applied sciences

MDPI

Article

Driver Behavior Analysis via Two-Stream Deep Convolutional Neural Network

Ju-Chin Chen *, Chien-Yi Lee, Peng-Yu Huang and Cheng-Rong Lin

Department of Computer Science and Information Engineering, National Kaohsiung University of Science and Technology, Kaohsiung city 8078, Taiwan; jc.strength8@gmail.com (C.-Y.L.); t1240370@gmail.com (P.-Y.H.); i07152110@nkust.edu.tw (C.-R.L.)

* Correspondence: jc.chen@nkust.edu.tw

Received: 31 October 2019; Accepted: 16 February 2020; Published: 11 March 2020

Abstract: According to the World Health Organization global status report on road safety, traffic accidents are the eighth leading cause of death in the world, and nearly one-fifth of the traffic accidents were cause by driver distractions. Inspired by the famous two-stream convolutional neural network (CNN) model, we propose a driver behavior analysis system using one spatial stream ConvNet to extract the spatial features and one temporal stream ConvNet to capture the driver's motion information. Instead of using three-dimensional (3D) ConvNet, which would suffer from large parameters and the lack of a pre-trained model, two-dimensional (2D) ConvNet is used to construct the spatial and temporal ConvNet streams, and they were pre-trained by the large-scale ImageNet. In addition, in order to integrate different modalities, the feature-level fusion methodology was applied, and a fusion network was designed to integrate the spatial and temporal features for further classification. Moreover, a self-compiled dataset of 10 actions in the vehicle was established. According to the experimental results, the proposed system can increase the accuracy rate by nearly 30% compared to the two-stream CNN model with a score-level fusion.

Keywords: CNN; two-stream convolutional neural network; driver behavior analysis

1. Introduction

According to the World Health Organization (WHO) global status report on road safety, in 2018, traffic accidents were the eighth leading cause of death in the world [1]. Nearly 135 million people were killed and injured every year. According to the report, nearly one-fifth of the traffic accidents were cause by driver distractions [2]. Although the mortality rate of traffic accidents in the world declined in recent years, compared with other countries, the mortality rate is much higher in Taiwan. In addition, due to the development of science and technology in recent years, in-vehicle information systems (IVISs) such as navigation and media devices were installed in cars. These devices would introduce more driver distractions and lead to more accidents [2]. Road traffic accidents cause huge damage, and the number of accidents due to distractions is increasing. According to the National Highway Traffic Safety Administrator of the United States (NHTSA) reports [2], human errors caused approximately 90% road accidents in the United States and represented a dominant factor for vehicle crashes [3]. Among them, the major cause of these accidents was the use of mobile phones [4]. As a result, the government enacted several laws to punish drivers and prevent distractions caused by the use of hi-tech products while driving.

In fact, using a smartphone is not the only reason for driver distractions. NHTSA defines a distraction during driving as "any activity that diverts attention of the driver from the task of driving", including talking or texting on one's phone, eating and drinking, or talking to people in the vehicle [4]. The Centers for Disease Control and Prevention (CDC) [5] classified distracted driving as cognitive,

visual, and manual distractions. Cognitive distraction means that the driver's mind is off driving. In other words, even though the driver is in a safe driving posture, they might be lost in their thoughts and mentally distracted from safe driving. Visual distractions refer to situations where the driver's eyes are off the road because of fatigue, sleepiness, drowsiness, inattention, or the use of multimedia devices. Manual distractions are concerned with various activities where the driver's hands are off the driving wheel. Such distractions include using the cellphone, eating or drinking, adjusting hair and makeup, or talking to a passenger.

In order to prevent accidents, distraction detection systems become an important component in semi-autonomous or autonomous cars, which can alert the driver to potential problems. Motivated by the fact that the major cause of manual distractions is the usage of the cellphones [6], some researches focused on cellphone usage detection while driving. In 2011, Zhang et al. [7] extracted features of the face, mouth, and hands from the images captured by a camera installed on the dashboard, and a hidden condition random field (HCRF) model was applied to detect cellphone usage. In 2014, Berri et al. [8] applied an support vector machine (SVM) model to check the hand and face locations and detect the usage of cellphones with a frontal image view of the driver. In 2015, Craye et al. [9] used AdaBoost and hidden Markov models to classify driver distraction by analyzing the RGB-D data captured by Kinect sensors. However, in the data collection process, the experimental set-up missed two essential points: the lighting conditions and the distance between the sensor and the driver [2]. In real cases, a driver is exposed to a variety of lighting conditions, including sunlight and shadows [2]. Seshardi et al. [10] created their own dataset for cellphone detection and applied a supervised descent method (SDM) to track the locations of face landmarks for the extraction of regions of interest. Unlike previous methods, no assumption where a face or hands were expected to be found was made. The authors applied a histogram of gradients (HOG) and an AdaBoost classifier trained for each side of the face regions to classify cellphone usage as right hand, left hand, or no usage Das et al. [11] introduced a video-based hand detection dataset in an automotive environment and used the aggregate channel features object detector.

In recent years, with the great success achieved by deep learning networks in computer vision [12–14], some deep-learning-based object detection models were applied to detect hand location. In 2016, Le et al. [15] trained a faster R-CNN model [16] to classify whether the hands are holding a steering wheel or not, and the results showed that the model could achieve a higher accuracy rate than found in Reference [10]. In 2016, Yuen et al. applied AlexNet to perform head pose estimation and used the stacked hourglass network in the refinement stage to estimate facial landmarks and refine the face localization [17]. The detected results can provide a basis for the estimation of the driver's state in terms of distraction and drowsiness. In addition, some studies were proposed to recognize the driver's behavior to detect distraction. Then, in Reference [18], the authors proposed DarNet and investigated the mixing of different models, i.e., CNNs, recurrent neural network (RNNs), and SVMs, to detect driver distraction. In 2018, Majdi et al. proposed Drive-Net to classify 10 distracted behaviors, which was composed of a convolutional neural network (CNN) and a random decision forest [19]. Tran et al. [20] utilized four different CNN models including VGG-16 [21], AlexNet [12], GoogleNet [22,23], and a residual network to classify 10 distracted behaviors. In addition, they developed a warning system that can alert the driver in real time when a distraction behavior is detected. According to their results, the authors observed that the deeper models could provide higher detection accuracy, but the cost of inference time was increased. The trade-off between accuracy and efficiency remains an issue. In addition, only spatial information was considered, but temporal information, related to important cues for behavior recognition, was ignored in recent works [19,20,24–26]. Although behavior analysis or action recognition was studied for many years in the computer vision field, driver behavior analysis is a specific issue and is challenging due to the light changing, occlusion, clutter, and subtle actions.

