# Vehicle Classification with Audio and Video Modalities Using CNN and Decision-Level Fusion

Vijay Viswanath, Ben P. Babu
Department of Electronics and Communication Engineering
Rajiv Gandhi Institute of Technology, Kottayam, India
vijayv.6595@gmail.com, benpbabu@gmail.com

*Abstract*— **The paradigm of vehicle classification is ever evolving, with new requirements and challenges adding up at every juncture. It remains an important task in the efficient management of traffic and road infrastructure. Much work has been done over the years to perfect this task. Complementarity of information has been used before to improve classifier performance using various classifier fusion techniques. This has however been not explored in the case of a completely neural network based multi-modal classification system. In this work, a multi-modal vehicle classification system is proposed which utilizes the useful property of complementarity to achieve improved performance. Classification of vehicles is performed with the two modalities separately, using Convolutional Neural Network (CNN) classifiers. In the case of audio modality, sets of Mel Frequency Cepstral Coefficients (MFCC) are the feature vectors. The individual predictions from the base classifiers are fused at the decision-level to arrive at a final prediction. The results of both single and multi-modal classification are compared. The results show that decision-level fusion improves the accuracy of classification.**

**Keywords-Vehicle classification; Intelligent Transportation Systems (ITS); Image processing; Information complementarity; Deep learning; Convolutional Neural Networks (CNN); Visual Geometry Group 16 (VGG-16); Mel Frequency Cepstral Coefficients (MFCC); Classifier fusion**

## I. INTRODUCTION

Intelligent transportation systems (ITS) rely on accurate information from vehicle detection and classification systems. Such data can be used for managing traffic flow to control vehicle density. Vehicle density, when uncontrolled, leads to congestion, which affects the free movement of people and goods. There are micro and macro level factors responsible for road congestion [1]. Micro-level factors relate to traffic on the road. Traffic management is of high value to India's road based transportation scenario.

Any sensory root of information is called a *modality* in the case of a human-computer interface. Visual data can be termed as the visual modality. Another important modality is audio. Using information from different modalities for a single application can enhance our understanding about it. This is termed as *complementarity* of the sources of information. This work explores the use of CNNs as the base classifiers in the case of the two modalities.

The idea of utilizing two or more modalities for improved classification performance is not new [2]. Usage of the audio modality in conjunction with visual is expected to increase the overall efficiency of the system due to the property of complementarity. This work is set apart by the fact that both modalities have been put through CNN classifiers, which has not been done before and finally a combined decision-making is performed.

This paper is organised in the following manner. Section 2 deals with the related works. The proposed system is detailed in section 3. The results of experimentation are given in section 4. Finally the conclusions and future works are presented in section 5.

## II. RELATED WORKS

Many researchers have proposed using Background Subtraction to emphasize moving objects and specifically vehicles [3] [4] [5] in an image taken from a video. Moreover, further improvements have been made using the Adaptive Background Subtraction (ABS) algorithm [6] [7] [8] [9]. A technique called blob detection has been used for the purpose of object detection in a frame [9] [10]. The peaks in the Short Time Energy (STE) signal have been used for localization of vehicles with much success [9] [11] [12]. Fig. 2 is the STE of an audio signal from traffic. This will help reduce the computational requirements later in the process. Energy of a signal is defined as,

$$E = \sum_{m=-\infty}^{\infty} x^2[m] \tag{1}$$

And the Short Time Energy as,

$$E_{\hat{n}} = \sum_{m=-\infty}^{\infty} (x[m]w[[\hat{n}] - m])^2 \tag{2}$$

Bay et al. [13] presented an improved feature detector named Speeded Up Robust Feature (SURF). Classifying vehicles based on their audio signatures, or features like Short Time Energy (STE) and Mel-Frequency Cepstral Coefficients (MFCCs) have been explored before [11]. Classification based on MFCCs proves to be the most efficient way, for the audio scenario. Classification using MFCC as the features show high accuracy [14] in many

978-1-7281-9393-9/20/$31.00 ©2020 IEEE
DOI: 10.1109/CICN.2020.86

Figure 1. Sample frames from the input data

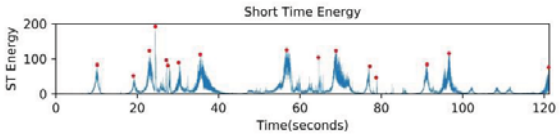


Figure 2. Short Time Energy (STE) of an audio signal

instances. Classification of vehicles based on multiple classifiers, specifically audio and visual have been studied earlier [11] [15]. But the combined usage of CNNs for both visual data and MFCCs of audio signal data was yet to be experimented with. Although, combination of such classifiers has been studied [2] and it has been shown that it can improve the overall classification performance, the works have not dealt with usage of purely, CNN classifiers. Also late-stage fusion or specifically decision fusion has not been performed on such a multi-modal CNN based setup.

