



# Hard Spatial Attention Framework for Driver Action Recognition at Nighttime

Karam Abdullah<sup>1,2</sup>, Imen Jegham<sup>3,4</sup>, Mohamed Ali Mahjoub<sup>3</sup> and Anouar Ben Khalifa<sup>3,5</sup>

<sup>1</sup>Université de Sousse, ISITCom, LATIS-Laboratory of Advanced Technology and Intelligent Systems,  
4011, Sousse, Tunisia

<sup>2</sup>University of Mosul, Collage of Education for Pure Science, Computer Science Department, Mosul, Iraq

<sup>3</sup>Université de Sousse, Ecole Nationale d'Ingénieurs de Sousse, LATIS- Laboratory of Advanced Technology and  
Intelligent Systems, 4023, Sousse, Tunisia

<sup>4</sup>Horizon School of Digital Technologies, 4023, Sousse, Tunisia

<sup>5</sup>Université de Jendouba, Institut National des Technologies et des Sciences du Kef, 7100, Le Kef, Tunisia

**Keywords:** Driver Action Recognition, Driving at Nighttime, Deep Learning, Hard Attention, Spatial Attention.


**Abstract:** Driver monitoring has become a key challenge in both computer vision and intelligent transportation system research fields due to its high potential to save pedestrians, drivers, and passengers' lives. In fact, a variety of issues related to driver action classification in real-world driving settings are present and make classification a challenging task. Recently, driver in-vehicle action relying on deep neural networks has made significant progress. Though promising classification results have been achieved in the daytime, the performance in the nighttime remains far from satisfactory. In addition, deep learning techniques treat the whole input data with the same importance which is confusing. In this work, a nighttime driver action classification network called hard spatial attention is proposed. Our approach effectively captures the relevant dynamic spatial information of the cluttered driving scenes under low illumination for an efficient driver action classification. Experiments are performed on the unique public realistic driver action dataset recorded at nighttime 3MDAD dataset. Our approach outperforms state-of-the-art methods' classification accuracies on both side and front views.


## 1 INTRODUCTION


The World Health Organization (WHO) estimates that over 1.3 million people lose their lives in road accidents yearly, and between 20 and 50 million people are injured as a result (WHO,2022). The main goal of safe system approaches to traffic safety is to make sure that the transportation system is safe for everyone who uses the roads; to make the safety systems more efficient in preserving the lives of drivers, passengers, and pedestrians, one of the important things is analysing the actions of drivers in realistic driving scenarios.


The main goal of Driver Actions Recognition (DAR) is to categorize the driver's normal actions and the various abnormal actions of the driver. There have

been many state-of-art studies in this field, where some of them have worked on images (Hu, 2019) (Koesdwiady, 2016) (Xing, 2019), and others treated the video (Hu, 2018). Distracted driving coupled with low illumination increase the accident probability, whereas, according to a survey conducted by the National Sleep Foundation (NSF), 60% of respondents have admitted they drive when fatigued, and 37% have admitted they doze off behind the wheel. 13% of those people claim that they nod off behind the wheel at least once a month, and 4% report that they have been responsible for an accident because of dozing off behind the wheel (NSC, 2022). The problem of being distracted when driving at nighttime is the primary subject of this research. Taking care of these issues will make it easier to

<sup>a</sup>  <https://orcid.org/0000-0003-2517-873X>

<sup>b</sup>  <https://orcid.org/0000-0003-1531-438X>

<sup>c</sup>  <https://orcid.org/0000-0002-8181-4684>

<sup>d</sup>  <https://orcid.org/0000-0002-9946-0829>

implement various applications, such as automatic driving at nighttime, safety driving systems in cars, and robotics in the near future.

Attention is a human cognitive process that is capable of promoting a selection of a few motives from a huge amount of information that reaches us constantly (da Silva Simões, 2016). Attention can be thought of as an essential component of both perceptual and cognitive processes. In light of the fact that our capacity to process information from multiple sources is limited, attention mechanisms select, modulate, and concentrate on the data that are most pertinent to behaviour (Correia, 2021). Attentional mechanisms consider making networks more powerful and simpler, reduce information bottlenecks caused by long spatial and temporal dependencies, and facilitate multimodality [10]. Thus, we can give a comprehensive definition of attention as a process of preferring a specific part of the data over other parts and whether they are image data.

