Detecting abnormal movement of driver's head based on spatial-temporal features of video using deep neural network DNN

Noor D. Al-Shakarchy[1], Israa Hadi Ali[2]
[1,2]College of Information Technology, University of Babylon, Iraq
[1]College of Science, University of Kerbala, Iraq

Article Info

Article history:

Received Jul 28, 2019
Revised Nov 20, 2019
Accepted Dec 26, 2019

Keywords:

3D CNN
Abnormal head Predictor
Deep Neural network (DNN)
Images
Video

ABSTRACT

The development of tracking and surveillance devices makes extracting useful information efficiently. Head tracking is an efficient method to obtain then analyze trajectory data and make a decision based on the spatiotemporal information of videos. Many applications are based on head tracking such as diseases some diagnosis, the gestures languages, and drowsiness detection and so on. Abnormal head movement detection can be achieved using spatial information based on a single image (one frame) at a time without considering the temporal information over time. In this paper, a new method based on multi-images is proposed to track head in order to detect abnormal head movement depending on spatiotemporal Feature using Deep Neural Network DNN that employed the 3-Dimensional Convolution Neural Networks 3D CNN. The proposed method extracts the spatial information as well as the temporal information available in a video then analysis this information to make the decision based on time series (sequences of frames); these time series provides the tracking to the head over time to make the decision. The new dataset created and gathered to implement with the proposed system and called Normal Abnormal Head Movement Dataset (NAHM) video dataset. The new dataset provides different subjects with different conditions that give more efficiency in the implementation of the proposed system. The accuracy of the training set that achieves by proposed system reach to 88% and of validation set reaches to 86%. The values of loss function reach to 0.3 for training set and 0.4 for the validation set.

Corresponding Author:

Noor D. Al-Shakarchy,
College of Information Technology,
University of Babylon,
51002, Babil, Iraq.
Email: noor.d@uokerbala.edu.iq

1. INTRODUCTION

Various applications in computer vision, gestures language, and diseases diagnosis take the head movement nature as a major role in this process. Driver Drowsiness Detection System (DDDS) is a collection of image processing and machine learning techniques to analyze, track, and classify video data in order to make predictions about drowsiness status. DDDS involves extracting salient and important features and then tracking these features to make an accurate classification about drowsiness status. Videos consist of spatial-temporal information which is presented as temporal features over the consecutive frames as well as the spatial features represented by each frame (still image) [1, 2] and the obtained trajectory from head tracking in a video is formed from a sequence of pixels in consecutive frames [3, 4]. Abnormal head movement detection in the video can be considered as a video classification problem based on Feature Extraction. Video classification methods can be categorized into two types: spatial stamp methods; where the decision makes based on spatial feature of one frame at a time; and time stamp methods based on a sequence of frames at a time to represent the dynamic changes in the features over the time called spatiotemporal feature [5, 6].

The robust and efficient classification methods depend on extracting salient and efficient spatiotemporal features that can be used in decision making as well as the possibility to deal with changing in luminance and resolution [7, 8]. Deep Neural networks and Deep learning presented an efficient approach to extract salient features [9] such as Convolution Neural Networks (CNN). CNN can extracts the spatial features as well as the temporal features (represented the tracking of the spatial features) simultaneously in convolution layers which play the role of spatiotemporal feature extractor by applying filters over all the frame sequence of video at a time [10].

In the implementation of video classification, a three Dimension Convolutional Neural Networks (3D CNN) take into account temporal information of video as well as spatial information of frames simultaneously. The two Dimension methods deal with spatial information based on the frames only and ignore the temporal features of the video; one frame is classified independently at a time [1, 11]. That's mean with compared to 3D CNN, the 3D CNN takes into account the dynamic changes in features over the time for each frame with next frame. This temporal information is represented by a sequence of frames over time which provide tracking to the head in a video that can use in classification [12, 13].

In 3D convolution the spatial relationships of objects encodes objects in the 3D space. The 3D filter move and slides in all 3-direction (height, width, channel of the image) and the output numbers are arranged in a 3D space also. The 3D convolution operation represents in Figure 1.

