Vehicle Classification Based on CCTV Video Recording Using Histogram of Oriented Gradients, Local Binary Patterns, and Hierarchical Multi-SVM

D A Navastara¹, M F Maulana², N Suciati³, and S T Sarena⁴

¹,²,³Department of Informatics, Faculty of Intelligent Electrical and Informatics Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
⁴Marine Electrical Engineering, Ship Building Institute of Polytechnic Surabaya, Indonesia

E-mail: ¹dini_navastara@if.its.ac.id, ²farhanm.official@gmail.com, ³nanik@if.its.ac.id, ⁴sryang.tera@ppns.ac.id

Abstract. Vehicle classification is a crucial element in a dynamic traffic management system. The system is needed since the number of vehicles gradually increase every year. There are many methods for vehicle classification, one of those is the use of CCTV camera. As it is multifunction and cheaper than other types of camera, the government tend to use the CCTV camera as monitoring equipment of vehicles on the road. Hence, an accurate vehicle classification system from CCTV records is needed for traffic management system as well as other systems. In this study, the method contains four steps: object detection using Gaussian Mixture Model, features extraction using Histogram of Oriented Gradients and Local Binary Patterns, model training and classification using Hierarchical Multi-SVM. The experimental results show that our proposed method works well with the accuracy of 80.28%, precision of 94.27%, recall of 73.79%, and F-measure of 82.76%.

1. Introduction

The level of density of road users in Indonesia can be categorized as crowded. The number of motorized vehicles in Indonesia always increases from year to year. The increase in the number of motor vehicles is motivated by the increasing number of Indonesia's population, which results in an increase in transportation needs in the daily life of the community. The last census in 2010 set Indonesia's population as 237,641,326 people, with an average increase of 1.54% per year [1]. As for the number of motorized vehicles, at the last calculation, there were 13,480,973 passenger car vehicles, 2,420,917 bus cars, 6,611,028 freight cars, and 98,881,267 motorbikes [2]. Data from the Central Statistics Agency shows an increase in the number of vehicles each year. Meanwhile, the growth of new roads cannot keep pace with the rate of increase in the number of vehicles each year. Therefore, we need a solution from the field of information technology in order to be able to analyze conditions on the highway.

There have been many research topics that prioritize the application and application of technology in transportation systems. One topic that is currently hot is the classification of vehicle types from CCTV recordings installed on the side of the road. This classification is very useful in daily life, such as helping to recognize the majority of vehicle types in certain areas, and enforcement of road regulations such as prohibiting road passes for certain types of vehicles. Some existing studies use CCTV recordings to detect vehicles, but only limited between "vehicles" and "non-vehicles" [3] [4]. Other studies have been...
able to classify more than one type of vehicle but are still weak against changes in vehicle direction [5] and tolerance to lighting [6] [7].

In this study, the process of classification of motorized vehicles in the daytime with sunny conditions. This classification process uses four stages. The first stage, object detection, is the stage for separating objects and backgrounds using the Gaussian Mixture Model, Median Filter, dilation, and erosion. The second stage, feature extraction is the stage for obtaining motor vehicle features using Histograms of Oriented Gradients and Local Binary Patterns. The third stage, the training model is the stage for creating a classification model with data from the feature extraction stage using the Support Vector Machine. The fourth stage, classification, uses Hierarchical Multi-SVM, a variation of how to implement the Support Vector Machine, to classify between types of vehicles.

2. Our Approach

2.1. Dataset

The data used in the form of two CCTV video footage on one of the highway sections with a size of 640 × 360 pixels and has an RGB image channel taken from the Youtube video sharing site [16]. Video 1 is 81 seconds long, with a total of 2030 frames. Video 2 is 55 seconds long, with a total of 1640 frames. Examples of recording frames used are shown in Figure 1, where the left frame is an example frame in Video 1, and the right frame is an example frame in Video 2.