Attention mechanisms can be broadly classified as belonging to one of the following groups: hard attention which determines whether part of the input will appear in the output or not, soft attention which determines whether a portion of the input appears in re-weighted form in the output, self-attention which is a technique that allows the input parts to interact with each other and determine which ones will be of the highest interest in the output by measuring the interconnection between the parts. There are also secondary categories as: global attention which is an approach that takes the traditional soft attention and makes it easier to implement within encoder-decoder frameworks, local attention which is a balance between soft attention and hard attention, co-attention which shifts the focus both from the context to the target and back to the context from the target, and hierarchical attention which provides mechanisms that have been adapted to engage with hierarchical levels at various levels of granularity (Santana, 2021). Visual attention can be clustered into two main categorized types of attention: spatial attention that focus on where concentration should be, and temporal attention which determines when paying attention should be (Guo, 2022). To improve overall performance, there is a hybrid method called spatial-temporal attention which was also proposed in (Du, 2017).

In this paper, we put forward a novel hard spatial attention mechanism to explain the behaviours of nighttime driving. This mechanism is based on the ability of humans to interpret visual sequences with high efficiency by focusing their attention exclusively on the most significant driver action information. We

develop a mechanism that uses image processing techniques to separate the driver's human body from the rest of the in-vehicle environment at nighttime. This is accomplished by creating a mask through which the driver's movements are extracted. This enables the model to focus on the most prominent parts of the driver's human body feature maps for spatial attention in the system. On the other hand, we use a hybrid network for classification and for getting the final decision.

Our main contributions can be listed as follows:

- We design a novel hard spatial attention model that captures the dynamic spatial dependency by focusing attention on related information present on infrared images.
- We develop a mechanism to separate the driver's body from the background by a piece-wise linear transformation function and morphological processes and the foreground shape dropped onto the original infrared image to obtain unique objects independently, without losing any pertinent information or any smallest visual details.
- Experiments on 3MDAD at nighttime benchmark datasets prove that the suggested method is efficient compared with state-of-the-art approaches.

## 2 RELATED WORKS

Several research are currently conducted on deep learning to recognize driver action systems (Jegham, 2020) (Hu, 2020) (Abdullah, 2022) (Hu, 2021). This is because deep learning can extract a large number of features quickly, with high accuracy, and in record time, compared to traditional methods (Kong, 2022) (Wang, 2019). There are many algorithms that use deep learning techniques by merging the depth shape with RGB visuals. Jegham et al. (Jegham, 2020) proposed a driver action detection system using a soft spatial attention network that was based on depth information. They suggested a depth-based spatial attention network to direct attention on person silhouettes, which would allow it to accurately identify the movements of the driver. Alotaibi et al. (Alotaibi, 2020), offered a strategy that improved the performance of detecting a distracted driver's behaviour by combining three of the most sophisticated models in deep learning. These models included the residual network, the inception module and the hierarchical recurrent neural network. In (Tran, 2018) Tran et al. proposed a system for the detection of distractions based on various deep learning architectures. They gathered a dataset that

included images of drivers in both their typical driving postures and in postures indicating that they were preoccupied. An embedded graphics processing unit platform was used to develop and test four deep convolutional neural networks, including VGG-16, AlexNet, GoogleNet, and the residual network. Kopuklu et al. (Kopuklu, 2021) suggested a contrastive learning strategy for acquiring a measure to distinguish between typical and unusual driving by using deep learning techniques. They employed ResNet-18 as the base encoder to determine the baseline results based on spatial attention. Previous work mainly dealt with driving distraction during daytime. This was due to the difficulty of obtaining a nighttime database. It was also difficult to interpret the very dark nighttime images because of lack of lighting and the type of noise that accompanied the nighttime image. Abdullah et al. (Abdullah, 2022) was the first to work on DAR at nighttime. They proposed multi-convolutional streams for a hybrid network that would effectively fuse multimodal data to efficiently categories drivers' activities in low visibility and a crowded driving scenario. We propose a novel hard spatial attention framework to separate the driver's body from the surrounding environment at nighttime by creating a mask through which the driver's actions are extracted. Then a hybrid network is used for classification and to getting the final decision.