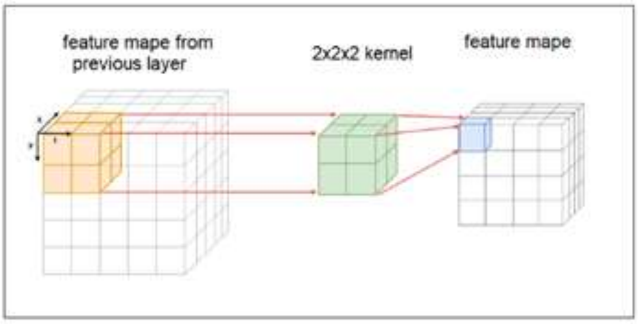

Figure 1. The 3D convolution operation in convolution layer

The lack in video datasets that concerned with head movement classification leads to create a new dataset to implement with the proposed system called Normal Abnormal Head Movement Dataset (NAHM) video dataset. The NAHM dataset contains videos of different subjects (volunteers) with different conditions to give all states of publically head movement and prevent the overfitting and under-fitting problem that may appear.

Some researches detected the head pose based on the "Appearance Template Models" such as J. Foytik et al. [14], V. N. Balasubramanian. [15] and J. Sherrah et al. [16] which saved a set of exemplar images, then the new unseen image is compared directly with these exemplar images and the head pose is detected with a most similar exemplar. This method is easy implementation but on the other side, it suffers from time-consuming and this method detecting only discrete poses.

Other researches employed the machine learning by depending on "Detector arrays" to detect the head pose such as E. Osuna et al. [17], M. Jones et al. [18] which trains multiple face detectors on diverse poses of the head. The head pose of a new unseen image can be defined based on the largest support of the detector. This method is more robust against the variations in appearance but it's poorer for training two classifiers on a similar pose.

M. S. L. Khan et al. [19] presented Geometric Methods to estimate the head pose by measuring the distances between facial features and deviation from bilateral symmetry. This method fails with the parts occluded such as when the eye occluded by glasses because the performance of this method depends on localization the facial features in an accurate manner.

A. Kumar et al. [20] estimate the head pose by using neural networks based on modified "Google Net" architecture to predict facial key points and then joint these key points to predict the head pose. The accuracy of this method depends on the right facial key point's prediction and fail with partial or total occluded.

J. G. X. Y. S. De et al. [21] used two combined networks a VGG network and a recurrent neural network to estimate the Euler angles of the head pose in a video. This method provides an improved pose prediction by leveraging the time dimension that takes into account a sequence of frames in pose prediction. This method provided a promising method for a video to estimate the head pose directly and classify this pose indirectly based on the Euler angles as well as it suffers from time-consuming.

Ines Sophia Rieger [22] implemented a residual network (ResNet) and LeNet-5 to improve the estimation of head pose. This method used spatial information for prediction without taking the time dimensions. The abnormal head movement cannot be detected directly but it computes from the estimated values of head pose. The proposed system presents an architecture of 3D CNN which directly predict the abnormal head movement in a video. In other words, the decision making of head movement produced directly without dependent on angles or any post-process.

In this paper, the DDDS make prediction based on abnormal head movement pose classification predictor that is predicted using Head Movement Status Deep Learning Neural Network (HMSDLNN) Model. The previous DDDS is based on the closed eye classification predictor presented in a previous published paper such that the closed eye is predicted by using the Eye Status Deep Learning Neural Network (ESDLNN) Model. The core of that work is greatly depending on extracted the face then the region of eyes, therefore if the face not extracted for any reason, the system failed to predict the eye's status. The contribution of this paper is to design and implemented a computer vision system to detect and classify the head movements in videos to abnormal and normal classes directly by considering the temporal information as well as the spacial information simultaneously in order to implemented for drowsiness detection. The head movement classification in the video is decided directly by implementing 3-Dimensional Convolutional Neural Networks (3D CNN) with the proposed architecture. The proposed system presented consists of three main parts: the preprocessing stage, the feature extraction stage achieved by employing 3D CNN, and finally decision making of classification.

The organization of this paper illustrates as follows. Section 2 illustrates 3 Dimension Convolution Neural Networks (3D CNN). Section 3 illustrates related works with the proposed system. Section 4 presented the proposed system, its architecture, its main stages and the inner steps of each stage. The experimental results presented in Section 5 and finally the conclusions obtained from this work present in Section 6.