In this study, we aimed to analyze driver behavior and detect manual distractions in order to develop a warning system for drivers. For action recognition, temporal information is important, as well as spatial information, but it was ignored in recent works of distracted behavior recognition.

Hence, inspired by the famous network architecture, using a two-stream CNN model [27] composed of a spatial and a temporal stream ConvNet, the spatial and temporal information was extracted using powerful CNN models for the classification of common 10 distracted behaviors inside the vehicle. For the spatial information, an average pooling in the temporal domain was performed on 10 consecutive RGB images, and then the resulting map was input to the following convolutional layers. The feature map from the last convolutional layer was taken as the spatial features for further processing. On the other hand, TVL[1] optical flow [28] was firstly applied to extract the horizontal and vertical motion information between input frames, and then two stacks of flow images were concatenated and input to the temporal stream ConvNet. The feature map from the last convolutional layer was extracted as the temporal features. Note that, instead of using a three-dimensional (3D) ConvNet, which would suffer from large parameters and the lack of a pre-trained model, two-dimensional (2D) ConvNet was used to construct the spatial and temporal ConvNet streams, and they were pre-trained using the large-scale ImageNet [12]. Moreover, in order to integrate different modalities, rather than using manually defined weights in the score-level fusion [27] or weights obtained via the optimization process [24], the feature-level methodology was applied and a fusion network was proposed to analyze the concatenated spatial and temporal features for classification.

The remainder of this paper is organized as follows: Section 2 reviews the issues of driver distraction; Section 3 presents the proposed system consisting of three modules, spatial deep network, temporal deep network, and integration module; Section 4 presents the experiment results of the proposed system and compares it with other systems; Section 5 concludes the paper.

2. Related Works

In the computer vision field, action recognition was developed for many years. Among these researches, how to extract action features is a key issue. In early years, studies focused on the hand-crafted local features that were designed by human experts to extract a given set of chosen characteristics [29]. Since temporal information includes important cues for action recognition, famous feature descriptors for image classification such as scale-invariant feature transform (SIFT) and histogram of oriented gradient (HOG) were extended to extract features for 3D data (2D spatial + one-dimensional (1D) temporal data). In 2003, Laptev et al. [30] proposed spatio-temporal interest points (STIPs) by extending Harris corner detectors. SIFT and HOG were also extended to SIFT-3D and HOG3D for action recognition. Dollar et al. [31] proposed the cuboid feature for behavior identification. Sadanand and Corso [32] established the ActionBank for action recognition. In 2013, Wang et al. [33] proposed improved dense trajectories (iDT). Although action recognition was studied for at least 20 years, the performance is still limited due to the difficulties including large appearance and pose variations, as well as a cluttered background [34]. With the development of convolutional neural networks (CNNs), CNN showed the advantages of performance improvement for action recognition compared to hand-crafted features [29]. It also makes real-life applications of action recognition possible. Among them, driver behavior analysis is a specific issue and becomes very important because in-vehicle information systems (IVISs) such as navigation and media devices are being developed rapidly but they introduce more distractions which cause accidents.

In early studies of driver distraction detection, motivated by the fact that the major cause of distractions is using a cellphone while driving [6], the detection of cellphone usage became a focus and many machine learning models were applied. For example, in 2011, Zhang et al. [7] extracted the features from the image captured by a camera installed on the dashboard and a hidden condition random field model was applied. Berri et al. [8] applied an SVM model to check the hands and face locations with the frontal image view of the driver. In order to cope with occlusion, Craye et al. [9] used the RGB-D data captured by Kinect sensors, but the distance between the sensor and the driver was missed in the data collection process. Unlike previous methods that assumed safe driving based on the location or state of the face or hands, Seshardi et al. [10] applied a supervised descent method

to track the locations of face landmarks, and an AdaBoost classifier trained for each side of the face regions was used to classify the cellphone usage as right hand, left hand, or no usage.

According to the definition of a distraction by the NHTSA as "any activity that diverts attention of the driver from the task of driving" [4], the CDC [5] classified distracted driving as cognitive, visual, and manual distractions, whereby not only cellphone usage detection was developed but more distracted behaviors were also considered. In 2014, Martin et al. [35] collaborated with the University of California, San Diego (UCSD) Laboratory of Intelligent and Safe Automobiles and presented a vision-based hand activity analysis system to detect three types of distractions: adjusting video, adjusting mirrors, and operating gear. The images are captured by two Kinect cameras that can provide the frontal and back views of the driver. Ohn-bar et al. [36] divided images into three regions: steering wheel, gear, and radio panel. A region-based model was designed to detect the presence of hands in certain pre-defined regions to classify three types of distractions. A more inclusive distracted driving dataset including four distractions (safe driving (holding steering wheel), operating shift, eating, and talking on the cellphone) was considered in Reference [37]. Zhao et al. [37] used contourlet transform for feature extraction, and different classifiers were applied including random forests, k-nearest neighbor, and multiplayer perceptron. Note that, in earlier years the datasets defined a limited number of distracted behaviors for distraction detection, and most datasets were not public. In 2016, StateFarm's distracted driver detection competition on Kaggle [38] was the first publicly available dataset for competition purposes. They defined 10 distractions to be detected: safe driving, texting using right hand, talking on the phone using right hand, texting using left hand, talking on the phone using left hand, operating the radio, drinking, reaching behind, doing hair and makeup, and talking to passenger. In 2017, Abouelnaga et al. [24] created a new AUC Distracted Driver dataset similar to StateFarm's dataset. The dataset was composed of the same 10 distraction behaviors and 31 participants from seven different countries in four different cars. They also proposed a real-time distracted driver posture classification system.

