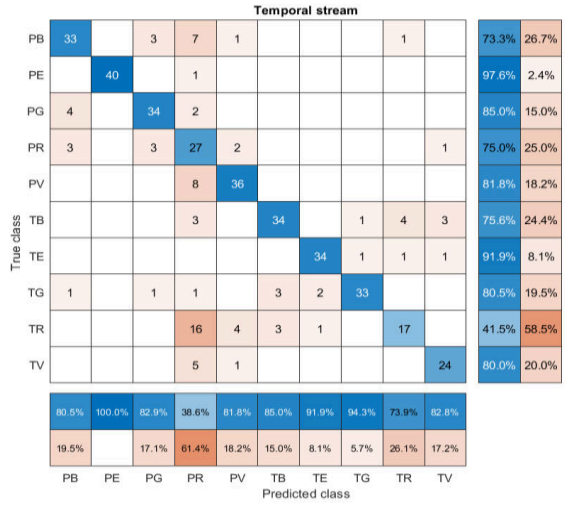



Fig. 7. Confusion matrix of NDAs recognition for the temporal stream.

presented in Fig. 6. It can be found that the phone-related activities can be easily distinguished with the tablet-related activities, evidenced by zero error. However, for the classification among the phone-related activities or the tablet-related activities, the performance is not satisfactory. For PB, PE, PG and PR, the recall is lower than 50%, more than half of the true instance has not been recognised. TB and TR are difficult to be differentiated as well. This indicates that the spatial stream is not able to offer a persuasive NDA classification for the same object. Besides, it can be observed that the value of both recall and precision for watching videos by phone (PV) and tablet (TV) are high, which suggests a reliable NDA classification. The reason is that the way how participants interact with objects is quite special. When participants are conducting some activities like browsing website or sending emails, they usually hold the phone or tablet vertically. However, for watching videos, most participants hold the phone or tablet horizontally. Comparing with the phone-related NDAs, the tablet-related NDAs classification shows a better performance in the spatial stream for both recall and precision. The content on the tablet's screen may have a

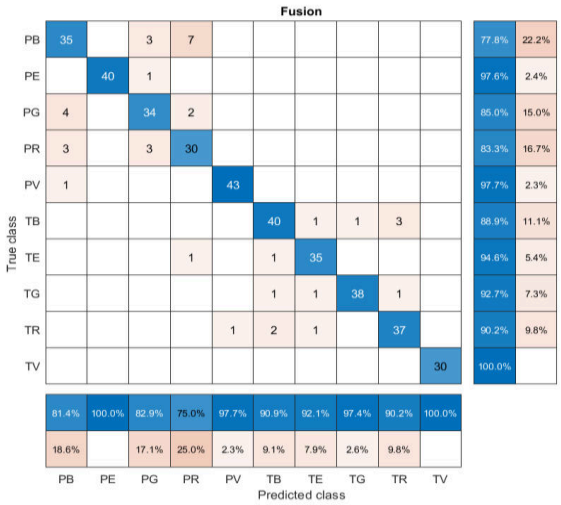



Fig. 8. Confusion matrix of NDAs recognition for the fusion of 2 streams.

TABLE III
OVERALL ACCURACY OF NDAS RECOGNITION

| Term | Spatial stream | Temporal stream | Fusion |
|---|---|---|---|
| P accuracy | 49.0% | 82.5% | 88.3% |
| T accuracy | 73.7% | 73.2% | 92.8% |
| Accuracy | 61.0% | 78.0% | 90.5% |
| Weighted F1 | 60.6% | 78.7% | 90.6% |
| Top-3 error | 10.5% | 4.3% | 0.5% |

TABLE IV
OVERALL ACCURACY OF NDA RECOGNITION WITHOUT ROI SELECTION

| Term | Spatial stream | Temporal stream | Fusion |
|---|---|---|---|
| Accuracy | 19.0% | 66.2% | 72.5% |
| Weighted F1 | 15.1% | 66.4% | 71.5% |
| Top-3 error | 32.5% | 10.2% | 5.5% |

contribution to the classification while that is not available for the phone, as shown in Fig. 5.

Fig. 7 presents the confusion matrix of the classification using the temporal stream only. The recall of most NDAs is around 75%, except TR. Almost half of the true instance has been predicted as PR, which is because both NDAs are lack of movement. The precision of most NDAs is above 80%, while the precision of PR is only 38.6% Both recall and precision of sending emails are the highest (above 90%) no matter using a phone (PE) and tablet (TE). This is contributed by the special interaction mode in comparison with others.