### 3 PROPOSED APPROACH

In this work, we develop a novel Hard Spatial Attention (HSA) architecture, which can recognize driver activities during the nighttime inside a vehicle. This framework is shown in Figure 1 and Algorithm 1.

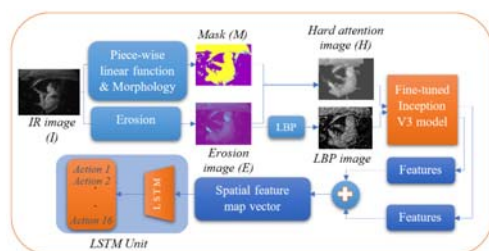


Figure 1: Proposed architecture of HSA.

---

Algorithm 1: Hard Spatial Attention Framework.

---

Require: IR test sequence:  $\{I_i, r_1, s_1, r_2, s_2\}$

Ensure: Classification result label:  $y_i$

```

1: for  $i \leftarrow 0 \dots t$  do

```

2:  $C_j \leftarrow$ 

Piecewise\_liner\_transformation ( $I_j, r_1, s_1, r_2, s_2$ );

```

3:    $M_i$ 
 $\leftarrow$  Morphological_operation(Dilation&Closing) ( $C_i$ );
4:    $E_i \leftarrow$  Morphological_operation(Erosion) ( $r_i$ );
5:    $H_i \leftarrow M_i \times E_i$ ;
6:    $L_i \leftarrow$  LBP ( $E_i$ );
7:   Features_ $L_i \leftarrow$  CNN ( $L_i$ );
8:   Features_ $H_i \leftarrow$  CNN ( $H_i$ );
9:   Spatial_Features_Map
       $\leftarrow$  (Features_ $L_i$  + Features_ $H_i$ );
10: end for
11:  $y_j \leftarrow$  LSTM_Classification (Spatial_Features_Map);
12: return  $y_j$ 

```

### 3.1 Hard Spatial Attention

Each image is driven through two series of image processors to produce a hard-attention image and a Local Binary Pattern (LBP)-image simultaneously. The two results of images pass through the fine-tuned Inception V3 model. The main idea behind the hard attention image is to create a mask and then drop it on the original image to focus on a most informative specific region, which is a main step in the spatial attention mechanism.

The mask is obtained by two steps: the first step is the center of interest because we work at nighttime, so the images suffer from low illumination, which causes low contrast images. Therefore, to get contrast stretching we use the piece-wise linear transformation function, which is not completely linear in its behaviour. On the other hand, although it is linear between specific intervals, contrast stretching is one of the most often utilized transformation functions. It effectively lowers the intensity of dark pixels while raising it for the light pixels, resulting in a stretching of the intensity levels. This step is illustrated by Figure 2 (GeeksforGeeks, 2019)(Gonzalez RC, 2017), where  $(r_1, s_1)$  and  $(r_2, s_2)$  are the two points.

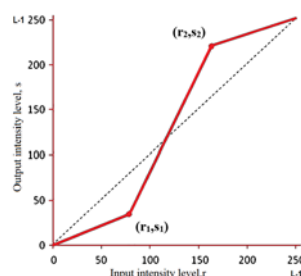


Figure 2: Piece-wise linear transformation function (Gonzalez RC, 2017).

This function makes the intensity values of the dark portion darker and the light portion lighter. It expands the range of the intensity in the image. Summing the number of segment lengths of discrete

linear functions can be expressed by Equation (1) (Gonzalez RC, 2017).

$$Y = \sum_{i=0}^n \sqrt{1 + \left(\frac{\Delta r_i}{\Delta s_i}\right)^2} \Delta s_i \quad (1)$$

The second step is done by using the morphological operation by applying dilation to expand the image pixels and fill the hole in the image. then a close operation to remove or close smallest objects. Figure 3 illustrates the steps of creating a mask.

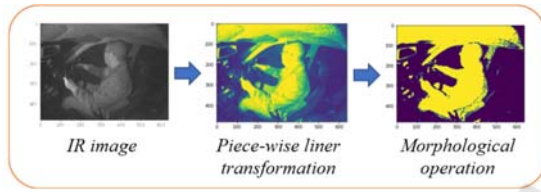


Figure 3: Example of creating mask steps.