2. RESEARCH METHOD

In this paper, a proposed head tracking model for abnormal head movement detecting based on a spatiotemporal feature using Deep Neural Network DNN is presented. This model design with proposed architecture of 3D CNN and trains and tests on created video dataset named Normal Abnormal Head Movement Dataset (NAHM) which achieve reasonable and efficient accuracy. The general stages and the main steps for each stage of the proposed system are illustrated in a block diagram of Figure 2. The general stages are categorized to two main stages: the pre-processing stage which responsible for preparing video data and extracting the Region of Interest ROI; which is the area around the face; to present it to the second stage. The second stage is abnormal movement detecting stage which presents in a Deep Neural Network system for spatiotemporal feature extraction step and decision making step. Which can be illustrated in Figure 3.

In the pre-processing stage, many functions run such as convert the video to corresponding frames, extract ROI around face area in each frame of a video, create samples depend on time series to represent the temporal dimension to support head tracking. In abnormal head movement detecting stage a 3 Dimension of Deep Neural Network model design and use to extract salient spatiotemporal features from the corresponding sequence of frames. At last these extracted features are used to recognize abnormal head movement.

As shown in Figure 3 above the number of filters (kernals) used in convolution layers are 32, 64, 128 and 128 respectevely. The increased in number of filters based on the concept that hierarchical features used in CNN processing pipeline. In lower layers the features are primitive while in upper layers those are made from combinations of lower-level features and are be high-level abstract features. Certain attributes of the input are measured from the kernels responses in the first hidden layer and in the second hidden needs to have an even larger number of kernels to properly measure the now richer projection of the input through hidden layer 1 and so on to avoide a serious bottleneck for upper layers. To be able to encode the increasingly richer and richer representations, the number of filters is incremented with the moves up the representational hierarchy due to avoid the bottleneck effect.

The sample size that used in first convolution layer is (100, 100, 8) which is pooled to (50, 50, 8) after first pooling layer with spatio-temporal extent (2, 2, 1) and spatio-temporal strid (2, 2, 1) to cant reduction the temporal features. The other pooling layers used spatio-temporal extent (2, 2, 2) and spatio-temporal strid (2, 2, 2) which reduct the spatio- temporal features togathers.