Driver behavior recognition is a kind of action recognition where the performance is rapidly increasing because of the rise of deep learning models [12–14], where some network architectures were developed for action recognition [39]. By extending the network architecture for image classification and the spirit of the bag-of-words model to extract features and identify each video frame [39,40], static CNNs were used to identify single or several video frames, and then the final recognition result was obtained by averaging scores across the whole video [40]. However, the motion information in the temporal domain between objects is ignored and the lack of specific temporal structure leads to only a slight accuracy improvement compared to methods based on hand-crafted features. Hence, some researchers improved the network architecture [39,41–46]. The studies [41,42] added a recurrent layer, such as an long short-term memory (LSTM), to the 2D ConvNets (2D ConvNets + LSTM) and, thus, the temporal ordering could be captured and encoded in the states. Although LSTMs on the features from the last layers of 2D ConvNets could model high-level variations, fine low-level motion could not be captured, and the detailed descriptor of the network was lost because of the backpropagation [39]. The training time was also increased because it required unrolling the network through multiple frames for backpropagation through time [27,39]. On the other hand, 3D ConvNets [43,44] seem to be a natural method to describe videos, and it can extract spatio-temporal features. However, there are two problems with this model [27]. Firstly, it is harder to train than 2D ConvNets because more parameters are needed. Secondly, due to the large network architecture and 3D filter usage, it cannot directly use the pre-trained model on large-scale datasets, such as ImageNet. As a result, it is necessary to train from scratch, and an overfitting situation would happen in the case of an increased number of parameters. In 2017, Simonyan and Zisserman [39] proposed a practical approach by modeling spatial and short temporal snapshots via two-stream networks. A single RGB image and a stack of 10 continuous optical-flow frames were input to the 2D ConvNets which were pre-trained on the ImageNet dataset, while, in the test process, multiple snapshots were sampled from the video and the action prediction was averaged. According to the experimental results, this network architecture

achieved high performance on the existing benchmarks, while the network was very efficient to train and test [39]. In addition, other models such as bidirectional LSTM [45] or neural hypergraph-based model [46] for sequence prediction in natural language processing could be applied as well.

For distraction behavior recognition, some deep-learning-based methods were proposed in recent years. In 2016, Le et al. [15] trained a faster RCNN model to classify whether the hands are holding the steering wheel or not, and the results showed that the work could achieve a higher accuracy rate than found in Reference [10]. In Reference [18], Streiffer et al. proposed DarNet and investigated the mixing of different models, i.e., CNNs, RNNs, and SVMs, to detect driver distraction. In 2018, Majdi et al. [19] proposed Drive-Net, which was composed of a convolutional neural network (CNN) and a random decision forest, for the classification of 10 distracted behaviors. Tran et al. [20] utilized four different CNN models including VGG-16, AlexNet, GoogleNet, and residual network to classify 10 distracted behaviors. Yen et al. [47] applied a CNN model to extract features and recognize four postures, including normal driving, using a cell phone call, eating, and smoking on the Southeast University (SEU) driving posture dataset [37]. Note that the authors also conducted the system on self-compiled infrared images, which could be invariant to illumination changes seen in daytime/night conditions for real driving environments. The results demonstrated better performance than conventional classifiers, e.g., SVM, with hand-crafted features. In 2017, Abouelnaga et al. [18,24] proposed a real-time distracted driver posture classification system. The input image was firstly pre-processed by the detection and segmentation methods that were applied to the regions of face, hands, and skin, and then a weighted ensemble of four different convolutional neural networks was applied for classification. In 2018, Masood et al. [26] applied VGG-16 and VGG-19 to recognize 10 distraction actions on the StateFarm dataset and showed that the impressive results achieved by the CNN models and the usage of the pre-trained model indeed saved training time. Baheti et al. [25] proposed a CNN-based system by modifying the VGG-16 network, including using the leaky rectified linear unit (Leaky ReLU) activation function instead of ReLU and applying various regularization techniques to cope with the problem of overfitting. Moreover, the authors proposed the network of modified-VGG16 by replacing the whole fully connected layer with a convolutional layer in order to reduce the number of network parameters. The results showed that the system could achieve 96.31% accuracy on the AUC Distracted Driver dataset. However, in recent works [19,20,24–26], only spatial information was considered, but temporal information, related to important cues for behavior recognition, were discarded. In Reference [48], Chuang et al. proposed a skeleton-based and a point-cloud approach with multiple views based on Kinect depth cameras for driver behavior recognition, and LSTM was adopted to train the behavior model. The authors evaluated the proposed system on the VAP multiple views dataset which was collected in the laboratory to simulate an in-vehicle scene. However, the distance between the sensor and the driver was missed, and there was a difficulty in the training process inherited from the LSTM model [27,39]. Additionally, there were related issues such as driver intention prediction to anticipate driver maneuvers. In Reference [49], Gebert et al. proposed an end-to-end network architecture which consisted of FlowNet [26] to extract optical flow, a 3D residual network for maneuver classification, and an LSTM model for handling temporal data with varying length. Note that FlowNet was used to extract the optical flow in the video interpolation as well; however, labeling the ground-truth flow data to train FlowNet for a specific task is hard work and time-consuming.

Moreover, in order to cope with illumination changes encountered in realistic driving scenarios, there are other methods of analyzing data captured by illumination-invariant sensors such as depth sensors [9,48,50], infrared cameras [47], or fusion of various sensor types [9,51]. In 2019, Martin et al. [52] presented the first large-scale dataset, Drive & Act dataset, for fine-grained categorization of driver behavior. The image frames were captured by six different views and three modalities that were collected by five near-infrared cameras and Kinect v2 cameras used to acquire color, infrared, and depth data. The authors aimed to facilitate researches for video- and pose-based action recognition.

3. Method

For action recognition, temporal information is important, as well as spatial information. Motivated by work [27] using temporal information, we proposed an architecture of two-stream convolutional networks for distracted detection as shown in Figure 1. The network is composed of three sub-networks: spatial stream ConvNet, temporal stream ConvNet, and a fusion network. Spatial stream ConvNet and temporal stream ConvNet are used to extract the spatial and temporal features, respectively, and the different features are integrated in the fusion network. By benefitting from transfer learning, the spatial stream ConvNet was designed based on the famous network configuration, VGG-16, and the pre-trained model on the ImageNet dataset can be applied. The average pooling in the temporal dimension is firstly performed on 10 consecutive RGB images, and the result is passed to the following layers of the spatial stream ConvNet. Via convolution operators performed in the convolutional layers, the spatial features are obtained from the feature map in the last convolutional layer. On the other hand, instead of using the famous 3D ConvNet that would suffer from large parameters and overfitting when there is lack of a large training dataset [27], the temporal features are obtained by extracting the motion information via the TVL[1] optical flow [28] from video frames and then a stack of consecutive flow images capturing vertical and horizontal motion information are input to the temporal stream ConvNet. Then, the temporal features are obtained from the feature map in the last convolution layer of temporal stream ConvNet. In the end, a fusion network, consisting of two convolutional layers and two fully connected layers, was designed to integrate the spatial and temporal features to classify 10 distracted behaviors.

