CITS4402 Research Project

Reproduction and Ablation Study of Human Detection using HOG and SVM on the DaimlerChrysler Pedestrian Classification Benchmark Dataset

Liweiwen Zhou
Student ID: 24100792
24100792@student.uwa.edu.au

Abstract

Human detection in static images is a fundamental problem in computer vision with applications in surveillance, autonomous driving, and human-computer interaction. This report presents a reproduction of the influential work by Dalal and Triggs [1], which proposed the use of Histograms of Oriented Gradients (HOG) for feature extraction, combined with a linear Support Vector Machine (SVM) classifier. The implementation follows their original pipeline closely, replicating key architectural decisions, and further incorporates three ablation studies to evaluate the impact of varying HOG parameters such as orientation bins, block size, and normalization schemes. In addition, a user-friendly GUI is developed to visualize model predictions and export results. The reproduction is based on the Daimler Pedestrian Benchmark Dataset [2], from which a balanced and preprocessed training and testing set to support evaluation was constructed.

1. Data Preparation

1.1 Background

This project adopts the Daimler Pedestrian Benchmark Dataset [2], a publicly available collection widely used for evaluating pedestrian classification systems. The dataset consists of grayscale images originally extracted from vehicle-mounted video sequences recorded under varied real-world conditions. Positive examples were generated by manually labeling upright pedestrians and extracting rectangular patches centered on their bounding boxes. To enhance robustness and compensate for minor localization errors, each labeled instance was mirrored and shifted by a few pixels in random directions, producing six variations per annotation. Negative examples were obtained from video frames known to contain no pedestrians by applying a shape-based pedestrian detector and extracting non-matching patches using distance-transformed edge maps.

Each cropped image was padded with a 2-pixel border and scaled to a uniform resolution of 18×36 pixels, preserving edge contour information. The full dataset is split into five disjoint subsets — three for training (labeled "1", "2", "3") and two for testing ("T1", "T2"). This split ensures that no pedestrian appears across multiple sets, preserving evaluation integrity.

1.2 Data Preprocessing and Structure

For this project, only subsets "1", "2" and "T1" were utilized. A balanced training set of 2000 images was formed by sampling 1000 pedestrian and 1000 non-pedestrian examples from folders "1" and "2". The testing set includes 100 positive and 100 negative samples drawn exclusively from "T1". All selected images were resized from their original 18×36 resolution to 128×64, conforming to the input dimensions specified by Dalal and Triggs [1] for HOG-based detection.

To facilitate efficient data management, a standardized directory structure was implemented. Raw samples are stored under `data/` with subfolders `train/pos`, `train/neg`, `test/pos`, and `test/neg`. Downstream Python scripts handle all copying, naming, and resizing procedures. Each image is labeled according to its folder of origin, and metadata is compiled into a tabular format to support later evaluation.

The resulting dataset adheres to assignment specifications: it ensures class balance, separates training and testing images without overlap, and supports reproducible preprocessing through automated scripts.

The figure below outlines the overall structure of the implementation:

```
CITS4402_Project/
|
|-- Others/                # Stores trained models and cached HOG features
|-- data/
|   |-- train/
|   |   |-- pos/           # 1000 positive samples
|   |   `-- neg/           # 1000 negative samples
|   `-- test/
|       |-- pos/           # 100 positive samples
|       `-- neg/           # 100 negative samples
|
|-- Baseline.py            # Original training script using HOG + SVM
|-- evaluate_model.py      # Model evaluation on test set
|-- ablation_bins.py       # Ablation study: orientation bins
|-- ablation_blocksize.py  # Ablation study: block sizes
|-- ablation_norm.py       # Ablation study: normalization techniques
|-- final_model.py         # Training best-performing model
|-- GUI.py                 # PyQt5-based interface for interactive testing
`-- Testing Images/        # Random 10+10 GUI testing samples (converted PNGs)
```

Code: Human Detection project Repo
Data: Daimler Pedestrian Dataset

This modular structure separates core components for training, ablation, testing, and GUI, supporting clarity and ease of replication. A preprocessing script ensures reproducibility by controlling sample selection and format normalization.

