ISSN: 2096-3246
Volume 54, Issue 03, August, 2022

Accurate Vehicular Movement Understanding in Aerial Images Using Convolutional Neural Network Architecture

Pushpalata[1] and M Sasikala[2]
Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering & Technology (Exclusively for women) Sharnbasva University Kalaburagi, Karnataka- India[1]
Professor, Department of Electrical and Electronics Engineering, Faculty of Engineering & Technology (Exclusively for women) Sharnbasva University Kalaburagi, Karnataka- India[2]

Keywords:
Convolutional Neural Network (CNN) architecture, Event Recognition Aerial (ERA) dataset, Classification, Vehicular data

ABSTRACT

With the faster development and advancement in surveillance and intelligence field such as the use of Unmanned Aerial Vehicles (UAVs) shows significant importance of object detection and movement understanding. However, data is gathered from varied locations, and the classification of that abundant amount of data is a challenging process. Thus, vehicle identification in aerial images is an essential and important research area. Therefore, a Selection Decision-Classification ($SelDes-Clas$) model in the Convolutional Neural Network (CNN) is adopted in this work to efficiently classify a large amount of vehicular data and correctly identify which data belongs to which class. The proposed $SelDesClas$ modelis transformed into the Selection and Decision Network ($SedeNet$) and Classification Network ($ClsNet$). Here, the $SedeNet$ model works on the essential image region selection and generation of preliminary weights and the $ClsNet$ model performs an efficient training and performs classification process. The event Recognition Aerial (ERA) dataset is utilized to evaluate the classification performance of the proposed $SelDesClas$ model. Binary classification is performed to test the model and its efficiency. The obtained classification accuracy using Traffic-Collision is 56.7%and Traffic-Congestion is 69.8%.

1. INTRODUCTION

Nowadays, the utilization of Unmanned Aerial Vehicles (UAVs) is massively enhanced for different purposes such as monitoring [1], research, tracking, and assistance [2-3]. Further, UAVs are majorly utilized in varied fields such as agriculture, filming, entertainment, defense, relief works in case of emergencies or disaster situations, item delivery, and so on. UAVs can provide real-time videos with high resolution at minimum cost. TheseUAVs are camera associated devices that can provide essential information for the purpose of object detection, movement understanding, and tracking in crowded places or events [4]. The performance of UAVs in terms of capturing desired information and accurate target identification is impeccable in real-time streaming. It is generally predicted that desired scenario descriptions, mobile target identification, and basic tracking remain absolutely precise. Thus, their capabilities to provide such useful information can enable faster decision-making [5]. However, there are some factors that may affect their performance such as weather conditions, cloud behavior, rainfall and

camera quality, etc. The main goal of adopting UAVs is their ability to move in those places where humans generally cannot reach especially in case of disastrous situations, relief works, or military operations. Because of their small size, low cost, and multiple advantages, there is an exponential growth in the utilization of UAVs. Especially, the adoption of UAVs is massively enhanced in cases where chances of collision or congestion, or crowd gathering are more such as musical concerts, political rallies, religious events gatherings, sporting events, and traffic situations in metropolitan cities. In those types of gatherings, there is a huge chance of collisions or crowd congestion and sometimes, the unwanted clash between two ultra-groups. So, in those situations, there is a necessity for continuous monitoring, object detection, movement understanding, and tracking on a regular basis. Therefore, UAVs have emerged as extremely powerful devices in recent times which can provide continuous information on desired location or event or crowd gathering while control remains in the hand of the operator and can operate from different locations.