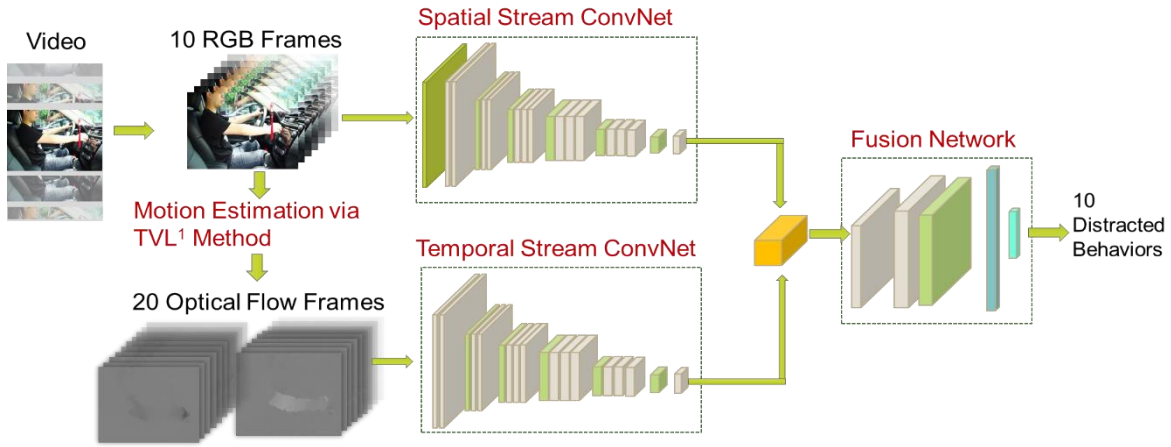

Figure 1. Driver distraction system flowchart.

3.1. Spatial Stream ConvNet

The convolutional neural network (CNN) plays an important role in deep learning models. In recent years, many studies relied on famous network architectures such as AlexNet [12], VGG-16 [21], and GoogleNet [22,23]. These networks usually consist of convolutional layers and pooling layers, followed by few fully connected layers. Some studies showed that better accuracy can be achieved using a deeper network. By considering the balance of performance and computation efficiency and the promising results [26], the spatial stream ConvNet was based on the network configuration of VGG-16 [21], which was designed for image classification and was proven to effectively extract the features of images layer by layer. Figure 2 shows the network configuration of spatial stream ConvNet. The network input is a stack of 10 consecutive RGB frames, each of which were resized to 224 × 224 pixels. By benefitting from the transfer learning of using the pre-trained model on the large-scale dataset, averaging pooling of the temporal direction is firstly performed on these input frames, and the resulting map with the size of 224 × 224 × 3 pixels is obtained. Following the average pooling layer, the network is composed of 13 convolution layers, while five max pooling and the ReLU

activation function are set in each layer. Unlike other CNN models which set different kernel size in convolutional layers, a small kernel size of 3×3 pixels is set in all convolution layers which keeps the scale-invariant feature transform after convolution by using the same padding mechanism. The max pooling can help to extract the feature information of a larger area. Although the configuration of the VGG-16 is simple and effective, the large number of parameters (140 million) on fully connected layers is the main problem [21]. This leads to high cost for the training and test process. Hence, in our work, unlike the original VGG-16 configuration, the fully connected layer was not used in the proposed spatial stream ConvNet. Here, a feature map with the size of $7 \times 7 \times 256$ pixels from the last convolution layer is extracted as the spatial features and further processed in the following fusion network.

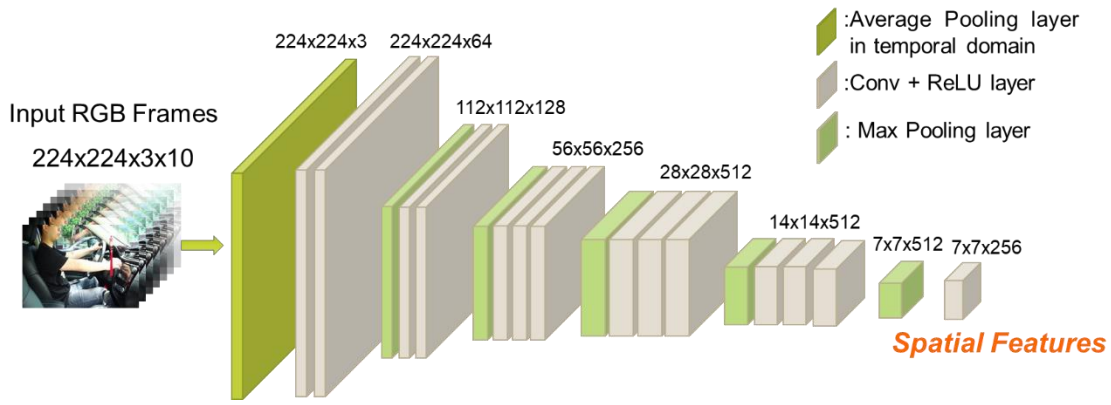

Figure 2. Configuration of spatial stream ConvNet.

Since the parameters of CNN models are large, transfer learning, using the pre-training model to initialize the network, is applied in the training process. For driver distracted behavior analysis, the hand movements are the main changes. Hence, rather than using the pre-trained model on the ImageNet, the spatial stream ConvNet was pre-trained on the dataset that contains the videos of N ($N = 10$ in our study) actions containing obvious hand movements from thUCF-101 dataset [53]. Additionally, data augmentation, including cropping, rotating, horizontal flipping, and shifting, was performed to enlarge the size of dataset. Note that, in order to obtain the pre-trained model with the ability to extract discriminative features, a softmax layer with the size of $1 \times 1 \times N$ was added after the last convolutional layer, which was removed in the fine-tuned process. The network was trained by stochastic gradient descent with a learning rate of 0.0001, decay rate of 10^{-6}, and momentum value 0.9. The batch size and number of epochs were set to 32 and 100, respectively.