After creating the mask, we drop it on the erosion image to get a hard attention image. Thereby,  $H_i$  can be expressed by applying Equation (2), where  $M_i$  is the mask image and  $E_i$  is erosion image.

$$H_i = \sum M_i \times E_i \quad (2)$$

### 3.2 Feature Extraction & Classification

LBP-images are obtained from the descriptor (LBP) to get hand craft features that will add more visual features to the recognition system. LBP is a form of a visual descriptor used in computer vision to classify objects. It can effectively summarize the local information of images, by labelling the pixels of an image with thresholding the pixels in the vicinity of each pixel and treating the result as a binary number (Huang,2011).

All images obtained from the hard attention and the LBP descriptor are simultaneously entered into a pre-trained Inception V3 network to extract 2048 features per frame from the hard attention line and 2048 features per frame from the LBP descriptor, and then combine them together to produce a spatial feature map vector, giving a clear understanding of the image content, Thus, it clearly describes the driver's actions inside the car cabin while driving at nighttime. It is known that most safety systems and modern human-action monitoring systems depend on deep learning in interpreting events and actions because of their high accuracy. One of the best

algorithms used in deep learning is the Inception V3 algorithm, which consists of 48 convolutional layers. It is based on the principle of the inception layer, with the basic concept of a sparse connected architecture. The inception layer contains  $(1 \times 1)$ ,  $(3 \times 3)$  and  $(5 \times 5)$  convolution layers and group together their result filter banks into a unique output vector that becomes the feed of the following stage (Szegedy, 2016), as shown in Figure 4.

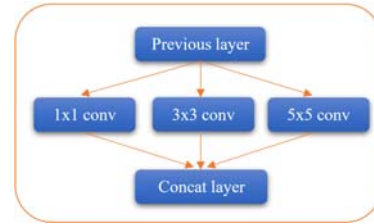


Figure 4: The sparsely connected layer of general inception architecture (Szegedy, 2016).

The presence of different sizes of filters in a particular layer trained within a model helps to find different sized parts within the image. For example, the parts of the image of the driver's actions inside the cockpit have different sizes. The next network layer will be trained on the overall object in an image, which allows the internal layers to select the right filter size to pick up the suitable information in the image. It is possible that the size of the human head in the first image is large, so we must use a large filter, while in the second image it is small, so we will use a smaller filter.

Figure 5 illustrates the inception module architecture that uses dimensionality reduction by  $1 \times 1$  convolution processes to reduce the cost of computing before more expensive  $3 \times 3$  and  $5 \times 5$  convolutions (Szegedy, 2015). It is possible to construct an Inception V3 network by stacking together the modules that are similar to the ones illustrated above, with the addition of some max pooling layers with stride "2" to reduce the resolution of the grid by half. This architecture has the advantage of allowing for considerable increment in the number of units for each step without causing an uncontrolled explosion in computational costs at later levels, which is an important feature. This is achieved by the ubiquitous use of dimensionality reduction prior to expensive convolutions with larger patch sizes.

The loss function in Equation (3) is utilized in multi-class classification endeavors, known as a categorical cross entropy. Calculating the loss of an example is done via the categorical cross entropy loss function, which does so by computing the following expression (Knowledge Center, 2022).



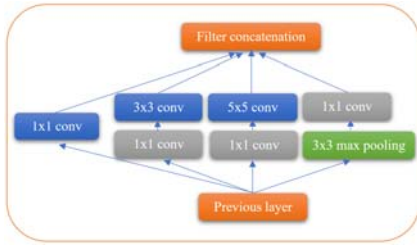


Figure 5: Inception module dimension reduction (Szegedy, 2015).

$$Loss = - \sum_{i=1}^s y_i \cdot \log \hat{y}_i \quad (3)$$

where  $s$  is the number of scalar values that are generated by the model,  $\hat{y}_i$  is the scalar value in the model output and  $y_i$  is the value that corresponds to the target. Because of the minus sign, the loss will be reduced when the distributions are brought into closer proximity with one another.