## III. SYSTEM DESCRIPTION

The system performs multi-modal classification using the modalities of visual and audio to improve the classification performance. Vehicles are classified into 4 classes, viz. 2-wheeler, 3-wheeler, 4-wheeler light-motor-vehicle and heavy vehicle.

### *A. Input Data*

A high frame-rate video (50 frames per second) is taken as the input. The video signal used is acquired from a camera mounted on a stable platform. The camera is faced approximately 45 degrees away from the imaginary road-parallel, into the traffic. A data acquisition vehicle can be easily stationed to obtain data from a road or bridge of interest within a matter of hours. The video signal is encoded by the camera in 'MPEG-4 Part 14' format. The signal is preferred in the maximum possible resolution keeping in mind the computational overheads that follow. It should be of at least a minimum resolution that does not warrant any reconstruction [16] or conditioning for obtaining good results. Each sample/image in the signal exists for 20 milliseconds (ms) in real-time. 'FFmpeg' is a popular framework for video processing. It is used here to extract an audio signal from the MPEG-4 file and is saved as a single-channel audio file utilizing the 'pcm s16le' audio codec. Examples of video frames are given in Fig. 1.

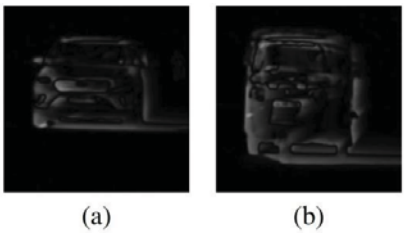


Figure 3 Foreground images from ABS (a) 4 wheeler (b) 3 wheeler

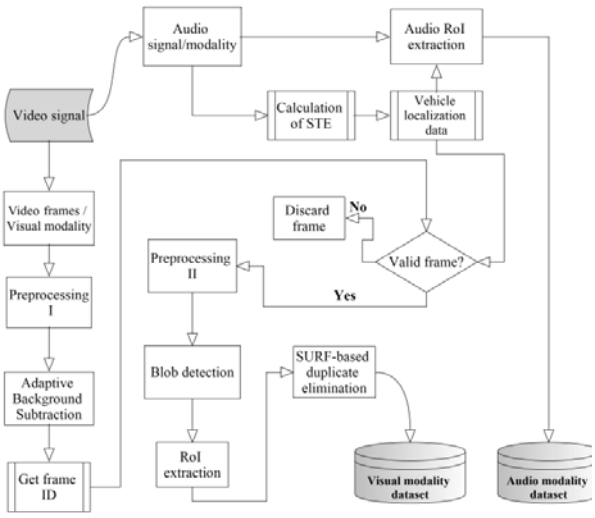


Figure 4. Pre-processing the input data

### *B. Classification using visual modality*

#### *1) Pre-processing*

*a) Localization and dataset generation:* Consider a frame $X_{if}$ at an instant of $i$ milliseconds and having frame identification number $f$. The frame is then processed using the ABS algorithm to obtain a foreground image $F_g$. Multiple morphological operations are performed on $F_g$. The binary image obtained after these operations indicates the moving vehicles in the scene as groups of white pixels. The processed frame is passed through a 'Blob detection' algorithm. This is to accurately locate the vehicles in the frame. A blob is also defined as a 'Binary Large Object' [10]. Once the blobs are detected, the frame identification number is also matched to data from audio based localization to reduce the computational requirements. The vehicle images are then extracted. Since there maybe multiple images of a given vehicle, they are then passed through a feature matching algorithm. The algorithm used here is Speeded Up Robust Features or SURF [13]. The dataset to be generated only needs one image per vehicle. Only an image with the most key points [13] is chosen. The processed images are saved for training/classification. The images are manually labelled into 4 classes as specified earlier and saved as a dataset.