From both videos, the object detection process will be carried out to obtain training data and testing data. The training data used are positive samples and negative samples on the results of the detection stage of randomly selected objects. Positive samples are motorized bus_truck, car, minivan, and motorcycle types. Whereas negative samples are non-motorized imagery, where non-motorized imagery can be in the form of streets, trees, buildings, and alike. Negative samples can also be in the form of images that have more than one motorized vehicle in them. Because the data used does not have a ground truth, the process of grouping training data into positive samples and negative samples is done manually. The results of the detection phase of objects not selected as training data will be used as testing data. Positive and negative samples for dimensions 64x64 pixels amounted to 3840 samples, consisting of 960 bus_truck positive samples, 960 positive car samples, 960 positive minivan samples, and 960 negative samples. While positive and negative samples for the dimensions of 32x64 pixels amounted to 720 samples, consisting of 360 positive motorcycle samples and 360 negative samples. The amount above is determined according to the ratio of 8:2 to class with the smallest number of samples.

2.2. System Design

The processes involved in implementing this system include the detection of motorized vehicle objects, the feature extraction stage, the training model stage, and the classification stage. The flowchart of the whole system process can be seen in Figure 2.

2.2.1. Object Detection. The first stage is the detection of vehicle objects. This stage uses the Gaussian Mixture Model (GMM) to separate vehicle objects from the background. The use of GMM was chosen because of its ability to distinguish lighting changes [8]. The results of GMM in the form of binary
images that still have some noise that has the potential to interfere with the detection results. For this reason, several filtering and morphological operations such as median filter, dilation, and erosion are performed to get better detection results. The median filter is useful for removing salt and pepper noise [9], and dilation and erosion are used to refine the image by connecting the fragments and eliminating small holes in the image [9]. The next step is cropping and resizing to get object segments as vehicle candidates based on GMM result image as shown in Figure 3. Because the feature extraction stage uses Histograms of Oriented Gradients (HOG) that depend on image size, and non-uniform cropping results, resizing is needed so that the size becomes uniform.

Figure 2. System Design

![Figure 2. System Design](image)

Figure 3. Original Image (top), result image of Gaussian Mixture Model (bottom)

The sizes used in the implementation of this stage are 64x64 pixels and 32x64 pixels, to adjust to the image of 4-wheeled vehicle segments that tend to be square and the image of 2-wheeled vehicle segments which tend to be rectangular in portrait orientation. From this stage, there will be some areas that have potential as vehicle objects. Figure 4 shows a positive sample resulting from this stage, and Figure 5 for a negative sample.

2.2.2. Feature Extraction. The feature extraction that used in this paper are Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBP). HOG is one method feature extraction commonly
used [10] - [12]. It counts the number of occurrences of the gradient orientation of a particular part of an image. This method is calculated from the grid of uniformly spaced cells and uses normalization of overlapping local contrast to improve accuracy [13]. Whereas LBP is one of the image feature extraction for classification. LBP calculates the pixel value of the image by thresholding neighbourhood pixels and storing the result as a binary number.

At this stage, the best HOG configuration will be searched by giving a variation of values in pixel configuration per cell and cell per block of HOG. The results of this stage are features that are ready to be used for the training model stage. Visualization of the results of the extraction of HOG and LBP features can be seen in Table 1.

![Image](image_url)

**Figure 4.** Example cropping positive samples. (a) 64x64 pixels, (b) 32x64 pixels

![Image](image_url)

**Figure 5.** Example cropping negative samples. (a) 64x64 pixels, (b) 32x64 pixels

2.2.3. **Training Model.** The training model stage is the stage for creating a classification model based on the features of the training data. This stage uses the Support Vector Machine (SVM) with a one-versus-rest configuration to produce a classification model that will be used at a later stage [14]. At this stage, the results of feature extraction will be classified into two classification models, namely for the effects of feature extraction from 64x64 pixel data and for feature extraction results from 32x64 pixel data. The 64x64 pixel classification model has four classes, namely bus_truck, car, minivan, and other_n. As for the classification model, 32x64 pixels have two classes, namely motorcycle, and other_m.
2.2.4. Classification. The last stage, the classification stage is the stage for classifying vehicle objects into prediction classes according to the classification model produced at the training model stage. Members of vehicle object prediction classes include Bus_Truck, Car, Minivan, and Motorcycle. This stage uses Hierarchical Multi-SVM, where this method is used to overcome heavy traffic conditions with vehicles that are side by side [15]. This method works on the principle that the prediction results of vehicle types do not have to be on the first try. At this stage, two SVM hierarchies are used, namely first-order SVM and second-order SVM. The results of this stage are vehicle objects that have been labeled with classes according to the predicted results. The flowchart of classification process using Hierarchical Multi-SVM can be seen in Figure 6.