The Long-Short Time Memory (LSTM) (Hochreiter, 1997) network and Gated Recurrent Units (GRU) (Chung, 2014) are a special type of Recurrent Neural Networks (RNN). They are a robust and effective sort of neural networks, and they are among the most interesting algorithms now in use due to the fact that they are the only ones that have an internal storage. Both models have a forget gate, but the GRU lacks an output gate, so it has fewer parameters. It is used for certain tasks, and it performs better on small datasets and less frequent data (Gruber, 2020). LSTM can learn long-range information and has been very successful and has been widely used. LSTM is relied upon to conduct the process of classifying the driver action, due to the fact that past events are used to infer subsequent events. This network includes the time factor, which is very necessary to understand the actions during a certain period of time. The network consists of four gates, which are dependent on the network in their operation, and which assist the network in remembering the most critical information, resulting in a significant improvement in the output quality. The LSTM formula can be expressed by Equations (4-9) (Ullah, 2017), where  $I$  is the input gate,  $W$  is the weight matrices,  $L$  is the incoming vector,  $b$  is the bias vector,  $F$  is the forgetting gate,  $O$  is the output gate,  $g$  is input node,  $C$  is the cell state,  $H$  is the short term memory and  $\odot$  is dot product operation.

$$I_t = \sigma(W^i [H_{t-1}, L_t] + b_i) \quad (4)$$

$$F_t = \sigma(W^f [H_{t-1}, L_t] + b_f) \quad (5)$$

$$O_t = \sigma(W^o [H_{t-1}, L_t] + b_o) \quad (6)$$

$$g_t = \tanh(W^g [H_{t-1} + L_t] + b_g) \quad (7)$$

$$C_t = C_{t-1} \odot F_t + g_t \odot I_t \quad (8)$$

$$H_t = \tanh(C_t) \odot O_t \quad (9)$$

For driver action recognition tasks, the performance metric is the classification accuracy as in the Equation (10):

$$Accuracy = \frac{TP + TN}{Total\ predictions} \quad (10)$$

where TP are true positives, TN is true negatives and total number of predictions is the total number of predictions made by the model (TP + TN + False Positives + False Negatives).

## 4 EXPERIMENTS AND RESULTS

To highlight the utility of our proposed HSA, we evaluate our approach on different views of realistic sequences collected at nighttime.

### 4.1 Dataset

The new distraction dataset called the Multiview, Multimodal, and Multispectral Driver Action Dataset 3MDAD (at nighttime) (Jegham, 2020) is used for the evaluation of HSA performance. To the best of our knowledge, this dataset is the first real-world dataset gathered from environments at nighttime in addition to daytime. The data are recorded from the side and front views at nighttime. It displays 19 drivers who are given instructions to carry out 16 actions. The first action is: “safe driving”, and 15 other distracting secondary actions are: “doing hair and makeup”, “adjusting radio”, “GPS operating”, “writing a message using the right hand”, “writing a message using left hand”, “talking phone using right hand”, “talking phone using left hand”, “having picture”, “talking to the passenger”, “singing or dancing”, “fatigue and somnolence”, “drinking using right hand”, “drinking using left hand”, “reaching behind” and “smoking”. These actions are respectively referred to as A1- A16. A total of 130,028 frames are included in this dataset.

### 4.2 Experimental Setup

The experiments are performed using an ASUS laptop model TUF F15 equipped with a core i7-11370H Intel processor, 40 GB of DDR4 RAM, and a clock speed of 3200 MHz. Windows 10 serves as a

primary operating system. We rely on libraries such as TensorFlow 2.6.0 and Keras 2.6.0, to run the code developed in Python 3.7 using the Spyder IDE 4.2.5. In order to complete the code, a massive Nvidia RTX 3070 graphics card equipped with 8 GB DDR6 is utilized. Through the use of the Adam optimizer. The 3MDAD dataset is divided into three parts, the training data make up 70% of the whole dataset, while the validation data are 10% and the test data are 20%.

### 4.3 Evaluation of HSA network

Many practical experiments are conducted to make an outstanding evaluation of our HSA network.

#### 4.3.1 Quantitative Results

Our comparison is conducted, on the one hand, using hybrid network-based methods called LRCN (Jegham, 2021) without any attention mechanism and using different pre-trained CNN architectures including Inception v3, VGG19 and VGG16. Our previous proposed spatial attention-based method MCSH (Abdullah, 2022) is used, on the other hand. The obtained results are depicted in Figure 6 for side view and in Figure 7 for front view. We notice that our proposed HSA outperforms the-state-of-the-art methods for the two views in terms of accuracy.