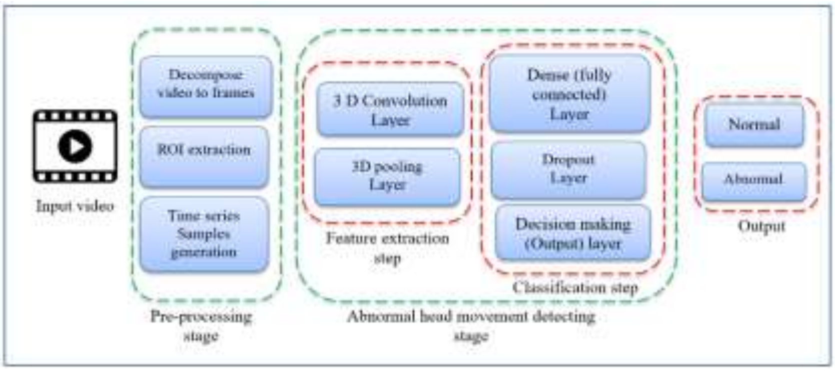

Figure 2. Block diagram of the proposed system

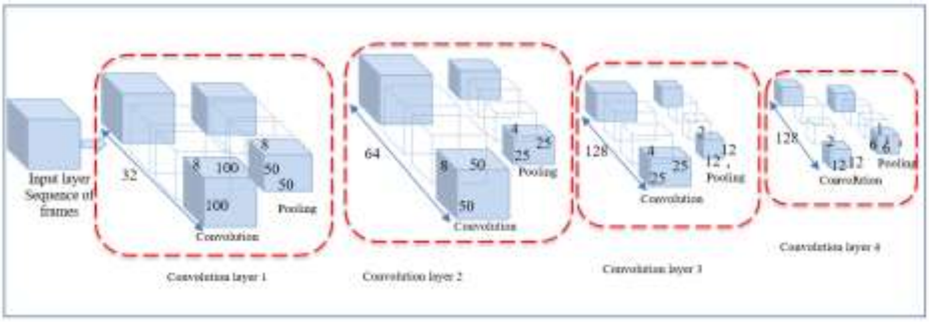

Figure 3. Proposed 3D CNN architecture for feature extraction

2.1. Dataset

The proposed model based on a new architecture of 3D CNN which is training and testing on created data set named Normal Abnormal Head Movement Dataset (NAHM). The total dataset consists of 20 subjects of both male and female acts two different scenarios of normal and abnormal head movements. The scenarios contain videos with barefaced (no Glasses) and Glasses and the videos in these scenarios are 30 frames per second. The total dataset divided into two sub-datasets; training dataset represented 85% from the original data set and testing dataset with 15% used for testing the system on unseen images. The training dataset also divided into actual training dataset represented 80% of the training dataset, and validation dataset represented 20% of the training dataset. To preparing dataset, each video converts to some samples represented sequences of frames with 25% overlapping. The overlapping used to increase the size of dataset and feeding to the model the movement at every moment.

2.2 Proposed system stages

2.2.1. Pre-processing stage

The first stage in the proposed system named pre-processing; all preparing works are presented on the input video to be suitable for implementing on proposed 3D CNN architecture. The core of the preprocessing stage is to perform a regular video segmentation based on time series by converting each video to some samples. Each sample represented sequences of frames with 25%overlapping. The overlapping used to feed to the model the movement at every moment. Gather the frames in a sequence of samples to represent the time series which is the temporal information of the video. In other words, the preprocessing stage creates the 5-dimensional array from video and a corresponding label. This 5D array is (number of samples, time series (sequences length), the height of frames, the width of frames, channels number of frames). The sequence length is chosen based on the fact that the human visual system has a reduction time about 1-2 seconds, which means about 30-60 frames since FPS is normally 30. So that 40 frames is likely to be reasonable as an initial interval. Figure 4 shows the block diagram of sequence generation with overlapping.

The other function that is done during the pre-processing stage is Region of Interesting (ROI) extraction by using Upper body Haar Cascade extractor; which is one of open CV functions. This step is most important because the variation in head scale and plan rotation can be eliminated. The captured frames in each sample input to the extractor and output samples containing heads.

Data resizing (scale) and data normalization (data modelling to DLNN) are also done as a preprocessing stage functions. A resizing the dimensionality of cropped images in the samples is more important step to provide the generality to samples and to be suitable for the prediction deep learning model. The image scaling interprets as an image resizing that involve reconstruction image from one pixel grid to another; which detected in the proposed system to 100 × 100, by increase or decrease the total number of pixels in images of samples. Normalize the pixels value is done across all channels by dividing all pixels values by the largest pixel value; that is 255, regardless of the actual range of pixel values present in the image.

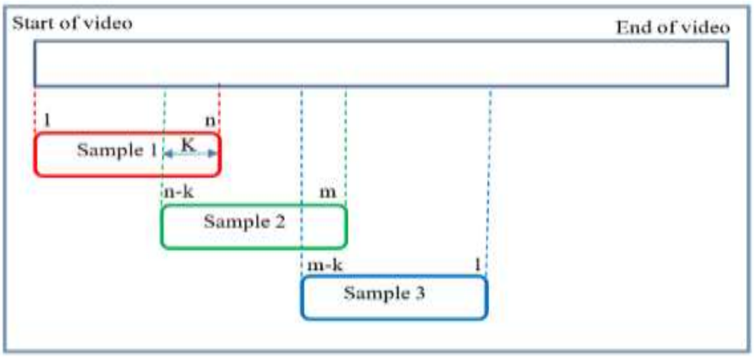

Figure 4. Block diagram of sequence generation with overlapping

2.2.2 The video classification stage

The next stage implements the feature extraction and video classification for abnormal head movement by using the same 3D CNN model. The main objective of the proposed model is to predict the drowsiness status of drivers based on detecting the driver abnormal head movements and tracking these movements over the time by the idea of deep learning in Neural Network. This model is also based on solving the problem completely (detecting and tracking the features) from the input until predicting the drowsiness state of the driver based on abnormal head movements feature.

This model consists of two main tasks (feature extraction task, and classification task) each task consist of different layers that done different functions according to specific objective to each layer. Each layer uses saves weights to produce the output of this layer to be the input to the next layer and so on until the final fully connected layer make the decision and produced the output of the network. These layers are 3D convolution layers, non-linear layers, pooling layers, dropout layers, batch normalization layers, and fully connected layers.