However, continuous analysis of obtained aerial videos is a completely different task to maintain peace and safety. Accurate detection of mobile objects, object movements and tracking of those objects can become an essential solution to understanding crowd behavior [6] and maintaining a peaceful environment so that situations related to collisions or congestion can be avoided. Furthermore, the detection of humans or objects in crowded places has gained a lot of attention in the research community, especially in the past two decades [4]. Object detection, movement understanding and tracking methods are majorly used in applications like crowd gatherings, law enforcement [7], traffic management, fall detection [8] and automated driver assistance [9], etc. Furthermore, real-time identification of objects in gathered aerial videos contains a few challenges which need to be addressed quickly in order to provide efficient object detection, movement understanding and tracking technique. For example, those challenges are a variation of altitude i.e. videos captured at different altitudes will provide different object sizes. This poses a serious challenge in the case of human identification [10]. Another issue is the impact of dynamic and environmental events such as illumination variations or noticeable degrees of motion and weather conditions, respectively. These situations can blur camera images and can cause jitter [11]. Thus, these issues are required to be addressed precisely by adopting an efficient object detection and tracking method based on the robust classification analysis which can identify humans in difficult situations as well. High-quality spatial and temporal features need to be generated with exact object boundaries so that objects can be identified accurately.

Several researchers have made their efforts in providing such object detection, movements understanding and analysis techniques from the captured aerial videos by UAVs. Some of the research works are discussed in this paragraph. A spatial pyramid context-aware approach [12] is adopted to analyze for the identification and tracking of moving vehicles from the aerial captured data. Here, visual object features are obtained using a collaborative spatial pyramid context approach, and temporal motion information of objects is captured. An event detection method [13] is presented to analyze crowd activities based on deep neural networks from captured aerial videos. This technique focuses on obtaining rich semantic features to detect objects accurately. The simulation results are obtained using a novel ERA dataset. Human-object Identification model [14] is adopted based on the CNN (Convolutional Neural Network) to analyze aerial captured data. Gamma correction and non-local means are determined using a filtering algorithm to pre-process video frames. Four varieties of features are obtained using the t-distributed stochastic neighbor embedding method. In [15], the CNN-based object detection technique is utilized to recognize human activities from the obtained aerial videos. A comparative analysis was presented between varied human detectors and feature extractor methods. The average accuracy obtained using this technique is 92.9%.

However, accurate vehicular movement understanding of classes is quite essential and challenging, when a large amount of data is present. As a result, there is a need for an efficient vehicular classification mechanism. Thus, a Selection Decision-Classification ($SelDesClas$) model in the Convolutional Neural Network (CNN) architecture is presented to analyze vehicular data and make a decision that which frame belongs to which class. Another objective is to improve classification accuracy and other performance metrics considering vehicular data. The proposed CNN architecture design consists of different layers (pooling layers, softmax layers, sequential layers, convolutional layers) and different blocks (dense blocks, residual blocks). To achieve better classification results, the proposed $SelDesClas$ model is isolated into two different models as Selection and Decision Network (SeDeNet model) and Classification Network ($ClsNet$) model, and both models remain linked to each other based on some parameters and functions. The $SeDeNet$ model is utilized to get the most important pixels near the central position and to analyze image regions to obtain feature weights. The $ClsNet$ model is used to analyze the feature weights which are generated in the $SeDeNet$ model to get the final feature maps and obtain classification results from feature maps. Classification results are obtained in terms of classification accuracy and precision and compared against varied vehicle classification methods.

This work is demonstrated in the following sections which remain as follows. In section 2, a mathematical representation of the vehicular movement understanding using CNN architecture is presented based on the feature extracted using CNN architecture. In section 3, simulation results are discussed and compared against varied classical object movement analysis methods and section 4 concludes the present work.

2. MODELING FOR VEHICULAR CLASSIFICATION ANALYSIS

In this section, a detailed classification analysis is performed using the proposed $SelDesClas$ model. Generally, aerial data is gathered from different domains or fields. Practically, segregation of this abundant amount of aerial data into different classes and particular data belongs to which class, that identification becomes a very hectic and tedious task. Therefore, an automatic deep learning model is required to classify those large datasets and accurately identify which particular vehicular data belongs to which class, and provide higher classification results. Therefore, in this work, a Convolutional Neural Network (CNN) model is adopted to automate the classification of aerial vehicular data and identify classes accurately. The proposed $SelDesClas$model consists of multiple layers like sequential layers, pooling layers, softmax layers, and convolutional layers of different sizes. The architecture also contains multiple blocks like dense blocks and residual blocks. The activation function utilized in the proposed $SelDesClas$ model is rectified linear activation function ($ReLU$) which is used to handle training computations and maintains multiple operations in the proposed $SelDesClas$ model. The next paragraph demonstrates how precise classification can be achieved using the proposed $SelDesClas$ model. Figure 1 shows the proposed $SelDesClas$model which contains convolutional blocks and dense blocks. The proposed architecture consists of varied layers, blocks, and activation functions.