The fusion result of the proposed 2-stream approach is shown in Fig. 8, which demonstrates a significant improvement for all NDAs in contrast to the results of any single stream. The classification error among the NDAs with the same object has been dramatically reduced. The overall accuracy is presented in TABLE III. The overall accuracy has been improved from 61.0% (the spatial stream only) to 90.5%. Specifically, for the phone-related activities, the accuracy has been improved from 49.0% to 88.3%. For the tablet-related activities, the accuracy has been improved from 73.7% to 92.8%. In terms of the performance of a single stream, the temporal stream performs much better for the phone-related activities. While for the tablet-related activities, the performance is similar. The weighted F1 scores

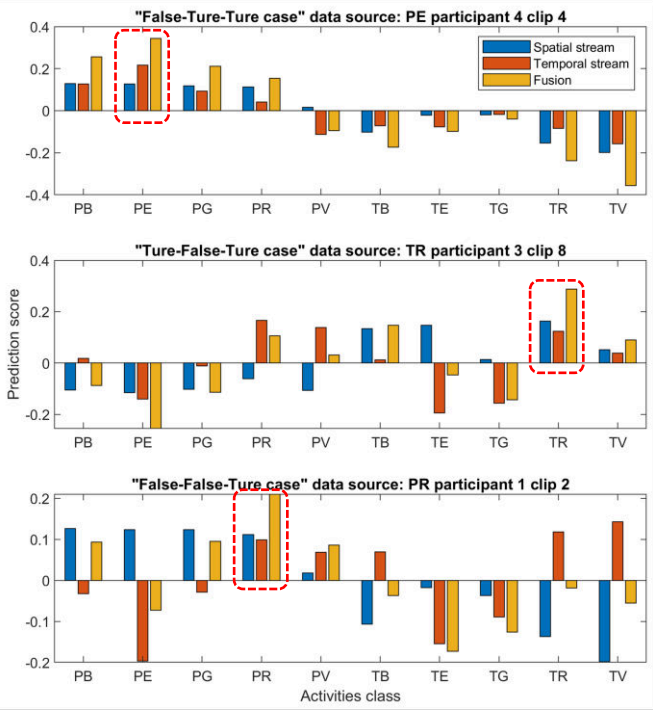



Fig. 9. Prediction results for inference cases. The true class is highlighted by a red block.

for all 3 terms are similar to the accuracy results. The top-3 error of the proposed method is only 0.5%. Specifically, for the spatial stream, the top-3 error is 10.5% while the weighted F1 value is only 60.6%. It suggests that the spatial stream could achieve a good performance on classifying the activities into some object-related groups, however, it can not further classify the specific class from groups with the spatial information only.

TABLE IV shows the overall performance when the ROI automatic selection is removed from the approach, which is similar to the work of [20]. It is suggested that the ROI automatic selection contributes almost 20% of accuracy. Furthermore, the performance of the spatial stream is especially sensitive to the ROI, where the accuracy drops from 61% to 19% in comparison to the temporal stream where the accuracy drops from 78% to 66%). This is probably because the spatial stream is easier to be interfered by the complex driving environment.

### C. Conflicted Cases Analysis

In this section, the details of conflicted cases are presented to further explain the reason why the fusion of two streams can help increase the accuracy of NDA recognition. Fig. 9 presents 3 cases where the fusion result is correct but the result from a single stream is not always right. It includes the “false-true-true case”, “true-false-true case” and “false-false-true case” for the spatial stream only, the temporal stream only and 2-steam respectively. The ground truth class is highlighted by a red block.

From the false-true-true case (the ground truth is PE), for the result of the spatial stream only, the scores of the first four classes are quite close. PB has the highest score that leads to a false result. However, both the temporal stream and 2-stream make the right decision. This is because the interaction mode of

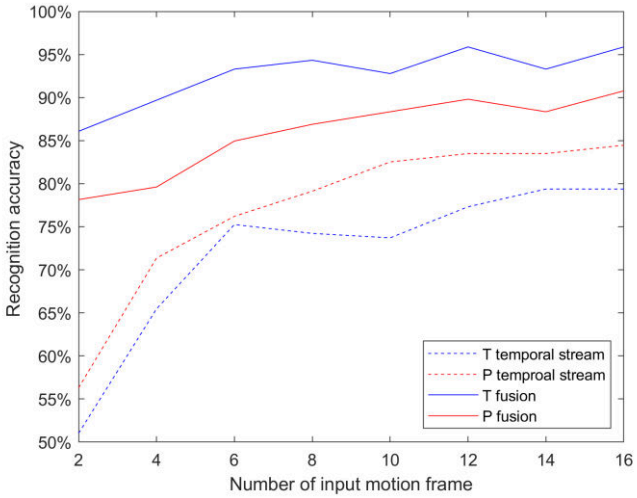



Fig. 10. Impact of number of input motion frame on performance of temporal stream and fusion.

writing email is relatively unique from the others. For the true-false-true case (the ground truth is TR), with the help of the content extracted from the screen, the spatial stream achieves a true prediction although the scores of TB, TE and TR are similar. The prediction result of the optical flow is false due to the interference of PR and PV. This is because hand movement information in these activities is limited. After fusing these 2 streams, the prediction result is true. The bottom subfigure of Fig.9 presents the false-false-true case. Similar to the last case, the temporal stream cannot provide a true prediction due to the similarities between TR and TV. It means that it is hard to differentiate reading and watching videos purely from the optical flow for the same reason above. Meanwhile, the spatial stream also suffers from the interference of PB, PE and PG. However, after combining the two streams, the score of PR is significantly higher than the others, which demonstrates the superiority of the proposed solution.