The first kind of layers in the model is 3D convolution layer which use 3D kernel (filter) on sequences of frames to extract feature and produced feature maps. These filters represent the depth of the layer and uses (32, 64, 128, and 128) respectively for the convolution layers. These filters are applied in spatial and temporal dimension for each image in input sample and the stride of these kernels used are (3, 3, and 3).The saved weights represented the kernel coefficient which can be decided during the training process. The 3D convolution layer use kernel stride of 3-dimensions. Non-linear layers are represented by using 'Rectified Linear Units (ReLUs) function' for all convolution layers and 'sigmoid function' for fully connected layers. These functions employed to determine the distinct features for each hidden layer.

The Max-pooling also called subsampling layer which reduces the resolution of the features to provide robust against noise and blurring. The reduction achieves on a spatial and temporal dimension by combining some neuron clusters at one layer into a single neuron in the next layer by applying a max-pooling function to produce the maximum value of these clustering neurons. The pooling layer of 3D CNN model deals with clusters of neurons which are three dimensions and non-overlapping spaces (that's mean this layer pooled in spatial and temporal dimensions).

Fully connected layers represent the final layers which are flattening the output of previous layers and apply the activation function to classify the sequence of frames. The last fully connected layer can be called output layer or decision-making layer which is implementing a sigmoid function to produce the final class of the sequences. To prevent the overfitting and provide the generalization on unseen data the proposed system provide the dropout layers. The dropout layer selects a portion of neurons randomly; determined as 25%; and set their weights to zero during the training process. The dropout layer is a simple way to effectively control the model sensitivity to the noise during the training process while maintaining the required complexity of the proposed model architecture.

3. RESULTS AND ANALYSIS

The metrics used in the proposed system are the accuracy function, loss and mean square error functions that use in each step of training and validation processes. The proposed system such as all Neural Network systems is implementing in two stages: learning stage and prediction stage. In the first stage, the proposed model is trained on all available training dataset to learn on the samples represented the desired labels by provided these training samples with corresponding labels to the model. The weights corresponding to each layer adjusted through training process and this updating of weights is continuous until the network is convergence to the minimum error; by using mean square error. After the network stable, the validation process implements (which can be consider a compatible part of the learning stage) on available validation dataset with corresponding labels and used the saved weights from training process to prove the performance and accuracy of the trained model and to decide if the model is prepared to predict the class of new data. The second stage represents the prediction of the trained proposed model to the output of unseen (untrained samples) inputs which is the application of the system. The implementation stages of the proposed system over total dataset illustrate in Figure 5.

The summary of each layer of the proposed system with its corresponding parameters and output shape is illustrated in Table 1. The learning behavior for the proposed model on specific training and validation dataset in each epoch is shown in plotting figures represent the results of metrics used in each epoch as shown in Figure 6.

Gradient descent is used in training the neural networks and the weights are updating based on error estimation of a subset of the training dataset. The number of examples in this subset of the training dataset is known a batch size. Batch size is considered an important hyperparameter that influences the dynamics of the learning algorithm, so it's controlling on the quickly that learning the algorithm. The experimental results of the batch size and learning rate hyperparameters tuning in the model are represented by Table 2 for batch size choose and Table 3 for learning rate choose.