2. Optimized HOG Feature Extraction and SVM Training

2.1 Final Feature Extraction Settings

The optimized configuration for HOG feature extraction was selected based on ablation results across multiple parameter groups. The chosen settings are listed in Table 1, which reflect the best-performing combination on the testing set. Gradient features were extracted from raw grayscale images resized to 128×64 pixels, using a standard `[-1, 0, 1]` derivative filter.

Table 1: Optimized HOG Parameters for Final Model

Parameter	**Value**
Detection Window Size	128 × 64 pixels
Gradient Filter	`[-1, 0, 1]`
Orientation Bins	18
Pixels per Cell	8 × 8
Cells per Block	4 × 4
Block Size	32 × 32 pixels
Block Stride	8 pixels
Normalization Scheme	L2-norm
Preprocessing	Raw grayscale (no gamma)

As shown in Figure 1, comparison of baseline, ablation-optimal, and final models across three evaluation metrics, where performance improvements from each ablation are reflected in accuracy, precision, and recall, culminating in the final optimized configuration.

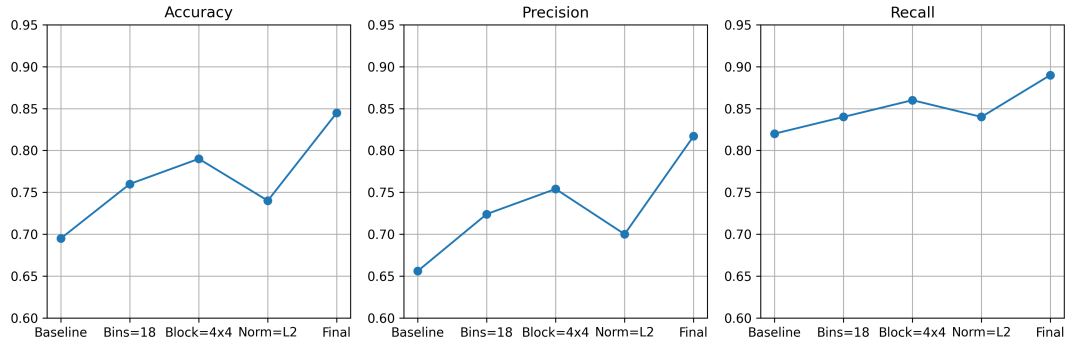

Figure 1: The baseline model uses the original HOG configuration from [1]: 9 orientation bins, 2×2 block size (16×16 pixels), and L2-Hys normalization.

2.2 Training and Evaluation

HOG descriptors were extracted from all 2000 training samples (1000 positive and 1000 negative), resulting in 324-dimensional feature vectors per image. A linear SVM was trained using the `LinearSVC` implementation from the `scikit-learn` library, with default regularization settings. The model achieved an accuracy of 84.5% on a hold-out testing set of 200 images (100 positive and 100 negative), with a precision of 81.7% and recall of 89.0%.

These results demonstrate a significant improvement compared to baseline settings and confirm the efficacy of increased orientation granularity, larger block size, and L2 normalization. The trained model was subsequently deployed in the final GUI interface to support visual testing and prediction export on unseen samples.

3. Ablation Study on HOG Parameters

Ablation studies were conducted to explore how varying individual Histogram of Oriented Gradients (HOG) parameters affect human detection performance. Three key parameters were selected for investigation: the number of orientation bins, block size, and normalization technique. In each study, a single parameter was varied while others were fixed to baseline values, enabling isolated analysis. Performance was assessed on the testing set using accuracy, precision, and recall.

3.1 Orientation Bins Ablation

The first ablation experiment varied the number of orientation bins across values $\{3, 4, 6, 8, \mathbf{9}, 12, 15, 18\}$, where 9 corresponds to the baseline configuration used in [1]. A pseudo-grid-search style loop was used to retrain and evaluate each variant. The visualization results are shown in Figure 2.