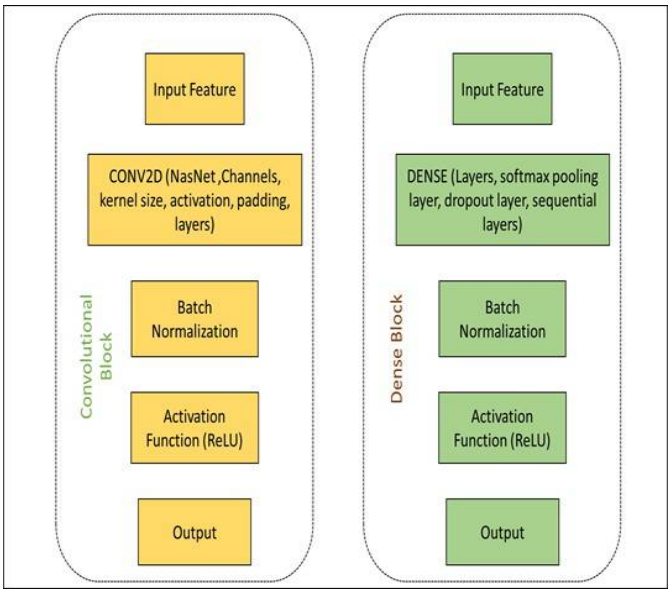


Figure 1: Proposed CNN Architecture

Here, the proposed $SelDesClas$model gets divided into separate phases to obtain high classification accuracy. Those two phases are Selection and Decision Network ($SeDeNet$) and Classification Network ($ClsNet$) model. Both models remain linked with each other through some commonly used functions, layers, and parameters. The main objective of forming the $SeDeNet$ model is to analyze the given input image and select the most important and significant regions from those given images automatically based on the parameters and functions incorporated in the architecture. As a result, high-quality feature weights are generated and the $SeDeNet$ model auto-tunes some specific parameters and functions to get maximum efficiency in the $ClsNet$model. Further, those generated feature weights are further processed while training of proposed $SelDesClas$model in the $ClsNet$model, and those trained model provides the final feature maps to get the best testing results in terms of varied performance metrics.

Each image is analyzed in the $SedeNet$ model in every iteration and based on the analysis, specific regions are identified and these regions contain pixels that have the most significant information related to the given image. Based on this information, some specific decisions are taken which improve the training efficiency of the $ClsNet$ model. Thus, each image is batch processed in two sub-phases which are the selection phase and the decision phase. Here, the main focus of the selection stage inside the $SedeNet$ model is to choose the most significant image regions. Thus, significant information is captured related to the image based on defined parameters and functions. The next step is a deep analysis of those obtained image regions to get the feature weights and some essential decisions are taken regarding architectural design and parameter selection so that efficient training is performed in the $ClsNet$ model. Finally, detailed feature maps are generated and labels are assigned to each image to obtain high-quality classification performance.

In the $SedeNet$ model, essential image regions are captured based on the given coordinates. Then, the number of steps is selected as r, and based on the central coordinates y_{r-1} image regions m_r are obtained. Further, based on the essential image information, the $SedeNet$ model is parameterized by a modeling parameter Θ_q and based on the certain parametric decisions $\pi_{\Theta q}(K_r)$, the discriminative features are obtained.