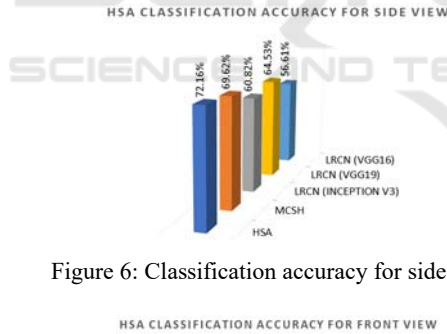


Figure 6: Classification accuracy for side view.

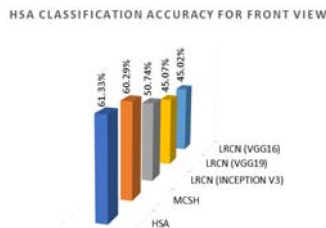


Figure 7: Classification accuracy for front view.

When compared to its rival, the HSA classification accuracy statistics demonstrate a significant improvement. In point of fact, compared to other methods, the accuracy of the HSA recognition system has increased by 3% to 25% in side view and by 1% to 15% in front view. It is important to note that the

side view is always more accurate than the front view. This is because of the occlusion present in the front view. The accuracy has reached more than 72 % in the side view and more than 61 % in the front view. These are very significant percentages in driver action classification at night.

#### 4.3.2 Qualitative Results

When comparing our proposed HSA to other deep learning techniques, it achieves very promising results. This is due to its reliance on highly sensitive spatial attention for night images. It is noticeable that the proposed algorithm provides remarkable results through the confusion matrices. Figure 8 and Figure 9 illustrate the confusion matrices of our proposed HSA, whereas Figure 10 and Figure 11 depict the confusion matrices of the best deep learning method without attention block LRCN (Inception V3). According to Figure 8 and Figure 9, general improvement is noticed. Some actions achieve a 100% discrimination, for example A2 and A7 in side view and A2 in front view. In addition, for the side view, the A5 and A14 have more than 80% classification accuracy and the rest of actions range between 52% and 79%. For the front view no action recorded less than 51% accuracy despite the wide range of naturalistic driving issues, principally the high interclass similarity and interclass variability present. Finally, as it is clear, there was a significant

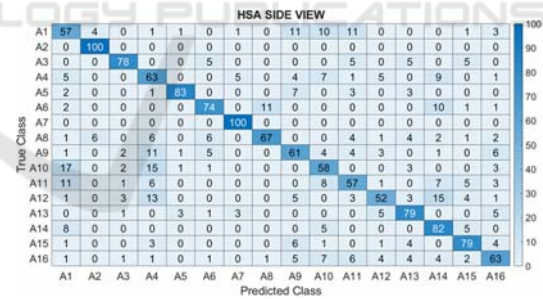


Figure 8: Confusion matrix of HSA for side view.

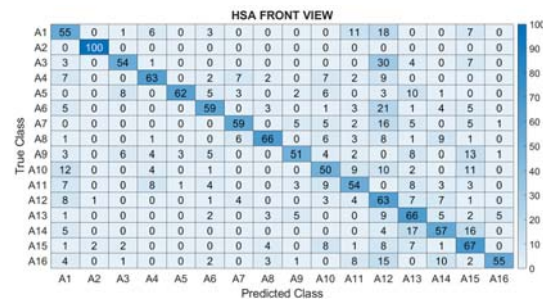


Figure 9: Confusion matrix of HSA for front view.

improvement in most of the action in the side view like A1, A12 and A16, as well as in the frontal view also; there was a significant improvement in the A1, A9 and A16.

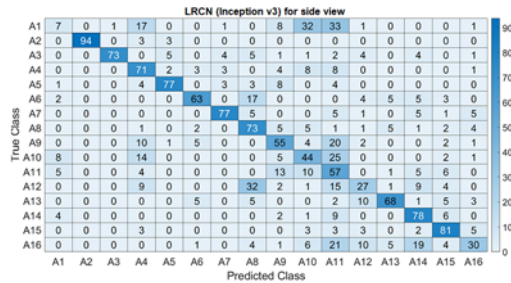


Figure 10: Confusion matrix of LRCN (Inception V3) network for side view.