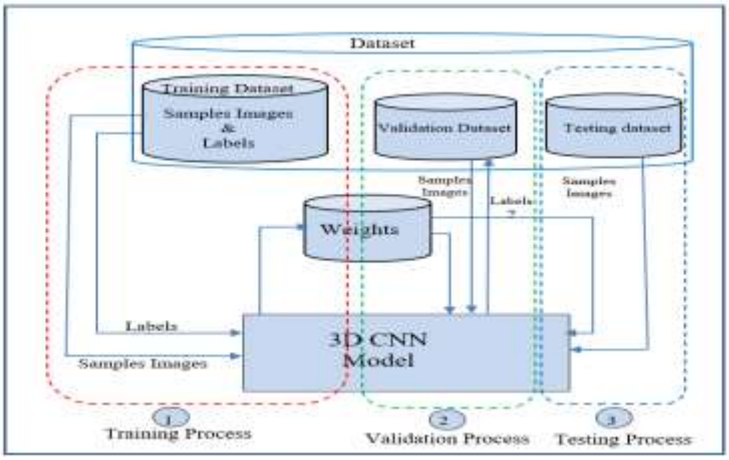

Figure 5. The proposed system implementation stages

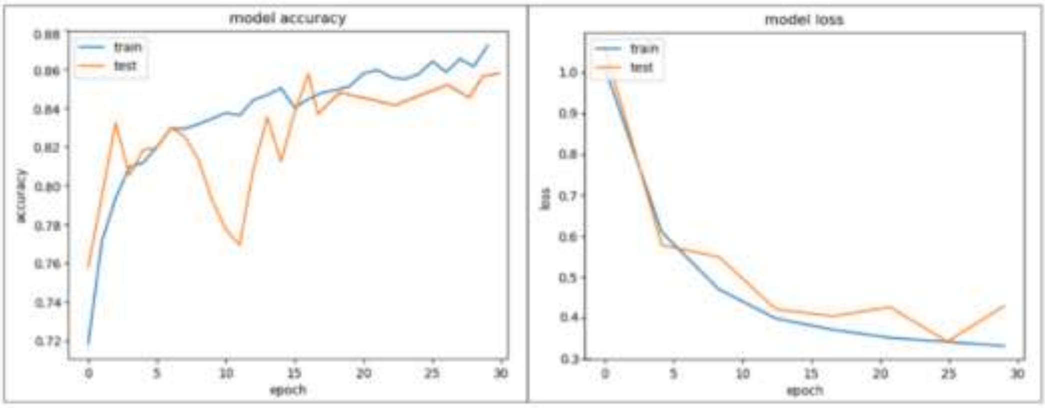

Figure 6. Accuracy and loss functions of model

Table 1. Proposed system summary

Block no.	Layer Type	Output Shape	Params no.
1	Conv1 (Conv3D-size (3,3,3))	(None, 8, 100, 100,32)	2624
	Pool1 (MaxPooling3D)	(None, 8, 50, 50,32)	0
		(None, 8, 50, 50,32)	0
2	Conv2 (Conv3D-size (3,3,3))	(None, 8, 50, 50,64)	5560
	Batch_normalization_2	(None, 8, 50, 50,64)	256
	Pool2 (MaxPooling3D)	(None, 4, 25, 25, 64)	0
	Dropout_2 (Dropout)	(None, 4, 25, 25, 64)	0
3	Conv3 (Conv3D-size (3,3,3))	(None, 4, 25, 25, 128)	221312
	Batch_normalization_3	(None, 4, 25, 25, 128)	512
	Pool3 (MaxPooling3D)	(None, 2, 12, 12, 128)	0
	Dropout_3 (Dropout)	(None, 2, 12, 12, 128)	0
4	Conv4 (Conv3D-size (3,3,3))	(None, 2, 12, 12, 128)	442496
	Batch_normalization_3	(None, 2, 12, 12, 128)	512
	Pool4 (MaxPooling3D)	(None, 1, 6, 6, 128)	0
	Dropout_4 (Dropout)	(None, 1, 6, 6, 128)	0
5	Flatten_1 (Flatten)	(None, 4608)	0
	Dense_1 (Dense)	(None, 1000)	4609000
	Batch_normalization_4	(None, 1000)	4000
	Dropout_5 (Dropout)	(None, 1000)	0
6	Dense_2 (Dense)	(None, 50)	50050
	Batch_normalization_5	(None, 50)	200
	Dropout_6 (Dropout)	(None, 50)	0
7	Dense_3 (Dense)	(None, 1)	51
Total Params:		5,386,373	
Trainable params:		5,383,633	
Non-trainable patams:		2,740	

Table 2. Batch size tuning of HMSDLNN

Batch Size	Mode	Loss Func.	MSE	Accuracy
5	Training	0.5095	0.1734	72.54
	Testing	0.5503	0.1760	0.7329
10	Training	0.4802	0.1629	74.93
	Testing	0.6950	0.2056	75.04
25	Training	0.4661	0.1592	75.39
	Testing	0.5968	0.1887	74.77
32	Training	0.4727	0.1615	74.73
	Testing	0.5188	0.1681	75.80

Table 3. Learning rate tuning of ESDLNN model

Learning Rate	Mode	Loss Func.	MSE	Accuracy
0.01	Training	0.3526	0.1157	83.06
	Testing	0.8300	0.2766	0.6027
0.001	Training	0.3991	0.1327	0.8060
	Testing	0.5606	0.2023	0.6351

4. EVALUATION MODEL

In this stage, the proposed system is evaluated based on the testing dataset. The performance metrics such as Accuracy, Convution Matrix, Recall, F1-Measure, Specificity, and Precision are used to evaluate the proposed model. The concept of Cross Validation CV is employed to become evaluation that is more reliable. This technique employing the smart ruse by performing K rounds of training-validation-testing on, different, non-overlapping, equally-proportioned Training (Tr = 80%), Validation (Va = 10%), and Testing (Te = 10%) sets.