ISSN: 2096-3246
Volume 54, Issue 03, August, 2022

The obtained discriminative features are expressed by $K = (D_P, D_G)$, and contain significant information such as analysis report representation D_G from selected image regions and input image statistics D_P. This analysis report consists of information related to the preliminary feature weights which can be used for the further training process so that efficient classification results are obtained. Whereas input statistics contain information related to the image and ground truth labels and central coordinate information y_{r-1}. After further processing of discriminative features K, these features are retransformed into Recurrent Layer Analysis (RLA) features S and Rich Attentive Information (RAI) features N: $K = S \cup N$. Encoding is performed using a fully linked layer to compress RLA features so that their dimensionality is reduced,

$$S_j^{'} = \Phi(U_s(S_j) + h_s) \quad (1)$$

Where dimensions of RLA features S consist of essential information as $U_s \in \mathbb{A}^{y \times y'}$, $h_s \in \mathbb{A}^{y'}$, y. The dimension of the compressed features S' is represented by y'. The activation function used in the proposed $SelDesClas$ model is $ReLU$ and represented by $\Phi(\cdot)$. In the same way, Rich Attentive Information (RAI) features N are compressed using another fully linked layer into encoded features N'.

$$N_j^{'} = \Phi(U_n(N_j) + h_n) \quad (2)$$

Where dimensions of RAI features N consist of essential information as $U_n \in \mathbb{A}^{e \times e'}$, $h_n \in \mathbb{A}^{e'}$, e. The dimension of the compressed features N' is represented by e'. Then, regenerate the discriminative features by rejoining both S' and N' features.

$$i_j = \Phi(U_k(N_j^{'} || S_j^{'}) + h_k) \quad (3)$$

Where dimensions of rejoined discriminative features consist of essential information as $U_k \in \mathbb{A}^{(y'+e') \times t}$, $h_k \in \mathbb{A}^{t'}$, t and $||$ is a rejoining function. Furthermore, activation function, softmax layers, and hidden layers are utilized to obtain some specific decisions related to the image analysis which can be utilized in the training of the proposed model.

The $ClsNet$ model is parameterized as Θ_d using a classification-enabled assessment mechanism $d_k(m_r; \Theta_d)$ and the obtained image regions m_r and obtained decision-related information are also compressed. These image regions contain the most valuable information about the image. Then, by analyzing this encoded information, feature maps are generated in the $ClsNet$ model as follows,

$$B(m_r) = (1 + C(m_r) * R(m_r)) \quad (4)$$

Where feature maps $B(m_r)$ are the combination of two separate information packages in which the first package $C(m_r)$ contains information related to encoded image regions and the second package $R(m_r)$ contains information regarding obtained decisions from the $SedeNet$ model. Generally, equation (4) is achieved by learning about residual blocks and pooling layers. Furthermore, fully linked layers and activation functions are utilized to generate feature vectors from the feature maps $B(m_r)$. Finally, the classification process is performed on obtained feature vectors using soft-max layers to get an accurate class prediction.

Furthermore, the optimization of the proposed $SedeNet$ model and $ClsNet$ model is performed to improve classification accuracy by minimizing entropy loss as follows,

$$\eta_s(\hat{w_j}, w_j) = -\frac{1}{X}\sum_{j=1}^{x}[w_j log\,(\hat{w_j}) + (1-w_j) log\,(1-\hat{w_j})] \quad (5)$$

Where ground-truth labels obtained using the $SedeNet$model are expressed by $w_j$and predicted class labels obtained using the $ClsNet$ model are represented by $\hat{w_j}$. The optimization function $\pi_\Theta(b_r|k_{1:r})$ is utilized to learn the efficiency of training models and convergence rate and can be improved as the following equation,

$$L(\Theta) = \mathbb{M}v(s1{:}\psi;\Theta)\left[\sum_{r}^{R}\left(f_v^R + f_s^R\right)\right] = \mathbb{M}_v(s_1{:}\psi;\Theta)[F] \quad (6)$$

The approximation of $L(\Theta)$ is achieved by the following equation,

$$\nabla_\Theta L = \sum_{r=1}^{R}[\nabla_\Theta \log \pi_\Theta(b_r|k_{1:r})F] \quad (7)$$

$$\nabla_\Theta L = 1/Z \sum_{l=1}^{Z}\sum_{r=1}^{R} \nabla\Theta \log \pi\Theta\,(brl|k1l{:}r)Fl \quad (8)$$