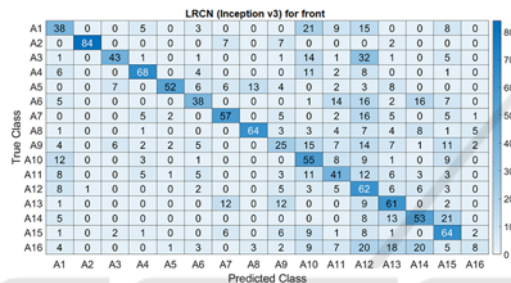


Figure 11: Confusion matrix of LRCN (Inception V3) network for front view.

## 5 DISCUSSIONS

Our proposed algorithm provides high performance and outstanding accuracy results due to the correct interpretation of nighttime images by adopting spatial attention and using a hybrid deep learning algorithm. Because of low illumination, image quality is bad. Thus, we must use a function that is not completely linear in its behaviour as a piece-wise linear transformation function with a morphological operation to expand image pixels and fill the hole in the image and then remove or close the smallest objects to create a mask that will be used to perform a hard cut mechanism. Thus, we obtain rich, focused information that can be useful in issuing the decision and shortening a significant amount of useless data as a black area, thus shortening the time, which significantly speeds up the implementation process. We perform the HSA on a multiview dataset (side and front views). In order to ensure the best accuracy and reliability of our proposed method, we obtained more than 72% accuracy, which is the highest percentage

that has been achieved so far in the literature in this field in a realistic and uncontrolled environment.

## 6 CONCLUSIONS

In this paper, we propose a novel hard spatial attention network for driver action recognition at nighttime by performing a projection of the mask performed by a series of advanced image processors and using an LBP descriptor to reach a high level of accuracy. In fact, the described process adds a cognitive mechanism to a hybrid network based on Inception V3 to mainly focus on relevant information in the driving scene. The proposed approach achieved our contribution that was mentioned in the introduction by improving classification accuracy on the 3MDAD dataset by up to 72% in the side view and 61% in the front view.

## REFERENCES

- Abdullah, K., Jegham, I., Khalifa, A. B., & Mahjoub, M. A. (2022, May). A Multi-Convolutional Stream for Hybrid network for Driver Action Recognition at Nighttime. In 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT) (Vol. 1, pp. 337-342). IEEE. DOI: 10.1109/CoDIT55151.2022.9804013.
- Alotaibi, M., & Alotaibi, B. (2020). Distracted driver classification using deep learning. *Signal, Image and Video Processing*, 14(3), 617-624. <https://doi.org/10.1007/s11760-019-01589-z>
- Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*. <https://doi.org/10.48550/arXiv.1412.3555>
- Correia, A. D. S., & Colomhini, E. L. (2021). Attention, please! a survey of neural attention models in deep learning. *arXiv preprint arXiv:2103.16775*. <https://doi.org/10.48550/arXiv.2103.16775>
- da Silva Simões, A., Colomhini, E. L., & Ribeiro, C. H. C. (2016). CONAIM: A conscious attention-based integrated model for human-like robots. *IEEE Systems Journal*, 11(3), 1296-1307. DOI: 10.1109/JSYST.2015.2498542
- Du, W., Wang, Y., & Qiao, Y. (2017). Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Transactions on Image Processing*, 27(3), 1347-1360. DOI: 10.1109/TIP.2017.2778563
- GeeksforGeeks, "Piecewise-Linear Transformation Functions," 2019, last accessed 1/08/2022. [Online]. Available: <https://www.geeksforgeeks.org>