3.2. Temporal Stream ConvNet

Although distracted behavior analysis was studied in recent works [19,24–26], only spatial information was considered while temporal information was discarded. In previous studies of action recognition, many network configurations were designed to integrate the spatial and temporal information, for example, 2D ConvNets + LSTM [41,42], 3D ConvNet [43,44], and two-stream network [27]. Instead of using the famous 3D ConvNet [43,44] that would suffer from large parameters and overfitting when there is lack of a large training dataset [27], the two-stream network was used to design the temporal stream ConvNet. In order to extract the motion information, TVL[1] optical flow [28] with default parameters in OpenCV is firstly applied to obtain vertical and horizontal flow frames between two consecutive frames.

Then, the optical frames of two directions are concatenated, and a stack of 20 flow images with the size of $224 \times 224 \times 20$ pixels are input to the temporal stream ConvNet. Figure 3 shows the configuration of the temporal stream ConvNet. Following the input layer, the network is composed of 13 convolution layers with a kernel size of 3×3 pixels, while five max pooling and the ReLU activation function are set in each layer. Note that, in our work, unlike the original VGG-16 configuration, the

fully connected layers were not used in the proposed temporal stream ConvNet. Here, a feature map with the size of $7 \times 7 \times 256$ pixels from the last convolution layer is extracted as the temporal features and fused with the spatial features in the following fusion network.

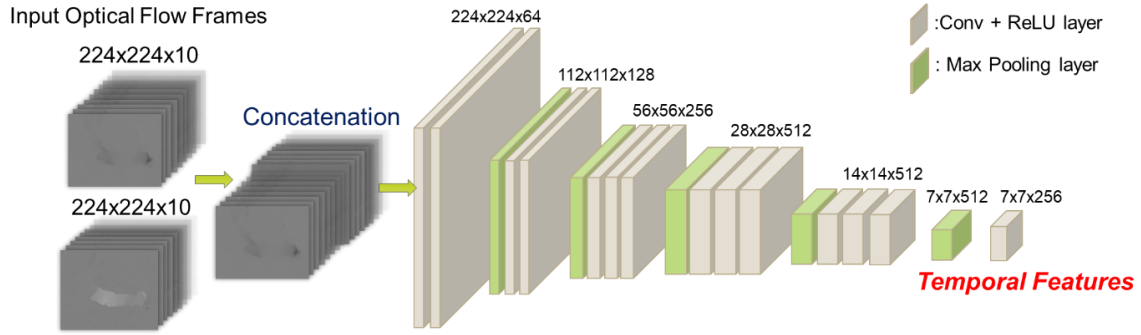

Figure 3. Configuration of temporal stream ConvNet.

In order to train the temporal stream ConvNet, the pre-trained dataset, consisting of videos of N ($N = 10$) actions involving hand motions from the UCF-101 dataset, as used in the training process of the spatial stream ConvNet, was applied. Data augmentation involving random cropping and horizontal flipping was also performed to enlarge the dataset. The vertical and horizontal motion information was estimated between two frames via TVL[1] optical flow. Additionally, in order to reduce the impact caused by the camera movement, the mean flow frame of each direction is calculated, and the flow frames were subtracted from the corresponding mean frame. For the pre-training process, a softmax layer with the size of $1 \times 1 \times N$ was added after the last convolutional layer, which was removed in the fine-tuned process, and the hyper-parameters were the same as used in the spatial stream ConvNet. The pre-trained model was then obtained, which was used to initialize the temporal stream ConvNet in the training process of the whole network.

3.3. Classification of Distractions by a Fusion Network

After the spatial and temporal stream ConvNet, the spatial and temporal feature maps with a size of $7 \times 7 \times 256$ pixels are obtained. In other to fuse different modalities, score-level and feature-level fusion are common methodologies [54]. Since the dimension of features is high, i.e., 25,088, rather than using either the manual-defined weights or weights obtained via the optimization method, e.g., genetic algorithm (GA) [24], at the score level, a fusion network was designed to fuse features in the feature level. Figure 4 shows the configuration of the fusion network. Two kinds of feature maps are concatenated in the third dimension, and then the resulting feature map with a size of $7 \times 7 \times 512$ pixels is input to the following convolutional layer. Since the CNN model as used in the spatial and temporal stream ConvNet has promising ability of feature extraction, the fusion network was not designed with deep layers in order to reduce the risk of overfitting. In our study, the fusion network consisted of two convolutional layers, as well as one pooling layer and two fully connected layers with sizes of $1 \times 1 \times 512$ and $1 \times 1 \times 10$, respectively. The kernel size of the first and second convolutional layers was 1×1 and 3×3 pixels, and the ReLU activation function was applied to both convolutional layers. The classification result of 10 distracted behaviors is obtained in the softmax layer.

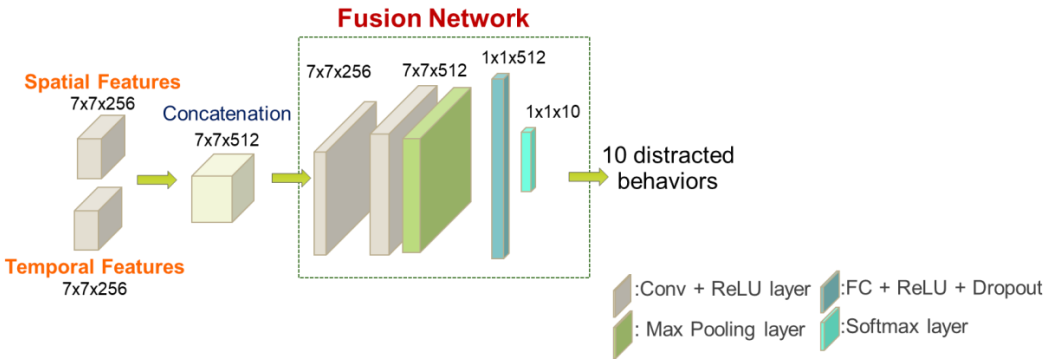

Figure 4. Configuration of fusion network.