The evaluation results of the HMSDLNN model are presented through the Table 4 for confusion matrix and Table 5 presents the evaluation of the model in the term of Performance Metrics. The 10-Fold CV implemented and the experimental results can be represented through the Table 6. And finally, a comparison of the proposed system with state-of-the-art methods presented in Table 7 below. The bold numbers represent the best accuracy.

Table 4. The confusion matrix CM

N=4103		Predictive model	
Actual Recommended	Yes	Yes	No
	No	639	1715

Table 5. The evaluation measures valus

Measure	Value
Accuracy	0.8406044357786985
Sensitivity (TP Rate)	0.9914236706689536
Specificity (FP Rate)	0.27145284621920135
Precision	0.7307206068268015
Recall	0.9914236706689536
F1-Score	0.841339155749636

Table 6. 10-fold CV for HMSDLNN model

No of fold	Accuracy of each fold	Model accuracy
1	82.21%	85.37%(+/-1.54%)
2	84.89%	
3	83.78%	
4	85.83%	
5	87.12%	
6	85.51%	
7	82.06%	
8	82.92%	
9	84.32%	
10	84.43%	

Table 7. Drowsiness detection accuracies (%) for different scenarios of the NTHU-DDD database

Scenarios	Human [23]	3D-DCNN [24]	3-nets DDD [23]	MLP [25]	MT-DMF [26]	Seq MT-DMF [26]	Proposed system
Bareface	82.04	75.10	69.83	87.12	76.04	84.46	**88.13**
Glasses	78.83	72.30	75.93	**84.85**	74.17	77.35	83.43
Sunglasses	80.89	70.90	69.86	75.11	72.39	**86.43**	85.77
Night-bareface	82.54	68.40	74.93	81.40	77.16	82.48	**84.58**
Night-glasses	79.87	68.30	74.77	76.15	78.56	**87.18**	84.28
Average	80.83	71.20	73.06	80.93	75.73	83.44	**85.238**

5. CONCLUSION

The proposed system presented in this paper performs drowsiness video classification based on determining efficient feature maps. The employment of deep learning used to extract high-level image features from the input sample (sequence of frames) by implementing a series of non-linear operation, then classifying these input sample (sequence of frames) depending on the extracted feature. Spatial-temporal information of the video is exploited using a 3D CNN, this temporal information represents the time stamp and can be considered to make an efficient decision of drowsiness video classification as well as give the proposed system better accuracy and loss functions than system based on 2D CNN model. The comparison and outperformance of 3D CNN system with 2D CNN system can be present which prove the accuracy enhancement with 3D CNN as shown in the experimental results.

REFERENCES

[1] I. T. Teivas, “Video event classification using 3D convolutional neural networks,” Thesis, Tampere University of Technology, 2017.

[2] A. A. Abdulkadhem and T. A. Al-assadi, “An important landmarks construction for a GIS-Map based on indexing of dolly images,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 1, pp. 451–459, July 2019.

[3] G. Cai, K. Lee, and I. Lee, “A framework for mining semantic-level tourist movement behaviours from geo-tagged photos,” *Australasian Joint Conference on Artificial Intelligence*, vol. 9992, pp. 519–524, November 2016.

[4] H. Wang et al., “A robust and efficient video representation for action recognition,” *International Journal of Computer Vision Manuscript*, 2016.

[5] B. Reddy, Y. H. Kim, S. Yun, C. Seo, and J. Jang, “Real-Time Driver Drowsiness Detection for Embedded System Using Model Compression of Deep Neural Networks,” *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Honolulu, HI, pp. 438-445, 2017.

[6] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, “Time-Delay Neural Network for Continuous Emotional Dimension Prediction From Facial Expression Sequences,” in *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 916-929, April 2016.

[7] M. Patacchiola and A. Cangelosi, “Head pose estimation in the wild using Convolutional Neural Networks and adaptive gradient methods,” *Pattern Recognition,* vol. 71, pp. 132–143, November 2017.