- Gonzalez RC, Woods RE. Digital Image Processing. 4th edition. New York, NY: Pearson; 2017. <http://14.139.186.253/2354/1/67747.pdf>
- Gruber, N., & Jockisch, A. (2020). Are GRU cells more specific and LSTM cells more sensitive in motive classification of text?. *Frontiers in artificial intelligence*, 3, 40. <https://doi.org/10.3389/frai.2020.00040>
- Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 1-38. <https://doi.org/10.1007/s41095-022-0271-y>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780. DOI: 10.1162/neco.1997.9.8.1735
- Hu, Y., Lu, M., & Lu, X. (2018, November). Spatial-temporal fusion convolutional neural network for simulated driving behavior recognition. In 2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV) (pp. 1271-1277). IEEE. DOI: 10.1109/ICARCV.2018.8581201
- Hu, Y., Lu, M., & Lu, X. (2019). Driving behaviour recognition from still images by using multi-stream fusion CNN. *Machine Vision and Applications*, 30(5), 851-865. <https://doi.org/10.1007/s00138-018-0994-z>
- Hu, Y., Lu, M., & Lu, X. (2020). Feature refinement for image-based driver action recognition via multi-scale attention convolutional neural network. *Signal Processing: Image Communication*, 81, 115697. <https://doi.org/10.1016/j.image.2019.115697>
- Hu, Y., Lu, M., Xie, C., & Lu, X. (2021). Video-based driver action recognition via hybrid spatial-temporal deep learning framework. *Multimedia Systems*, 27(3), 483-501. <https://doi.org/10.1007/s00530-020-00724-y>
- Huang, D., Shan, C., Ardabilian, M., Wang, Y., & Chen, L. (2011). Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6), 765-781. DOI: 10.1109/TSMCC.2011.2118750
- Jegham, I., Khalifa, A. B., Alouani, I., & Mahjoub, M. A. (2020). Soft spatial attention-based multimodal driver action recognition using deep learning. *IEEE Sensors Journal*, 21(2), 1918-1925. doi: 10.1109/JSEN.2020.3019258.
- Jegham, I., Khalifa, A. B., Alouani, I., & Mahjoub, M. A. (2020). A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3MDAD. *Signal Processing: Image Communication*, 88, 115960. doi: 10.1016/j.image.2020.115960.
- Knowledge Center, "Categorical crossentropy" 2022, last accessed 1/08/2022. [Online]. Available: <https://peltarion.com/knowledge-center/documentation/modeling-view/build-an-ai-model/loss-functions/categorical-crossentropy>
- Koesdwiady, A., Soua, R., Karray, F., & Kamel, M. S. (2016). Recent trends in driver safety monitoring systems: State of the art and challenges. *IEEE transactions on vehicular technology*, 66(6), 4550-4563. DOI: 10.1109/TVT.2016.2631604
- Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5), 1366-1401. DOI <https://doi.org/10.1007/s11263-022-01594-9>
- Kopuklu, O., Zheng, J., Xu, H., & Rigoll, G. (2021). Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 91-100). <https://doi.org/10.48550/arXiv.2009.14660>
- NSC, "The Most Dangerous Time to Drive", 2022, last accessed 26/08/2022 [Online]. Available: <https://www.nsc.org/road/safety-topics/driving-at-night>.
- Santana, A., & Colombini, E. (2021). Neural Attention Models in Deep Learning: Survey and Taxonomy. *arXiv preprint arXiv:2112.05909*. <https://doi.org/10.48550/arXiv.2112.05909>
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9). [foundation.org/openaccess/content\\_cvpr\\_2015/papers/Szegedy\\_Going\\_Deeper\\_With\\_2015\\_CVPR\\_paper.pdf](https://openaccess.thecvf.com/content_cvpr_2015/papers/Szegedy_Going_Deeper_With_2015_CVPR_paper.pdf)
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818-2826).
- Tran, D., Manh Do, H., Sheng, W., Bai, H., & Chowdhary, G. (2018). Real - time detection of distracted driving based on deep learning. *IET Intelligent Transport Systems*, 12(10), 1210-1219. <https://doi.org/10.1049/iet-its.2018.5172>
- Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE access*, 6, 1155-1166. DOI: 10.1109/ACCESS.2017.2778011
- Wang, K., Chen, X., & Gao, R. (2019, December). Dangerous driving behavior detection with attention mechanism. In *Proceedings of the 3rd International Conference on Video and Image Processing* (pp. 57-62). <https://doi.org/10.1145/3376067.3376101>
- WHO, "Road traffic injuries," 2022, last accessed 1/08/2022 [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
- Xing, Y., Lv, C., Wang, H., Cao, D., Velenis, E., & Wang, F. Y. (2019). Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE transactions on Vehicular Technology*, 68(6), 5379-5390. DOI: 10.1109/TVT.2019.2908425