Figure 5. Image processing stages: (a) Original, (b) ABS output (c) Morphological operations (d) Blob detection

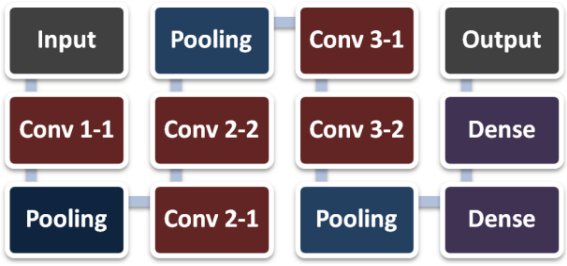


Figure 6. A representation of the smallerVGGNet network layers

## C. The Classifier

Classification of the saved images containing the vehicles is performed using a Convolutional Neural Network (CNN). The classifiers output a set of probability values which indicates the appropriate assignment of the sample to any one of the 4 classes (2-wheeler, 3-wheeler, 4-wheeler light-motor-vehicle and heavy-vehicle).

*a) VGG-16:* Convolutional Neural Networks such as VGG-16 [17] have high accuracy when it comes to image classification. These networks have been pre-trained using millions of image from the 'ImageNet' dataset. The VGG-16 network (Fig. 7) therefore is used here in the Transfer Learning mode.

*b) SmallerVGGNet:* A reduced version of VGG16 is also utilized without Transfer Learning, and is trained on the previously generated dataset. It is called the smallerVGGNet (sVGGNet). This reduced version has 5 convolutional layers (Fig. 6). This is suitable for training on small datasets because of its shallowness.

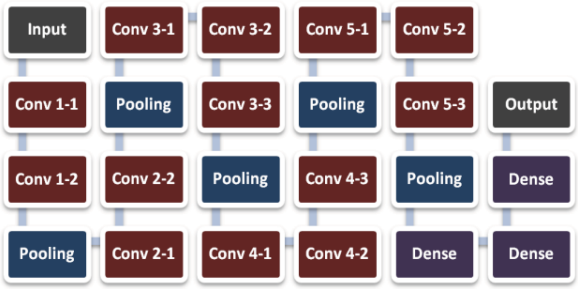


Figure 7. A representation of the VGG-16 network layers

## D. Classification using audio modality

### 1) Pre-processing:

*a) Localization:* Audio track extracted from the video is put through an algorithm to find the Short Time Energy (STE). The vehicles indicate their presence by the peaks in the STE values [12]. This is used to locate the vehicle. This information is used to reduce computational costs in the system.

*b) Dataset generation:* Audio signal extracted from the video is processed further to create a dataset of audio snippets for the purpose of classification based on the audio modality. The extracted snippets are of 2 seconds duration. The MFCC of these around the STE peaks are extracted in classification. The clips of audio are labelled into the 4 classes specified earlier.

## E. The Classifier

Classification of the audio signals is performed after extracting the Mel-Frequency Cepstral Coefficients (MFCCs). The Mel-scale mimics the non-linear human-ear perception of sound. The frequency in Hertz ($f$) is related to the Mel-frequency ($m$) by,

$$m = 2595 * log_{10}\left(1 + \frac{f}{700}\right) \quad (3)$$

The classifier used in this case is a CNN with 14 layers, of which 4 are convolutional layers. The audio signals corresponding to each vehicle is converted to an MFCC array and reshaped into a 2-dimensional array for passing it through the CNN. The classification result is a decision vector just like in the case of the visual modality. This vector is used in decision-level fusion.

## F. Multi-modal classification

*a) Decision fusion/Late fusion:* The final part of this work is to combine the results from both the audio and visual modalities' classification to arrive at a combined result. Let the prediction vector of the base classifiers be

$$C_i(x) = \left[d_{i,1}(x) \ \ldots \ d_{i,c}(x)\right]^T \quad (4)$$

For 4 classes, it becomes,

$$C_i(x) = \left[d_{i,1}(x) \ \ d_{i,2}(x) \ \ d_{i,3}(x) \ \ d_{\ i,4}(x)\right]^T \quad (5)$$

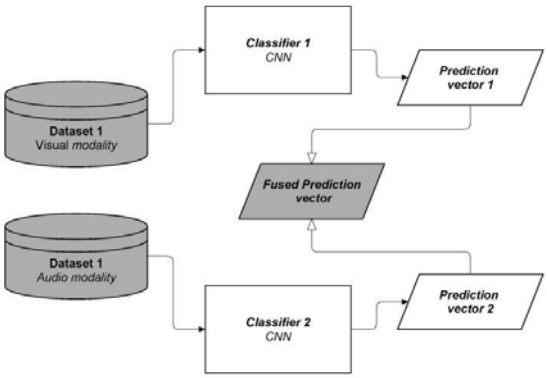


Figure 8. Decision-level classifier fusion

where,

$$C = (C_1, C_2) \tag{6}$$

are the two classifiers.