[8] I. Hadi and A. Mahdi, “Generating images of partial face using landmark based k-nearest neighbor,” *Indonesian Journal of Electrical Engineering and Computer Science,* vol. 17, no. 1, pp. 420–428, January 2020.

[9] E. Al-Shamery and A. A. H. Al-Shamery, “A New Deep Neural Network Regression Predictor Based Stock Market,” *Journal Engineering and Applied Sciences*, vol. 13, no. 5, pp. 4794–4801, 2018.

[10] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, “Deep spatio-Temporal features for multimodal emotion recognition,” *2017 IEEE Winter Conference on Applications of Computer Vision (WACV),* Santa Rosa, CA, pp. 1215-1223, 2017.

[11] Q. Lan, Z. Wang, M. Wen, C. Zhang, and Y. Wang, “High Performance Implementation of 3D Convolutional Neural Networks on a GPU,” *Computational Intelligence and Neuroscience*, vol. 2017, pp. 1–8, November 2017.

[12] M. Khosla, K. Jamison, A. Kuceyeski, and M. R. Sabuncu, “3D Convolutional Neural Networks for Classification of Functional Connectomes,” *Arxiv,* pp. 137–145, 2018.

[13] A. Khvostikov, K. Aderghal, J. Benois-Pineau, A. Krylov, and G. Catheline, “3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies,” *Arxiv*, January 2018.

[14] J. Foytik and V. K. Asari, “A two-layer framework for piecewise linear manifold-based head pose estimation,” *International Journal of Computer Vision,* vol. 101, pp. 270-287, 2013.

[15] V. N. Balasubramanian, J. Ye and S. Panchanathan, "Biased Manifold Embedding: A Framework for Person-Independent Head Pose Estimation," *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, pp. 1-7, 2007.

[16] J. Sherrah, S. Gong, and E. J. Ong, “Face distributions in similarity space under varying head pose,” *Image and Vision Computing,* vol. 19, no. 12, pp. 807–819, October 2001.

[17] E. Osuna, R. Freund, and F. Girosit, “Training support vector machines: an application to face detection,” *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, USA, pp. 130-136, 1997.

[18] M. Jones and P. Viola, “Fast multi-view face detection,” *Mitsubishi Electric Research Laboratories,* August 2003.

[19] M. S. L. Khan, S. U. Réhman, Z. LV, and H. Li, “Head Orientation Modeling: Geometric Head Pose Estimation using Monocular Camera,” *Proceedings of the 1st IEEE/IIAE International Conference on Intelligent Systems and Image Processing,* pp. 149-153, 2013.

[20] A. Kumar, A. Alavi, and R. Chellappa, “KEPLER: Keypoint and Pose Estimation of Unconstrained Faces by Learning Efficient H-CNN Regressors,” *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017),* Washington, DC, pp. 258-265, 2017.

[21] J. Gu, X. Yang, S. De Mello, and J. Kautz, “Dynamic facial analysis: From Bayesian filtering to recurrent neural network,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Honolulu, HI, pp. 1531-1540, 2017.

[22] I. S. Rieger, T. Hauenstein, and S. H. Bamberg, “Head Pose Estimation using Deep Learning,” Master’s Thesis, Faunhofer-Institute for Integrated Circuits IIS, April 2018.

[23] S. Park, F. Pan, S. Kang, and C. D. Yoo, “Driver drowsiness detection system based on feature representation learning using various deep networks,” *Asian Conference on Comouter Vision*, vol. 10118, pp. 154–164, March2017.

[24] Jongmin YuSangwoo ParkSangwook LeeMoongu Jeon, “Representation Learning, Scene Understanding, and Feature Fusion for Drowsiness Detection,” *Asian Conference on Computer Vision (ACCV),* vol. 10118, pp. 165-177, March 2017.

[25] R. Jabbar, K. Al-khalifa, M. Kharbeche, and W. Alhajyaseen, “ScienceDirect Real-time Driver Drowsiness Detection for Android Application Using Deep Neural Networks Techniques,” *Procedia Computer Science,* vol. 130, pp. 400-407, 2018.

[26] L. Celona, L. Mammana, S. Bianco, and R. Schettini, “A Multi-Task CNN Framework for Driver Face Monitoring,” *2018 IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin),* Berlin, pp. 1-4, 2018.