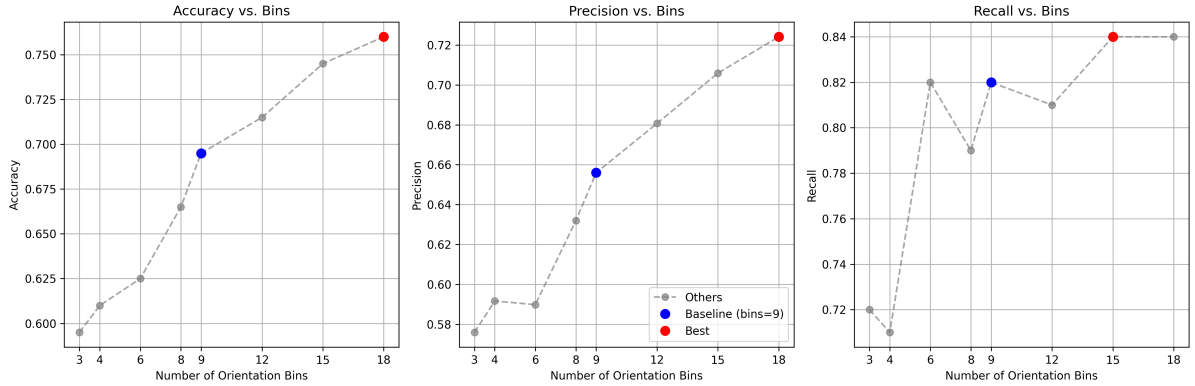

Figure 2: Performance comparison across different HOG orientation bin counts. Blue marks the baseline model (9 bins), red marks the best-performing configuration (18 bins).

As observed, the model using 18 bins achieved the highest accuracy (0.760) and precision (0.7241), with recall matching that of several other configurations. The performance improves steadily from 3 to 18 bins, suggesting that increased angular granularity contributes positively to detection performance.

3.2 Block Size Ablation

Next, the impact of changing the block size was studied. The baseline setting was 2×2 HOG cells per block, with evaluations performed using 3×3 and 4×4 block sizes. Results are summarized in Table 2.

Table 2: Performance across block sizes (cells per block)

Block Size	Accuracy	Precision	Recall
2×2 (Baseline)	0.6950	0.6560	0.8200
3×3	0.7550	0.7217	0.8300
4×4	**0.7900**	**0.7544**	**0.8600**

A clear improvement trend was noted as block size increased. The largest block (4×4) yielded the best results across all metrics. This implies that larger local regions provide more robust gradient summaries.

3.3 Normalization Method Ablation

Lastly, the normalization technique used in HOG feature construction was tested. The baseline uses L2-Hys, while alternatives include L2-norm, L1-norm, and L1-sqrt. Table 3 presents the results.

Table 3: Performance comparison across normalization techniques

Normalization	Accuracy	Precision	Recall
L2-norm	**0.7400**	**0.7000**	0.8400
L1-norm	0.7000	0.6515	**0.8600**
L1-sqrt	0.7350	0.6975	0.8300

The L2-norm normalization produced the highest accuracy and precision, while L1-norm slightly favored recall. Compared to the baseline, all alternatives yielded competitive results, though L2-norm was selected for the final model due to overall balanced performance.

In summary, ablation studies revealed that using 18 orientation bins, 4×4 blocks, and L2 normalization substantially improved classification outcomes over the original settings proposed in [1].

4. Conclusion

This project successfully reproduced the human detection pipeline proposed by Dalal and Triggs, utilizing Histogram of Oriented Gradients (HOG) features and a linear SVM classifier. Through a systematic ablation study framework, I explored how modifications to orientation bins, block sizes, and normalization schemes affect classification accuracy. Among all tested configurations, a setup with 18 orientation bins, a 4×4 block size, and L2 normalization yielded the best performance, achieving an accuracy of 84.5% and recall of 89% on the test set.

To support user interpretation and testing, a lightweight GUI was developed to visualize predictions and export results. The modular design of the training and evaluation pipeline allows easy integration of additional datasets or more advanced classifiers in the future.

Possible directions for future improvement include adopting more robust gradient filters (e.g., Sobel or Scharr), leveraging soft-margin or kernel-based SVMs, and applying data augmentation to expand the diversity of training samples. Additionally, performance on real-world cluttered backgrounds could be enhanced through hard negative mining or fine-tuning on more varied pedestrian datasets. Overall, this reproduction confirms the robustness of HOG-SVM methods and highlights the value of empirical hyperparameter tuning in classic computer vision pipelines.

References

[1] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[2] S. Munder and D.M. Gavrila. An experimental study on pedestrian classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1863–1868, 2006.