## IV. Discussions

For the proposed NDA classification system, the performance could be affected by a few factors including the camera position and the number of frames for the temporal stream (N). A few other camera positions have been tested in the experiment including the windscreen in front of the driver, the side window near the front passenger seat. On those positions, a clear view of the object and hands could not be obtained due to occultation caused by human body or steering-wheel. It is essential to recognise the driver and the object from the captured images. The selected camera position achieved the best performance of the tested positions. Although the side window is included, the ROI module can successfully remove this type of noise.

A stack of optical flow frames is regarded as the input of the temporal stream. The performance of the single temporal stream and 2-stream against the number of frames in the stack is presented in Fig. 10, where P indicates the phone-related activities and T indicates the tablet-related activities. It can be observed that, in general, with the increment of N, the recognition accuracy increases due to the consideration of increasing temporal information. However, a larger number of frames also indicates that the system takes more time to determine the type of NDA, which is not helpful for real-time system deployment in the future. In this experiment, the number was set as 10 for the balance.

It should be noted that all analysis of this study are off-line based and the real-time performance is not evaluated. From our point of view, it is not necessary and unlikely to output a decision for every frame because an activity usually is defined as a period of interaction. Using the mentioned PC, the average processing rate is 3.07, 16.38 and 126.17 fps for ROI selection, optical flow estimation and two-stream CNN activity recognition, respectively. It is our notion that the system can update the outcome for every 1 second. Furthermore, the experiments were conducted on a stationary vehicle. There will be some challenges to deploy it to a driving vehicle. For example, camera vibration could introduce the noise to the optical flow estimation. As a computer-vision approach, the rapid variation of illumination will also introduce extra noise for object recognition.

## V. Conclusions

This paper proposed a single-camera-based NDA classification method using a 2-stream CNN benefiting from both spatial and temporal information of an automatically selected RIO. The spatial stream extracts the spatial features of the driver and the engaged object, and the temporal stream characterises the pattern of the interaction behaviour. With this method, different tasks with the same object can be differentiated. The key findings of this study are listed below.

1. The spatial stream achieves good performance in the action recognition dataset like UCF-101, HMDB-51, since the scenario of each action category is quite different. However, for the fine recognition of NDA in this paper, this stream is not sufficient.

2. The content of the tablet screen can help increase the classification accuracy in the spatial stream. However, this is not applicable for small-size objects like phones due to reflection.

3. The temporal stream shows good performance on NDAs involving high-frequency interaction like sending emails or playing games, but low performance on NDAs with very limited interaction such as watching videos or reading.

4. For the conducted experiments, the accuracy of NDA recognition was improved from 61% using the spatial stream and 78% using the temporal stream to 90.5% using the two streams.

5. The inclusion of the ROI automatic selection improves the overall performance from 72.5% to 90.5%.

It should be noted that the proposed system can only be applied to NDAs required physical interaction with the device or object, such as drinking, playing an instrument. A further study is required to tackle other NDAs such as listening to music where other sensors are required.

## Acknowledgment

This work was partly supported by Jaguar Land Rover and

the UK-EPSRC grant EP/N012089/1 as part of the jointly funded Towards Autonomy: Smart and Connected Control (TASCC) Programme; partly supported by Cranfield's EPSRC Impact Accrelate Account EP/R511511/1.

## REFERENCES


[1] L. Yang, K. Dong, A. J. Dmitruk, J. Brighton, and Y. Zhao, "A Dual-Cameras-Based Driver Gaze Mapping System With an Application on Non-Driving Activities Monitoring," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–10, 2019.

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

[3] T. Ersal, H. J. A. Fuller, O. Tsimhoni, J. L. Stein, and H. K. Fathy, "Model-Based Analysis and Classification of Driver Distraction Under Secondary Tasks," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 692–701, Sep. 2010.

[4] J. Kim, W. Kim, H.-S. Kim, and D. Yoon, "Effectiveness of Subjective Measurement of Drivers' Status in Automated Driving," in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, vol. 2018-Augus, pp. 1–2.

[5] M. Bueno, E. Dogan, F. Hadj Selem, E. Monacelli, S. Boverie, and A. Guillaume, "How different mental workload levels affect the take-over control after automated driving," in *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016, pp. 2040–2045.