In the training process, the whole network, including the spatial stream ConvNet and the temporal stream ConvNet with the pre-trained weights and the fusion network, was trained by the self-compiled dataset. Note that all layers of the spatial stream ConvNet and the temporal stream ConvNet were fine-tuned. A higher dropout value of 0.9 was set in the fully connected layers to reduce interdependent learning amongst the neurons, which is an efficient way of coping with overfitting.

4. Experiment Results

We evaluated the performance of the proposed driver distraction detection system in the self-compiled dataset. The experimental set-up is firstly introduced, including the dataset and the computing environment. Then a performance comparison with other existing systems is performed.

4.1. Dataset and Experimental Setting

A limited set of distraction classes were included in earlier datasets and most datasets were not public. The dataset of StateFarm on Kaggle [38] was the first publicly available dataset containing 10 distraction classes, used for competition purposes. The AUC Distracted Driver dataset [24], shown in Figure 5, is similar to the StateFarm dataset with the same 10 distraction classes. However, the sampling rate is low, and the motion information between consecutive frames is unstable. Hence, in this study, we referred to the StateFarm and AUC Distracted Driver datasets to compile our own dataset as shown in Figure 6. Ten driver distraction classes were also defined [24,38], including safe driving, text right, phone right, text left, phone left, adjusting radio, drinking, reaching behind, hair or makeup, and talking to passenger. We had 20 participants and the videos were shot in the same car. The videos were divided into 198 films, 28,779 RGB images, and 57,558 optical flow images of horizontal and vertical components. All images were resized to 344 × 356 pixels before the training and test process.

Figure 5. AUC Distracted Driver dataset [24].

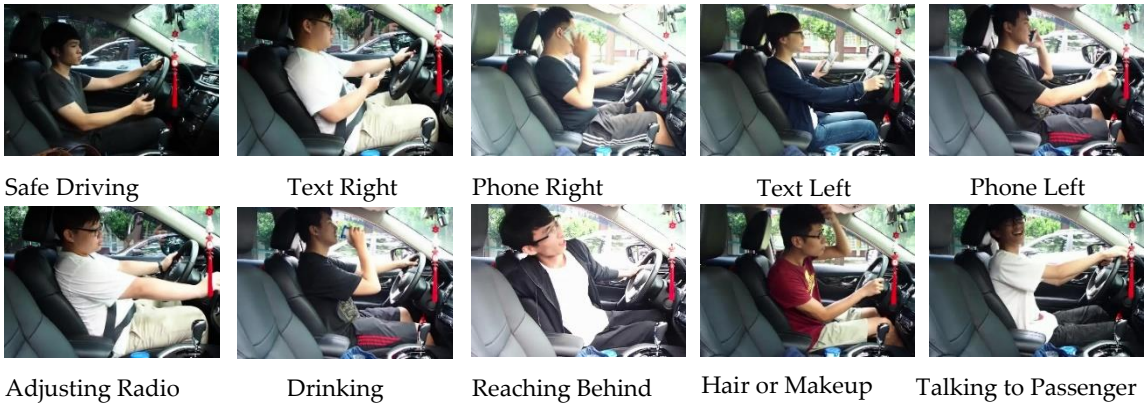

Figure 6. Self-compiled distraction dataset.

Due to the limited amount of data in our dataset, the pre-trained model on UCF-101 [53] was used for network initialization, and then all the layers of spatial and temporal ConvNet streams were fine-tuned. In order to reduce the training time, videos of the 10 actions with large hand movements were selected from the UCF-101 dataset. All experiments were performed on a personal computer (PC) with an NVIDIA GeForce GTX TITAN X graphics processing unit (GPU) having 3072 CUDA cores with 12 GB of random-access memory (RAM), using an Intel i7-6700K central processing unit CPU and 16 GB of RAM; the frameworks of Keras and Tensorflow were used.

4.2. Comparison with Existing Systems

We firstly performed experiments on the AUC Distracted Driver Dataset to evaluate the accuracy of the spatial network. As with the train/test data split proposed in Reference [17], it was found that there was a high correlation between training and test data. In other words, similar images of a person performing one distraction behavior would be in the training and test data. In order to avoid this situation, we re-split the AUC dataset according to the drivers' identifier (ID), and the IDs used in training and test data were separated. The experimental results were compared with Reference [25] as shown in Table 1. "Original Split" means the original data split used in Reference [25], and "Re-Split" means the data was re-split according to the drivers' IDs. The study applying VGG-16 with regularization was compared. Note that we only evaluated the performance of the spatial stream ConvNet because, in the AUC Distracted Driver Dataset, the sampling rate was low which resulted in discontinuous motion. This would cause very large errors for motion estimation, and the proposed temporal stream ConvNet was not suitable for this case. According to the results, it was found that the data split affected the performance much. The results of using different data splits were quite different. The regularization methodology could increase the accuracy rate by about 1% to 2%.

Table 1. Performance evaluation and comparison for spatial stream ConvNet in the AUC Distracted Driver Dataset.

Model	Dataset	Accuracy (%)
Spatial Stream ConvNet	Original Split	94.44
VGG-16 with Regularization [25]	Original Split	96.31
Spatial Stream ConvNet	Re-Split	76.25
VGG-16 with Regularization [25]	Re-Split	77.15

Since the displacement of the objects between two consecutive frames is large in the AUC Distracted Driver Dataset, it might cause errors in the process of motion estimation. We compiled our distraction dataset for performance evaluation. In the training process, stochastic gradient descent was used for an optimization learning rate of 0.0001, decay rate of 10^{-6}, and momentum value of 0.9. The batch size and number of epochs were set to 16 and 500 respectively. We compared the results

with the two-stream method [27], and the results are shown in Table 2. Firstly, it was found that the two-stream method [27] and the proposed method indeed increased the accuracy rate of the pre-trained model, especially for the temporal stream ConvNet in the proposed method. Secondly, fusion results were better than using either spatial or temporal stream ConvNet alone. Hence, fusing information is recommended. In addition, the proposed method increased the accuracy rate by nearly 30% compared with Reference [27].

Table 2. Accuracy rate in the self-compiled dataset.