| Class        | Sample          | HOG          | LBP          |
|--------------|-----------------|--------------|--------------|
| Bus/Truck    | ![Bus/Truck](image) | ![HOG](image) | ![LBP](image) |
| Car          | ![Car](image)   | ![HOG](image) | ![LBP](image) |
| Minivan      | ![Minivan](image) | ![HOG](image) | ![LBP](image) |
| Motor Cycle  | ![Motor Cycle](image) | ![HOG](image) | ![LBP](image) |
| Negative 64x64 Piksel | ![Negative 64x64](image) | ![HOG](image) | ![LBP](image) |
| Negative 32x64 Piksel | ![Negative 32x64](image) | ![HOG](image) | ![LBP](image) |

3. Result and Discussion
In the experiment, some scenarios have been applied to obtain the best model of our proposed method. The initial values of parameters should be set, such as the LBP type used is LBP uniform 3x3 pixels. In hierarchical multi-SVM, we set the kernel as RBF and the regularization value, C=1.

3.1. Pixel Cell HOG Test per Cell
In this scenario, a feature extraction trial is performed using the Histogram of Oriented Gradients (HOG). In this scenario, the pixel value per cell is used as an independent variable, and cells per block are fixed variables. The pixel values per cell used are 4x4, 8x8, and 16x16. The cell value per block as a fixed variable is 2x2, according to the HOG standard configuration, trials using training data obtained
from object detection results. Then the accuracy calculation process is done using 10-fold Cross-Validation. Table of trial results can be seen in Table 2.

**Figure 6.** flowchart of classification process using Hierarchical Multi-SVM

**Table 2.** HOG pixel per cell trial results

| n-Fold | Accuracy of Pixel per Cell (%) | 64x64 Pixels | 32x64 Pixels |
|--------|--------------------------------|--------------|--------------|
|        | 4x4 | 8x8 | 16x16 | 4x4 | 8x8 | 16x16 |
| 1      | 83.33 | 92.71 | 92.19 | 95.83 | 95.83 | 98.61 |
| 2      | 90.36 | 91.41 | 91.41 | 80.56 | 79.17 | 100  |
| 3      | 80.47 | 87.76 | 93.49 | 98.61 | 100  | 94.44 |
| 4      | 89.32 | 92.19 | 94.27 | 100  | 100  | 100  |
| 5      | 92.97 | 94.01 | 96.61 | 91.67 | 97.22 | 97.22 |
| 6      | 89.84 | 91.93 | 92.19 | 86.11 | 95.83 | 94.44 |
| 7      | 87.50 | 90.36 | 90.62 | 76.39 | 94.44 | 98.61 |
| 8      | 89.32 | 90.62 | 87.24 | 95.83 | 94.44 | 98.61 |
| 9      | 91.41 | 91.93 | 89.32 | 94.44 | 98.61 | 98.61 |
| 10     | 83.07 | 83.85 | 82.29 | 98.61 | 98.61 | 98.61 |
| **Average** | **87.76** | **90.68** | **90.96** | **91.81** | **95.42** | **97.92** |

3.2. **Trial HOG Cells per Block**

In this scenario, a feature extraction trial is performed using the Histogram of Oriented Gradients. In this scenario, the value of cells per block is used as an independent variable, and pixels per cell are fixed variables. Cell values per block used are 1x1, 2x2, and 4x4. The selection of cell values per block is based on factors on the dimensions of the object detection results, namely 32x64 and 64x64. Then the accuracy calculation process is done using 10-fold Cross-Validation. The table of trial results can be seen in Table 3.

3.3. **Comparative Trials of HOG, LBP and HOG + LBP**

In this scenario, a comparison of the results of the trial feature extraction of vehicles using Histogram of Oriented Gradients, Local Binary Patterns, and a combination of Histogram of Oriented Gradients and Local Binary Patterns. Trials using training data obtained from object detection results. Then the accuracy calculation process is done using 10-fold Cross-Validation. Table of trial results can be seen in Table 4.
3.4. Discussion

From the first and second test scenarios, it is found that the best accuracy value for the extraction of HOG features is the configuration of 16x16 pixels per cell and 2x2 cell per block, with values of 90.96% for dimensions of 64x64 pixels and 97.92% for dimensions of 32x64 pixels. However, after the third test scenario, it was found that the accuracy value for the 64x64 pixel dimension decreased after combined with LBP to 90.68%. In contrast, the configuration of 8x8 pixels per cell has increased accuracy from 90.68% to 90.99%. For this reason, it was concluded that HOG + LBP with an arrangement of 8x8 pixels per cell is the most optimal configuration for extracting object features with dimensions of 64x64 pixels.