$d_{i,j}(x)$ is the degree of support given by classifier $C_i$ that $x$ comes from class$j$. The outputs of the two classifiers constitute a Decision Profile ($DP(x)$).

$$DP(x) = \begin{bmatrix} d_{1,1}(x) & d_{1,2}(x) & d_{1,3}(x) & d_{1,4}(x) \\ d_{2,1}(x) & d_{2,2}(x) & d_{2,3}(x) & d_{2,4}(x) \end{bmatrix} \tag{7}$$

The final Prediction vector can be derived from the above Decision Profile by aggregating its columns. This aggregation can take the form of summation, product, averaging, etc. A sample is assigned with the class in which the prediction vector gives the maximum score.

## IV. EXPERIMENTATION AND RESULTS

### A. Image and audio processing

The video signal is pre-processed in the audio and visual modality using various open source libraries in Python 3.7.8. Each vehicle is represented by a 2 second audio extracted around each STE peak. The Mel-Frequency Cepstral Coefficients (MFCC) are calculated for audio classification.

### B. Classification

In the dataset there are 477 samples in total, of which 384 are designated for training the classifier, in the case of both visual and audio modality classification. The visual modality was used as input to the sVGGNet and VGG-16 CNNs, with the VGG-16 in Transfer Learning mode. The audio modality was also put through the classification process using a CNN. The results of the individual base classifiers are depicted in the Table I. For the visual modality classification, from Table I, it is clear that the sVGGNet and the VGG-16 network gave very good and close results. From Table II, it is evident that the result of decision-level fusion, using VGG-16 and sVGGNet for visual modality, and a CNN for audio modality, is superior compared to the case of single-modality classification.

TABLE I
SINGLE-MODALITY CLASSIFICATION PERFORMANCE

| Modality | Method | Average Accuracy | f1 score |
|---|---|---|---|
| Visual | sVGGNet | 89.54% | 0.88 |
| | VGG-16 (Transfer Learning) | **90.32%** | **0.92** |
| Audio | CNN | 61.29% | 0.61 |

TABLE II
MULTI-MODAL CLASSIFICATION PERFORMANCE

| Case | Modality | Classifier | Fusion results | |
|---|---|---|---|---|
| | | | *Average Accuracy* | *f1 score* |
| 1 | Visual | sVGGNet | 86.84% | 0.86 |
| | Audio | CNN | | |
| 2 | Visual | VGG-16 (Transfer Learning) | **91.40%** | **0.92** |
| | Audio | CNN | | |

### C. Decision-level fusion

The results from the classifiers in both the modalities were combined using two methods. First, the two predictions were multiplied with each other. This resulted in a new prediction vector. The performance metrics of which was calculated. In the next method, the predictions resulting from both the modalities were combined using weighted sum. Among these only the product method produced improvements over the single-modality case.

The performance of a classification system can be measured using the Confusion matrix. It measures the number of correct predictions and provides an intuitive graphical interpretation of the classifier's performance. In Fig. 9, the most accurate vehicle classification system achieved here is visualized, which is the result of decision-level fusion of the base classifiers. Therefore, compared to the results from the single-modality vehicle classification, the multi-modal method with decision-level fusion provides superior accuracy.

## V. CONCLUSION

Vehicle classification systems are in need for a multitude of applications like road infrastructure planning, traffic management, law enforcement, etc. This work was aimed at devising a system of vehicle classification using multiple modalities of data for improved classification performance. Based on the experiments, out of the two methods of decision level fusion for CNN based multi-modal classification, the product method gave the best results.

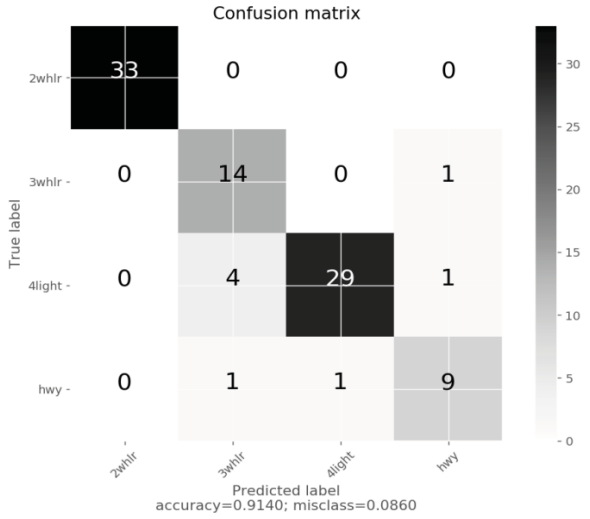


Figure 9. Results of Multi-modal decision fusion with VGG16

Even though the classification performance is noteworthy while using a custom dataset, further improvements can be made by, increasing the size of the dataset, utilizing night-time data and using a superior decision fusion technique.