Where processing epochs are represented by $l = 1 \ldots \ldots \ldots .. Z$. The value of $\nabla_\Theta L$ gives probability predictions for each image considering all given classes. More the value of $\nabla_\Theta L$, predictions are better and vice versa. So accordingly, parameters can be set. However, equation (7) and equation (8) generate large variance and can make training unstable. Thus, the training instability can be handled efficiently by baseline subtraction using the following equation,

$$\nabla_\Theta L = Z\sum_{l=1}^{}\sum_{r=1}^{R} \nabla\Theta \log \pi\Theta\,(brl|k1l{:}r)(Fl - h_r) \quad (9)$$

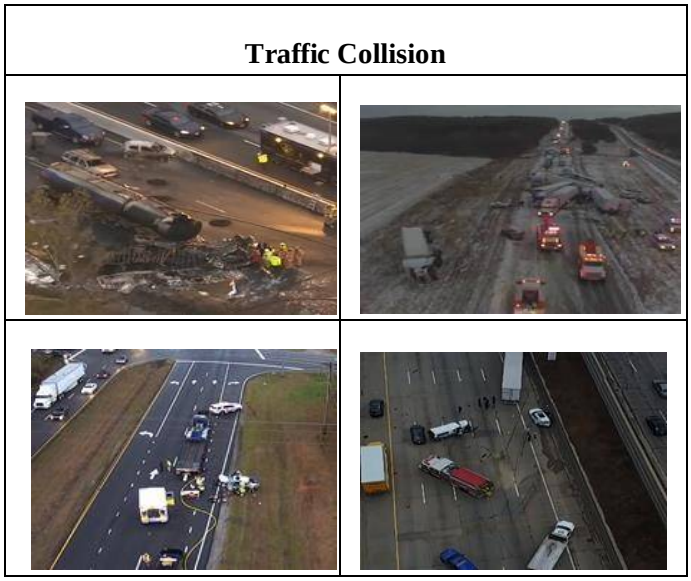
Where h_r is the average performance value of all the epochs. Here, equation (9) produces less variance and obtains a similar value which produces in equation (8). Finally, the obtained training models are utilized for the testing of the proposed $SelDesClas$model to get classification results in terms of classification accuracy and precision with minimum loss.

Testing of the proposed $SelDesClas$ modelis performed by comparing assigned class labels and predicted class labels.

3. RESULTS AND DISCUSSION

This section provides details about the vehicular dataset, testing performance, and comparison of obtained classification performance. A detailed comparison of testing performance obtained using the proposed $SelDesClas$ modelis presented in this section in terms of classification accuracy. First of all, the base of high classification results in testing is efficient training and model performance. First of all, aerial vehicular data is gathered with different classes. Then, an input image is passed to the proposed model, and an analysis is performed on the given images. Further, with the help of the $SeDeNet$ model, the regions which contain the most significant image information, are selected. Afterward, a deep analysis using varied convolutional layers, pooling layers, and dense blocks is performed, and some essential decisions which can improve training efficiency are derived and preliminary feature weights are generated. Based on those obtained feature weights and essential parametric and functional decisions, efficient training of vehicular data is performed using the $ClsNet$ model. Then, model feature maps are obtained which can be used for generating testing results in terms of classification accuracy and other performance metrics. The proposed $SelDesClas$ model efficiently identifies which class belongs to which class.

The vehicular dataset utilized to test the performance of the proposed $SelDesClas$ modelis the Event Recognition Aerial Dataset (ERA) dataset [16]. There are a total number of 25 classes present in this dataset which are gathered from YouTube and captured from different domains like sports, event gathering, traffic, construction, crowd gathering, etc. There are a total number of 2864 aerial videos available in this dataset of different domains and the length of each video is 5 seconds. The size of each image frame is 640 × 640. Figure 2 demonstrates obtained testing frames considering Traffic Collision and Traffic Congestion class. In Figure 2, the first two rows demonstrate images of the Traffic Collision class and the next two rows represent images of the
Traffic Congestion class.