[6] S. H. Yoon, Y. W. Kim, and Y. G. Ji, "The effects of takeover request modalities on highly automated car control transitions," *Accid. Anal. Prev.*, vol. 123, pp. 150–158, Feb. 2019.

[7] K. Zeeb, A. Buchner, and M. Schrauf, "Is take-over time all that matters? The impact of visual-cognitive load on driver take-over quality after conditionally automated driving," *Accid. Anal. Prev.*, vol. 92, pp. 230–239, Jul. 2016.

[8] B. Wandtner, G. Schmidt, N. Schömig, and W. Kunde, "Non-driving related tasks in highly automated driving - Effects of task modalities and cognitive workload on take-over performance," *AmE 2018 - Automot. meets Electron. 9th GMM-Symposium*, pp. 1–6, 2018.

[9] C. Wu, H. Wu, N. Lyu, and M. Zheng, "Take-Over Performance and Safety Analysis Under Different Scenarios and Secondary Tasks in Conditionally Automated Driving," *IEEE Access*, vol. 7, pp. 136924–136933, 2019.

[10] N. Li and C. Busso, "Detecting Drivers' Mirror-Checking Actions and Its Application to Maneuver and Secondary Task Recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 980–992, Apr. 2016.

[11] L. Jin, B. Guo, Y. Jiang, F. Wang, X. Xie, and M. Gao, "Study on the Impact Degrees of Several Driving Behaviors When Driving While Performing Secondary Tasks," *IEEE Access*, vol. 6, pp. 65772–65782, 2018.

[12] M. Martin, J. Popp, M. Anneken, M. Voit, and R. Stiefelhagen, "Body Pose and Context Information for Driver Secondary Task Detection," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018, vol. 2018-June, no. Iv, pp. 2015–2021.

[13] Y. Xing *et al.*, "Identification and Analysis of Driver Postures for In-Vehicle Driving Activities and Secondary Tasks Recognition," *IEEE Trans. Comput. Soc. Syst.*, vol. 5, no. 1, pp. 95–108, Mar. 2018.

[14] M. Sivak, B. Schoettle, "Motion Sickness in Self-Driving Vehicles," Transportation Res. Inst., Ann Arbor, Univ. Michigan, Ann Arbor, MI, USA, Tech. Rep. UMTRI-2015-12, Apr. 2015

[15] Bangpeng Yao and 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.*, vol. 34, no. 9, pp. 1691–1703, Sep. 2012.

[16] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: A literature review," *Pattern Recognit.*, vol. 48, no. 8, pp. 2329–2345, Aug. 2015.

[17] T. H. N. Le, C. Zhu, Y. Zheng, K. Luu, and M. Savvides, "DeepSafeDrive: A grammar-aware driver parsing approach to Driver Behavioral Situational Awareness (DB-SAW)," *Pattern Recognit.*, vol. 66, no. December 2016, pp. 229–238, Jun. 2017.

[18] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, vol. 2015 Inter, pp. 4489–4497.

[19] S. R. Sreela and S. M. Idicula, "Action Recognition in Still Images using Residual Neural Network Features," *Procedia Comput. Sci.*, vol. 143, pp. 563–569, 2018.

[20] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos," *Biochem. Pharmacol.*, vol. 32, no. 5, pp. 849–855, Jun. 2014.

[21] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2556–2563.

[22] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," no. November, Dec. 2012.

[23] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 4724–4733.

[24] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[25] T. Pfister, J. Charles, and A. Zisserman, "Flowing convnets for human pose estimation in videos," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2015 Inter, pp. 1913–1921, 2015.

[26] Joe Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, vol. 07-12-June, pp. 4694–4702.

[27] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the Integration of Optical Flow and Action Recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11269 LNCS, 2019, pp. 281–297.

[28] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "DeepFlow: Large Displacement Optical Flow with Deep Matching," in *2013 IEEE International Conference on Computer Vision*, 2013, no. Section 2, pp. 1385–1392.

[29] C. Bailer, B. Taetz, and D. Stricker, "Flow Fields: Dense Correspondence Fields for Highly Accurate Large Displacement Optical Flow Estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1879–1892, Aug. 2019.

[30] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, vol. 2017-Janua, pp. 1647–1655.

[31] G. Chantas, T. Gkamas, and C. Nikou, "Variational-Bayes Optical Flow," *J. Math. Imaging Vis.*, vol. 50, no. 3, pp. 199–213, Nov. 2014.

[32] J. Donahue *et al.*, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.

[33] K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, vol. 2016-Decem, pp. 770–778.

[35] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of COMPSTAT'2010*, Heidelberg: Physica-Verlag HD, 2010, pp. 177–186.