## VI. ACKNOWLEDGEMENTS

The authors are thankful to members of the Centre for Advanced Signal Processing (CASP) Lab at the Department of Electronics and Communication Engineering of Rajiv Gandhi Institute of Technology, Kottayam, Kerala, India, for their assistance and support towards the completion of this work.

## VII. REFERENCES

[1] A. M. Rao and K. R. Rao, "Measuring urban traffic congestion – a review," *International Journal for Traffic and Transport Engineering*, 286–305, Dec. 2012. [Online]. Available: https://doi.org/10.7708/ijtte.2012.2(4).01

[2] B. Selbes and M. Sert, "Multimodal vehicle type classification using convolutional neural network and statistical representations of mfcc," in 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2017, pp. 1–6

[3] H. Unno, K. Ojima, K. Hayashibe, and H. Saji, "Vehicle motion tracking using symmetry of vehicle and background subtraction," in 2007 IEEE Intelligent Vehicles Symposium, 2007, pp. 1127-1131.

[4] H. Zhang and K. Wu, "A vehicle detection algorithm based on threeframe differencing and background subtraction," in 2012 Fifth International Symposium on Computational Intelligence and Design, vol. 1, 2012, pp. 148–151.

[5] N. Seenouvong, U. Watchareeruetai, C. Nuthong, K. Khongsomboon, and N. Ohnishi, "A computer vision based vehicle detection and counting system," in 2016 8th International Conference on Knowledge and Smart Technology (KST), 2016, pp. 224–227.

[6] N. S. Sakpal and M. Sabnis, "Adaptive background subtraction in images," in 2018 International Conference On Advances in Communication and Computing Technology (ICACCT), 2018, pp. 439–444.

[7] E. Komagal, A. Vinodhini, Archana, and Bricilla, "Real time background subtraction techniques for detection of moving objects in video surveillance system," in 2012 International Conference on Computing, Communication and Applications, 2012, pp. 1–5.

[8] C.-L. Huang and H.-N. Ma, "A moving object detection algorithm for vehicle localization," in 2012 Sixth International Conference on Genetic and Evolutionary Computing. IEEE, Aug. 2012. [Online].Available: https://doi.org/10.1109/icgec.2012.23

[9] D. Renganathan, L. Mary, and A. George, "Detection and classification of vehicles from heterogeneous traffic video data collected using a probe vehicle," in 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information Communication Technology (RTEICT), 2018, pp. 1979–1982.

[10] T. Jia, N. liang Sun, and M. yong Cao, "Moving object detection based on blob analysis," in 2008 IEEE International Conference on Automation and Logistics. IEEE, Sep. 2008. [Online]. Available:https://doi.org/10.1109/ical.2008.4636168

[11] C. Daniel and L. Mary, "Fusion of audio visual cues for vehicle classification," in 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), 2016, pp. 1–4.

[12] S. Anuja Prasad and L. Mary, "A comparative study of different features for vehicle classification," in 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), 2019, pp. 1–5.

[13] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool], "Speeded-up robust features (surf)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346 – 359, 2008, similarity Matching in Computer Vision and Multimedia.

[14] Ziyou Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Comparing mfcc and mpeg-7 audio features for feature extraction, maximum likelihood hmm and entropic prior hmm for sports audio classification," in 2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698), vol. 3, 2003, pp. III–397.

[15] Piyush P., R. Rajan, L. Mary, and B. I. Koshy, "Vehicle detection and classification using audio-visual cues," in 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), 2016, pp. 726–730.

[16] Y. Pathak, K. V. Arya, and S. Tiwari, "An efficient low-dose CT reconstruction technique using partial derivatives based guided image filter," Multimedia Tools and Applications, vol. 78, no. 11, pp. 14 733– Nov. 2018. [Online]. Available: https://doi.org/10.1007/s11042-018-6840-5

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.1556