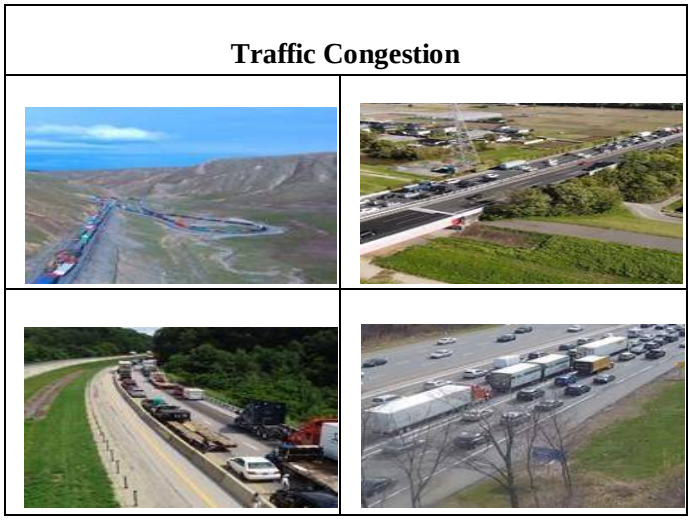


Figure 2: Overview of ERA Dataset

The performance of the proposed CNN architecture is compared against varied state-of-art techniques 3-D Convolutional Network (C3D) and Pseudo-3-D Residual Network (P3D ResNet) in terms of classification accuracy. The C3D models generate spatiotemporal features with the help of pooling layers and convolutional filters. This model preserves temporal information from input images. The filter size of the convolutional layer is 3 × 3 × 3. However, classification performance is comparably lesser that the proposed CNN architecture. Another conventional CNN model is the P3D ResNet model where spatial and temporal information is acquired using convolutional layers. This model reduces network size and is compared against the proposed CNN architecture. However, the classification accuracy is much higher than the conventional P3D ResNet model. Another classification model is Inflated 3-D ConvNet (I3D). This classification model of CNN architecture is designed using pooling and convolutional layers. Here, the pre-trained weights are utilized to get testing results, and results are obtained based on the ImageNet dataset.

There are a total number of two classes are utilized in which gathered data which is related to vehicles and traffic is used. Those two classes are Traffic-Collision and Traffic Congestion. The obtained classification accuracy using Traffic-Collision is 56.7and Traffic-Congestion is 69.8. These results are quite higher than the classical classification methods. The comparison between the proposed CNN architecture and classical models is presented in graphical form considering classification accuracy.

Here, Figure 3 demonstrates the comparison of classification performance for traffic collision class considering varied classification models such as C3D++, P3D++, and I3D++ against the proposed CNN architecture. The obtained classification accuracy using Traffic-Collision is 56.7% which is superior to other classification models.
Similarly, Figure 4 demonstrates the comparison of classification performance for traffic congestion class considering varied classification models such as C3D++, P3D++, and I3D++ against the proposed CNN architecture. The obtained classification accuracy using Traffic-Congestion is 69.8% which is superior to other classification models.

ISSN: 2096-3246
Volume 54, Issue 03, August, 2022

Furthermore, Figure 5 demonstrates the graphical representation of classification performance using the proposed CNN architecture. The obtained overall classification accuracy and average precision results considering all input images of both classes of Traffic collision and Traffic-Congestion are 93.99% and 90.42%, respectively which is quite superior to other classification models. It is evident from classification results that the proposed $SelDesClas$model shows much higher classification performance compare to any other previous classification models.