Model	Spatial Stream ConvNet	Temporal Stream ConvNet	Fusion Result
Two-Stream Method 27	9.52	39.68	38.10
Two-Stream Method 27 **with Pre-trained model**	12.98	41.27	39.68
Proposed Method	33.34	9.52	34.92
Proposed Method with Pre-trained models	**49.21**	**65.08**	**68.25**

Table 3 shows the confusion matrix for distraction behavior analysis in the self-compiled dataset. The proposed system provided better results for the actions of *text left* and *phone left*, while the action of *phone right* was easily confused with the action of *hair or makeup*.

Table 3. Confusion matrix in the self-compiled dataset.

	Safe Driving	Text Right	Phone Right	Text Left	Phone Left	Adjusting Radio	Drinking	Reaching Behind	Hair or Makeup	Talking to Passenger
Safe Driving	**0.80**	0.0	0.0	0.0	0.20	0.0	0.0	0.0	0.0	0.0
Text Right	0.0	**0.75**	0.0	0.0	0.0	0.0	0.25	0.0	0.0	0.0
Phone Right	0.0	0.0	0.33	0.0	0.0	0.0	0.17	0.0	0.50	0.0
Text Left	0.0	0.0	0.0	**1.00**	0.0	0.0	0.0	0.0	0.0	0.0
Phone Left	0.0	0.0	0.0	0.0	**1.00**	0.0	0.0	0.0	0.0	0.0
Adjusting Radio	0.11	0.22	0.0	0.0	0.0	**0.56**	0.0	0.11	0.0	0.0
Drinking	0.0	0.13	0.25	0.0	0.0	0.0	**0.5**	0.0	0.12	0.0
Reaching Behind	0.0	0.0	0.0	0.0	0.0	0.25	0.0	**0.75**	0.0	0.0
Hair or Makeup	0.0	0.0	0.14	0.0	0.0	0.0	0.0	0.0	**0.86**	0.0
Talking to Passenger	0.0	0.0	0.11	0.0	0.0	0.0	0.0	0.23	0.0	**0.67**

5. Conclusions

Driver distraction is one of the major causes of traffic accidents, especially in Taiwan. It would be worth developing a system to detect driver distraction for automatic vehicles. Inspired by the famous two-stream CNN model, we proposed a driver behavior analysis system using CNNs to analyze the input consecutive frames, and the feature maps from the last convolution layer were extracted as the spatial and temporal features for further classification. Unlike previous studies using manually defined weights or weights obtained via the optimization process, a fusion network was designed to integrate the modality features for classification. In addition, a self-compiled dataset of 10 actions in the vehicle was established. According to the experimental results, the proposed system can increase the accuracy rate by nearly 30% compared to the original two-stream CNN model. In the future, a comprehensive network configuration could be designed for more fine-grained actions and various scene challenges, such as behavior recognition in the night.

Author Contributions: Methodology, J.-C.C. and C.-Y.L.; Software, C.-Y.L. and P.-Y.H.; Visualization, P.-Y.H.; Writing—original draft preparation, C.-Y.L.; Writing—review & editing, J.-C.C. and C.-R.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Science and Technology of Taiwan, R.O.C., under grant No. 108-2221-E-992-032-.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization (WHO). Global Status Report. Available online: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/ (accessed on 13 February 2019).
2. Eraqi, H.; Abouelnaga, Y.; Saad, M.H.; Moustafa, M. Driver Distraction Identification with an Ensemble of Convolutional Neural Networks. *J. Adv. Transp.* **2019**, *2019*, 1–12. [CrossRef]
3. Lee, J.D. Driving safety. *Rev. Hum. Factor Ergonom.* **2005**, *1*, 172–218. [CrossRef]
4. Traffic Safety Facts. Available online: https://www.nhtsa.gov/risky-driving/distracted-driving (accessed on 13 February 2019).
5. Distracted Driving. 2016. Available online: https://www.cdc.gov/motorvehiclesafety/distracted_driving/ (accessed on 13 February 2019).
6. National Highway Traffic Safety Administration. Traffic Safety Facts. Available online: https://www.nhtsa.gov/risky-driving/distracted-driving (accessed on 13 February 2019).
7. Zhang, X.; Zheng, N.; Wang, F.; He, Y. Visual recognition of driver hand-held cell phone use based on hidden CRF. In Proceedings of the 2011 IEEE International Conference on Vehicular Electronics and Safety, Beijing, China, 10–12 July 2011; pp. 248–251.
8. Berri, R.; Silva, A.G.; Parpinelli, R.S.; Girardi, E.; Arthur, R. A Pattern Recognition System for Detecting Use of Mobile Phones While Driving. In Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, 5–8 January 2014.
9. Craye, C.; Karray, F. Driver distraction detection and recognition using RGB-D sensor. *arXiv* **2015**, arXiv:1502.00250.
10. Seshadri, K.; Juefei-Xu, F.; Pal, D.K.; Savvides, M.; Thor, C.P. Driver cell phone usage detection on Strategic Highway Research Program (SHRP2) face view videos. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 35–43.
11. Das, N.; Ohn-Bar, E.; Trivedi, M.M. On Performance Evaluation of Driver Hand Detection Algorithms: Challenges, Dataset, and Metrics. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Gran Canaria, Spain, 15–18 September 2015; pp. 2953–2958.
12. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Pdf ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
13. Hoang, T.; Do, T.-T.; Le Tan, D.-K.; Cheung, N.-M. Selective Deep Convolutional Features for Image Retrieval. In Proceedings of the 2017 ACM on Web Science Conference—WebSci '17, Troy, NY, USA, 25–28 June 2017.
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
15. Le, T.H.N.; Zheng, Y.; Zhu, C.; Luu, K.; Savvides, M. Multiple Scale Faster-RCNN Approach to Driver's Cell-Phone Usage and Hands on Steering Wheel Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 46–53.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
17. Yuen, K.; Martin, S.; Trivedi, M.M. Looking at faces in a vehicle: A deep CNN based approach and evaluation. In Proceedings of the 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC), Rio de Janeiro, Brazil, 1–4 November 2016.
18. Streiffer, C.; Raghavendra, R.; Benson, T.; Srivatsa, M. Darnet: A deep learning solution for distracted driving detection. In Proceedings of the ACM/IFIP/USENIX Middleware Conference, Las Vegas, NV, USA, 11–15 December 2017; pp. 22–28.