Table 3. HOG cell test results per block

| n-Fold | Accuracy of Cell per Block (%) | 64x64 Pixels | 32x64 Pixels |
|--------|-----------------|--------------|-------------|
|        | 1x1  | 2x2  | 4x4  | 1x1  | 2x2  | 4x4  |
| 1      | 87.76| 92.71| 94.27| 95.83| 95.83| 95.83|
| 2      | 92.19| 91.41| 91.41| 77.78| 79.17| 87.50|
| 3      | 86.20| 87.76| 86.98| 98.61| 100  | 100  |
| 4      | 90.88| 92.19| 93.49| 98.61| 100  | 100  |
| 5      | 92.71| 94.01| 91.93| 98.61| 97.22| 100  |
| 6      | 88.28| 91.93| 91.67| 90.28| 95.83| 97.22|
| 7      | 88.28| 90.36| 87.76| 91.67| 94.44| 84.72|
| 8      | 90.36| 90.62| 89.06| 98.61| 94.44| 97.22|
| 9      | 92.97| 91.93| 91.67| 98.61| 98.61| 97.22|
| 10     | 83.30| 83.85| 83.85| 98.61| 98.61| 100  |
| Average| 89.30| 90.68| 90.21| 94.72| 95.42| 95.97|

Table 4. Test results of feature extraction comparisons

| Feature Extraction | Accuracy Average (%) | 64x64 Pixels | 32x64 Pixels |
|--------------------|----------------------|--------------|-------------|
| LBP                | 78.80                | 94.86        |
| HOG 4x4 pixel per cell | 87.76                | 91.81        |
| HOG 8x8 pixel per cell | 90.68                | 95.42        |
| HOG 16x16 pixel per cell | 90.96                | 97.92        |
| HOG 4x4 pixel per cell + LBP | 87.53                | 94.72        |
| HOG 8x8 pixel per cell + LBP | **90.99**            | **98.33**    |
| HOG 16x16 pixel per cell + LBP | 90.68                | 97.92        |

Table 5. System performance

| Video   | Accuracy (%) | Precision (%) | Recall (%) | F-measure (%) |
|---------|--------------|---------------|------------|---------------|
| Video 1 | 82.60        | 92.14         | 74.46      | 82.36         |
| Video 2 | 77.96        | 96.39         | 73.11      | 83.15         |
| Average | **80.28**    | **94.27**     | **73.79**  | **82.76**     |

Whereas for object dimensions of 32x64 pixels, the best accuracy values occur after the third test scenario, where the HOG + LBP configuration of 8x8 pixels per cell and HOG + LBP 4x4 cells per block both have an accuracy of 98.33%. So that the configuration used is the same as the configuration for the 64x64 pixel dimension, HOG + LBP with 8x8 pixel configuration is chosen for the extraction of 32x64 pixel object dimensions. Based on Table 5, the system performance shows the proposed method
works quite well with the accuracy, precision, recall, and F-measure of 80.28%, 94.27%, 73.79%, and 82.76%, respectively.

Examples of differences in detection results between Hierarchical Multi-SVM and standard SVM in heavy traffic can be seen in Figure 7, with the top frame using standard SVM, and the bottom frame using Hierarchical Multi-SVM. In the picture, it can be seen that there are two adjoining cars that are unable to be classified by standard SVM but can be sorted by Hierarchical Multi-SVM. Figure 8 shows the example of vehicle type misclassification. The car (red circle) misclassified as minivan because the characteristic of that car similar to minivan.