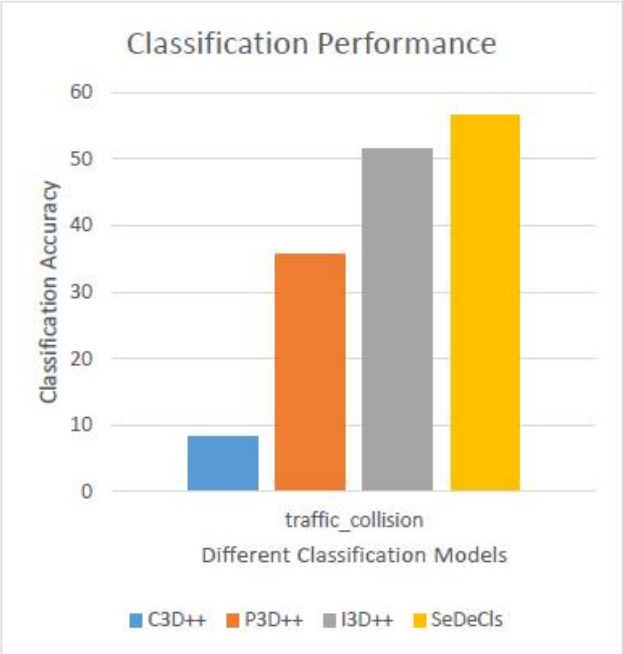


Figure 3 Classification Performance Comparison

Figure 4 Classification Performance Comparison

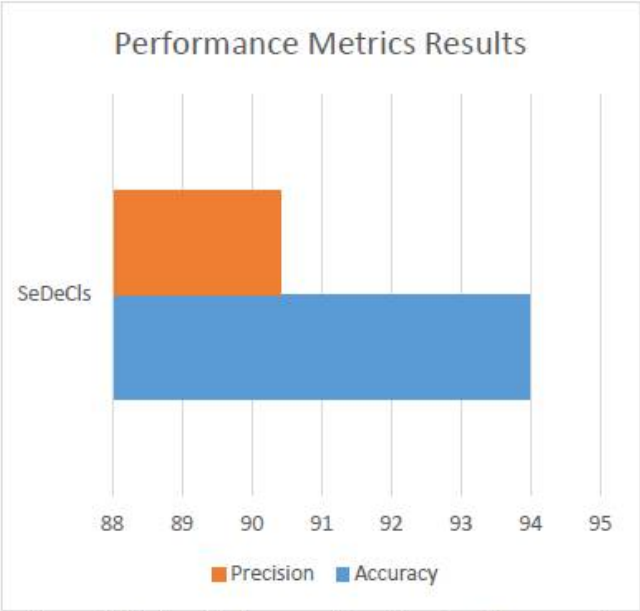


Figure 5 Testing Performance Results using the proposed *SelDesClas model*

4. CONCLUSION

Due to the high importance of vehicular data classification, in this work, a CNN-based classification model is presented to analyze vehicular data and identify which particular vehicle-related data belongs to which class among available multiple classes. The proposed CNN architecture design is segregated into two different phases ($SeDeNet$ model and $ClsNet$ model). Phase 1 discusses related to the image region selection and analysis to get preliminary weights to improve the training efficiency of the $ClsNet$ model. Phase 2 demonstrates the conduction of training and based on this efficient training, high-quality testing performance. Detailed mathematical modeling for $SeDeNet$ and $ClsNet$ model is presented. All the performance results while testing the proposed model are evaluated based on the ERA dataset. Here, the number of classes used from the given dataset to test the performance are two such as traffic Collision and Traffic Congestion. The obtained classification accuracy using Traffic-Collision is 56.7%and Traffic-Congestion is 69.8%. Furthermore, overall classification accuracy and precision considering all the images from both the classes of Traffic collision and Traffic-Congestion are 93.99% and 90.42%. It is evident from the classification results that the proposed $SelDesClas$ model performs significantly better than the previous CNN-based classification model.

REFERENCES

[1] B. Tamersoy and J. K. Aggarwal. Robust vehicle detection for trackingin highway surveillance videos using unsupervised learning. InIEEE AVSS, pages 529– 534, 2009. [2] P. V. K. Borges, N. Conci, and A. Cavallaro, “Videobasedhumanbehavior understanding: A survey,” IEEE Transactions on Circuits andSystems for Video Technology, vol. 23, no. 11, pp. 1993–2008, Nov2013.

[3] Y. Wu, J. Lim, and M. H. Yang, “Object tracking benchmark,” IEEETransactions on Pattern Analysis and Machine Intelligence, vol. 37,no. 9, pp. 1834–1848, Sept 2015.

[4] M. Paul, S. M. E. Haque, and S. Chakraborty, ``Human detection insurveillance videos and its applications_Areview,''EURASIP J. Adv.Signal Process., vol. 2013, no. 1, pp. 1_16, Dec. 2013, doi: 10.1186/1687-61802013-176.

[5] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, ``UAV-human: Alarge benchmark for human behavior understanding with unmanned aerialvehicles,'' Aug. 2021, arXiv:2104.00946. Accessed: Apr. 28, 2022.