19. Majdi, M.S.; Ram, S.; Gill, J.T.; Rodriguez, J.J. Drive-Net: Convolutional Network for Driver Distraction Detection. In Proceedings of the 2018 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI), Las Vegas, NV, USA, 8–10 April 2018; pp. 1–4.
20. Tran, D.; Do, H.M.; Sheng, W.; Bai, H.; Chowdhary, G. Real-time detection of distracted driving based on deep learning. *IET Intell. Transp. Syst.* **2018**, *12*, 1210–1219. [CrossRef]
21. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, Banff, AB, Canada, 14–16 April 2014.
22. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
23. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the AAAI, Phoenix, AZ, USA, 12–17 February 2016.
24. Abouelnaga, Y.; Eraqi, H.M.; Moustafa, M.N. Real-time distracted driver posture classification. In Proceedings of the Workshop on Machine Learning for Intelligent Transportation Systems, Montréal, QC, Canada, 8 December 2018.
25. Baheti, B.; Gajre, S.; Talbar, S. Detection of Distracted Driver Using Convolutional Neural Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 1032–1038.
26. Masood, S.; Rai, A.; Aggarwal, A.; Doja, M.; Ahmad, M. Detecting distraction of drivers using Convolutional Neural Network. *Pattern Recognit. Lett.* **2018**. [CrossRef]
27. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the NIPS, Montreal, QC, Canada, 8–13 December 2014.
28. Zach, C.; Pock, T.; Bischof, H. A Duality Based Approach for Realtime TV-L 1 Optical Flow. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4713.
29. Nanni, L.; Ghidoni, S.; Brahnam, S. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognit.* **2017**, *71*, 158–172. [CrossRef]
30. Laptev, I.; Lindeberg, T. Velocity adaptation of space-time interest points. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 14–17 October 2003; pp. 432–439.
31. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior Recognition via Sparse Spatio-Temporal Features. In Proceedings of the 2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2006.
32. Sadanand, S.; Corso, J.J. Action bank: A high-level representation of activity in video. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.
33. Wang, H.; Schmid, C. Action Recognition with Improved Trajectories. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013.
34. Kong, Y.; Fu, Y. Human Action Recognition and Prediction: A Survey. *arXiv* **2018**, arXiv:1806.11230.
35. Martin, S.; Ohn-Bar, E.; Tawari, A.; Trivedi, M.M.; Martin, S. Understanding head and hand activities and coordination in naturalistic driving videos. In Proceedings of the 2014 IEEE Intelligent Vehicles Symposium, Dearborn, MI, USA, 8–11 June 2014; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2014; pp. 884–889.
36. Ohn-Bar, E.; Martin, S.; Trivedi, M.M. Driver hand activity analysis in naturalistic driving studies: Challenges, algorithms, and experimental studies. *J. Electron. Imaging* **2013**, *22*, 41119. [CrossRef]
37. Zhao, C.; Zhang, B.; He, J.; Lian, J. Recognition of driving postures by contourlet transform and random forests. *IET Intell. Transp. Syst.* **2012**, *6*, 161–168. [CrossRef]
38. StateFarm's Distracted Driver Detection Dataset. Available online: https://www.kaggle.com/c/state-farm-distracted-driver-detection (accessed on 13 February 2019).
39. Carreira, J.; Zisserman, A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
40. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L.; Shetty, S.; Leung, T. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 24–27 June 2014; pp. 1725–1732.

41. Donahue, J.; Hendricks, L.A.; Guadarrama, S.; Rohrbach, M.; Venugopalan, S.; Darrell, T.; Saenko, K. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.
42. Ng, J.Y.-H.; Hausknecht, M.; Vijayanarasimhan, S.; Vinyals, O.; Monga, R.; Toderici, G. Beyond short snippets: Deep networks for video classification. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4694–4702.
43. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
44. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning Spatiotemporal Features with 3D Convolutional Networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4489–4497.
45. Lin, J.C.W.; Shao, Y.; Zhou, Y.; Pirouz, M.; Chen, H.C. A Bi-LSTM mention hypergraph model with encoding schema for mention extraction. *Eng. Appl. Artif. Intell.* **2019**, *85*, 175–181. [CrossRef]
46. Lin, J.C.W.; Shao, Y.; Fournier-Viger, Y.; Hamido, P.F. BILU-NEMH: A BILU neural-encoded mention hypergraph for mention extraction. *Inf. Sci.* **2019**, *496*, 53–64. [CrossRef]
47. Yan, C.; Coenen, F.; Zhang, B. Driving posture recognition by convolutional neural networks. *IET Comput. Vis.* **2016**, *10*, 103–114. [CrossRef]
48. Chuang, Y.-W.; Kuo, C.-H.; Sun, S.-W.; Chang, P.-C. Driver behavior recognition using recurrent neural network in multiple depth cameras environment. *Electron. Imaging* **2019**, *2019*, 56-1–56-7. [CrossRef]
49. Gebert, P.; Roitberg, A.; Haurilet, M.; Stiefelhagen, R. End-to-end Prediction of Driver Intention using 3D Convolutional Neural Networks. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 969–974.
50. Xu, L.; Fujimura, K. Real-Time Driver Activity Recognition with Random Forests. In Proceedings of the 6th international conference on Multimodal interfaces—ICMI '04, Seattle, WA, USA, 17–19 September 2014; Association for Computing Machinery (ACM): New York, NY, USA, 2014; pp. 1–8.
51. Martin, M.; Popp, J.; Anneken, M.; Voit, M.; Stiefelhagen, R. Body Pose and Context Information for Driver Secondary Task Detection. In Proceedings of the 2018 IEEE Intelligent Vehicles Symposium (IV), Changshu, China, 26–30 June 2018; pp. 2015–2021.
52. Martin, M.; Roiterg, A.; Haurilet, M.; Horne, M.; Reib, S.; Voit, M.; Stiefelhagen, R. Drive & Act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In Proceedings of the International Conference on Computer Vision, Thessaloniki, Greece, 23–25 September 2019.
53. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild. *arXiv* **2012**, arXiv:1212.0402.
54. Soleymani, S.; Dabouei, A.; Kazemi, H.; Dawson, J.; Nasrabadi, N.M. Multi-Level Feature Abstraction from Convolutional Neural Networks for Multimodal Biometric Identification. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3469–3476.