![Figure 7. Differences in the results of SVM detection (top) and Hierarchical Multi-SVM in heavy traffic (bottom)](image)

![Figure 8. Example of vehicle type misclassification](image)

4. Conclusion
Vehicle classification systems can be implemented using the Gaussian Mixture Model for object detection, Histogram of Oriented Gradients and Local Binary Patterns for feature extraction, and
Hierarchical Multi-SVM for classification. The best configuration for this system is with a distribution value of 8x8 pixels per cell for HOG, 2x2 cells per block for HOG, and LBP uniform 3x3 pixels. Based on the experimental results, our proposed method gives quite good performance with accuracy, precision, recall, and F-measure of 80.28%, 94.27%, 73.79%, 82.76%, respectively.

In the future, this system can be developed by extracting highways and tracking vehicles to improve vehicle detection results, as well as improving work efficiency so that the operation can do real-time classification.

Acknowledgment
This study was funded by Directorate of Research and Community Service (Direktorat Riset dan Pengabdian Masyarakat, DRPM) Institut Teknologi Sepuluh Nopemar (ITS) Surabaya with the grant number of 1665/PKS/ITS/2020. The source code is provided in a repository at https://github.com/mumulmaulana/VehicleClassification.

References
[1] BPS 2010 Jumlah dan Distribusi Penduduk Retrieved on 16 April 2018 from http://sp2010.bps.go.id/index.php/site/topik?kid=1&kategori=Jumlah-dan-Distribusi-Penduduk
[2] BPS 2016 Perkembangan Jumlah Kendaraan Bermotor Menurut Jenis 1949-2018 Retrieved on 16 May 2018 from http://www.bps.go.id/linkTableDinamis/view/id/1133
[3] Saiful A 2016 Deteksi Mobil dari Citra CCTV Menggunakan Fuzzy Morphology dan Gradient Boosting thesis (Surabaya: Teknik Informatika ITS)
[4] Prasetya I 2016 Deteksi Kendaraan Bermotor dari Citra CCTV Menggunakan Gray Level Co-Ocurrence Matrix dan Random Forest (Surabaya: Teknik Informatika ITS)
[5] Dong Z, Wu Y, Pei M and Jia Y 2015 Vehicle type classification using a semisupervised convolutional neural network IEEE Trans. Intell. Transp. Syst. 16 no 4 pp 2247–56
[6] Hsieh J, Yu S and Chen Y 2006 Automatic traffic surveillance system for vehicle tracking and classification IEEE Trans. Intell. Transp. Syst. 7 no 2 pp 175–87
[7] Xue X and Ling H 2011 Robust visual tracking and vehicle classification via sparse representation IEEE Transactions on Pattern Analysis and Machine Intelligences 33 no 11 pp 2259–72
[8] Stauffer C and Grimson W 1999 Adaptive background mixture models for real-time tracking IEEE Computer Society Conf. on Computer Vision and Pattern Recognition pp 246–52
[9] Gonzales R C and Woods R E 2008 Digital Image Processing 3rd edition (New Jersey: Pearson Prentice Hall)
[10] Dalal N and Triggs B 2005 Histograms of oriented gradients for human detection IEEE Computer Society Conf. on Computer Vision and Pattern Recognition pp 886–93
[11] Yan G, Yu M, Yu Y and Fan L 2016 Real-time vehicle detection using histograms of oriented gradients and AdaBoost classification Optik – Int. J. for Light and Electron Optics no 127 pp 7941–51
[12] Wang X, Han T X and Yan S 2009 An HOG-LBP human detector with partial occlusion handling IEEE 12th Int. Conf. on Computer Vision (ICCV) pp 32–39
[13] Navastara D A, Putra W Y M and Faticah C 2020 Drowsiness detection based on facial landmark and uniform local binary pattern J. of Physics: Conf. Series vol 1529 no 5 (Bristol: IOP Publishing)
[14] Faticah C, Alam S P and Navastara D A 2019 Optical flow feature based for fire detection on video data 1st Int. Conf. on Artificial Intelligence and Data Sciences (AiDAS) (New York: IEEE) pp 100–5
[15] Fu H, Ma H, Liu Y and Lu D 2016 A vehicle classification system based on hierarchical multi-SVMs in crowded traffic scenes *Neurocomputing* vol 211 pp 182–90

[16] Supercircuits 2014 *Alibi ALI-IPU3030RV IP Camera Highway Surveillance* Retrieved on 30 October 2020 from https://www.youtube.com/watch?v=PJ5xXXcfuTc