[6] F. Negin, M. Koperski, C. F. Crispim, F. Bremond, S. Cos¸ar, andK. Avgerinakis. A hybrid framework for online recognition of activitiesof daily living in realworld settings. In IEEE AVSS, pages37–43, 2016.

[7] B. Engberts and E. Gillissen, ``Policing from above: Drone use by thepolice,'' in The Future of Drone Use: Opportunities and Threats FromEthical and Legal Perspectives, B. Custers, Ed. The Hague, The Netherlands:T.M.C. Asser Press, 2016, pp. 93_113, doi: 10.1007/978-94-6265-132-6_5.

[8] S. Ezatzadeh, M. R. Keyvanpour, and S. V. Shojaedini, ``A human falldetection framework based on multicamera fusion,''J. Exp. Theor. Artif.Intell., pp. 1_20, Jul. 2021, doi: 10.1080/0952813x. 2021.1938696.

[9] K. Balani, S. Deshpande, R. Nair, and V. Rane, ``Human detection for autonomous vehicles,'' in Proc. IEEE Int. Transp. Elec-tri_c. Conf. (ITEC), Aug. 2015, pp. 1_5, doi: 10.1109/ITEC-India.2015.7386891.

N. AlDahoul, A. Q. M. Sabri, and A. M. Mansoor, ``Real-timehuman detection for aerial captured video sequences via deep models,''Comput. Intell. Neurosci., vol. 2018, pp. 1_14, Feb. 2018, doi:10.1155/2018/1639561

[10] N. AlDahoul, A. Q. M. Sabri, and A. M. Mansoor, ``Real-timehuman detection for aerial captured video sequences via deep models,''Comput. Intell. Neurosci., vol. 2018, pp. 1_14, Feb. 2018, doi:10.1155/2018/1639561.

[11] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr,``Staple: Complementary learners for realtime tracking,'' in Proc. IEEEConf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 1401_1409, doi: 10.1109/CVPR.2016.156.

[12] M. Poostchi, K. Palaniappan and G. Seetharaman, "Spatial pyramid context-aware moving vehicle detection and tracking in urban aerial imagery," 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1-6, doi: 10.1109/AVSS.2017.8078504.

[13] L. Mou, Y. Hua, P. Jin and X. X. Zhu, "Event and Activity Recognition in Aerial Videos Using Deep Neural Networks and a New Dataset," IGARSS 2020 – 2020 IEEE International Geoscience and Remote Sensing Symposium, 2020, pp. 952-955, doi: 10.1109/IGARSS39084.2020.9324182.

[14] Ghadi, Y.Y.; Waheed, M.; al Shloul, T.; A. Alsuhibany, S.; Jalal, A.; Park, J. Automated Parts-Based Model for Recognizing Human–Object Interactions from Aerial Imagery with Fully Convolutional Network. Remote Sens.2022, 14, 1492. https://doi.org/10.3390/rs14061492

[15] N. Aldahoul, H. A. Karim, A. Q. M. Sabri, M. J. T. Tan, M. A. Momo and J. L. Fermin, "A Comparison Between Various Human Detectors and CNN-Based Feature Extractors for Human Activity Recognition via Aerial Captured Video Sequences," in IEEE Access, vol. 10, pp. 63532-63553, 2022, doi: 10.1109/ACCESS.2022.3182315.

[16] L. Mou, Y. Hua, P. Jin and X. X. Zhu, "ERA: A Data Set and Deep Learning Benchmark for Event Recognition in Aerial Videos [Software and Data Sets]," in IEEE Geoscience and Remote Sensing Magazine, vol. 8, no. 4, pp. 125-133, Dec. 2020, doi: 10.1109/MGRS.2020.3005751.

[17] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEEInt. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489–4497.

[18] Z. Qiu, T. Yao, and T. Mei, “Learning spatio-temporal representation with pseudo-3D residual networks,” in Proc. IEEE Int. Conf. Comput.Vis. (ICCV), Oct. 2017, pp. 5533–5541.

[19] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the Kinetics dataset,” in Proc. IEEE Conf. Comput. Vis.PatternRecognit. (CVPR), Jul. 2017, pp. 6299–6308.