COMP61342

Report: Comparative Study between Traditional Computer Vision and Deep Learning Methods

Words: 2379 (Without Titles)

11457839

Introduction:

In recent years, the field of computer vision has rapidly evolved, becoming a cornerstone of modern robotics and artificial intelligence. The ability to interpret and understand visual data not only enriches the interaction between machines and the real world but also enhances the autonomy of robotic systems in diverse applications. From autonomous vehicles to intelligent surveillance, the implications of advanced object recognition technologies are profound and far-reaching.

This report is centrally focused on the exploration and comparison of two distinct approaches to object recognition: traditional computer vision methods and deep learning techniques. Specifically, the project employs Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) methods—both stalwarts in traditional image processing techniques—alongside a more contemporary deep learning approach using Convolutional Neural Networks (CNNs). By applying these methodologies to a benchmark vision dataset, the project aims to systematically evaluate and contrast their performance in recognizing objects.

The goal of this report is to not only develop a nuanced understanding of each technique's operational mechanics but also to assess their practical effectiveness and limitations within the domain of robotics. This comparison will allow for a critical analysis of how traditional and modern approaches can be optimized or combined to enhance object recognition capabilities in robotic systems.

Methods:

Dataset

For this project, we utilized the CIFAR-10 dataset, a well-known benchmark in the field of machine learning and computer vision. Comprising 60,000 32x32 color images evenly distributed across ten different classes, this dataset presents a balanced challenge for object recognition tasks. The classes include everyday objects such as airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks, providing a diverse array of shapes and textures for analysis.

Traditional Computer Vision Methods

SIFT (Scale-Invariant Feature Transform)

The SIFT method is pivotal in extracting distinctive invariant features from images which aid in object recognition, regardless of scaling, rotation, or illumination changes. For each image in the CIFAR-10 dataset, SIFT descriptors were computed to capture key local features. These descriptors not only represent the local appearance of an image but also provide robust matching across varied conditions, making them highly effective for recognizing objects in scenarios where the orientation or scale of the subject may vary.

HOG (Histogram of Oriented Gradients)

The Histogram of Oriented Gradients (HOG) method focuses on the gradient directions of localized portions of an image. By evaluating how local intensities of an image change, HOG captures edge, texture, and gradient structure that are essential for the representation of shape. Each image was processed to extract gradient information, which was then compiled into a descriptor capable of effectively capturing the outline and form of the objects within the CIFAR-10 dataset.

Deep Learning Approach

CNN Architecture

Our deep learning model was constructed using a Convolutional Neural Network (CNN), renowned for its prowess in image classification tasks. The architecture of the CNN used in this project is structured as follows:

Input Layer: Accepts a 32x32x3 image corresponding to the dimensions and three color channels of CIFAR-10 images.
Convolutional Layers: The network includes multiple convolutional layers with 32, 64, and 128 filters of size 3x3, each followed by a ReLU activation function to introduce non-linearity, making the network capable of learning complex patterns in the data.
Pooling Layers: After each set of convolutional layers, a 2x2 max pooling layer reduces the spatial dimensions, helping to decrease computation and control overfitting.
Dropout Layers: Dropout at a rate of 0.2 after each pooling layer and before the final dense layer reduces overfitting by randomly omitting subsets of features during training.

Flatten and Dense Layers: A flatten layer converts the 3D feature maps to 1D feature vectors, followed by dense layers that perform classification. The network ends with a softmax activation layer that classifies the inputs into one of the ten possible classes.

To improve the model's performance, we tuned the hyperparameters. The hyperparameters that were optimized were dropout rate, number of layers, number of epochs, batch size, and optimizer. Various combinations of these hyperparameters were tried, and for each simulation mode's accuracy, it was noted. Based on the highest accuracy among different simulations, the best hyperparameters were chosen. The final hyperparameters used were a
dropout rate of 0.2, 16 layers, 25 epochs, a batch size of 64, the Adam optimizer, and a kernel size of 3x3.

Model: "sequential"

Layer (type)	Output Shape	Param #
conv2d (Conv2D)	(None, 32, 32, 32)	896
conv2d_1 (Conv2D)	(None, 32, 32, 32)	9,248
max_pooling2d (MaxPooling2D)	(None, 16, 16, 32)	0
dropout (Dropout)	(None, 16, 16, 32)	0
conv2d_2 (Conv2D)	(None, 16, 16, 64)	18,496
conv2d_3 (Conv2D)	(None, 16, 16, 64)	36,928
max_pooling2d_1 (MaxPooling2D)	(None, 8, 8, 64)	0
dropout_1 (Dropout)	(None, 8, 8, 64)	0
conv2d_4 (Conv2D)	(None, 8, 8, 128)	73,856
conv2d_5 (Conv2D)	(None, 8, 8, 128)	147,584
max_pooling2d_2 (MaxPooling2D)	(None, 4, 4, 128)	0
dropout_2 (Dropout)	(None, 4, 4, 128)	0
flatten (Flatten)	(None, 2048)	0
dense (Dense)	(None, 128)	262,272
dropout_3 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 10)	1,290

Total params: 550,570 (2.10 MB)

Trainable params: 550,570 (2.10 MB)

Non-trainable params: 0 (0.00 B)

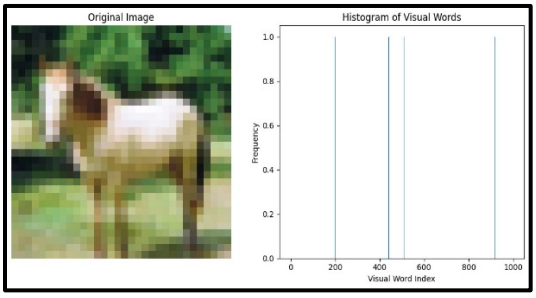
Results:

SIFT Features

The application of SIFT features on the CIFAR-10 dataset resulted in moderate performance metrics across the board. For the validation set, the classification accuracy stood at approximately 30%, with similar figures observed for precision, recall, and F1-score across different classes. The best performance was noted in classes with distinct geometric features and high contrast, such as class '0' (airplanes) and class '9' (trucks), where the precision and recall were slightly higher due to more distinct shapes and boundaries. The test set mirrored these results closely, confirming the consistency of the SIFT method but also highlighting its limitations in handling images with less pronounced features or more complex backgrounds.

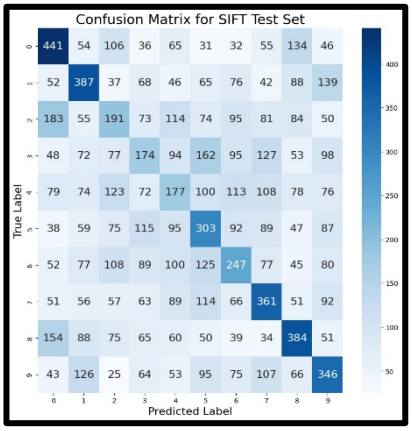
Classification report for SIFT features- TEST set:

	precision	recall	f1-score	support
0	0.39	0.44	0.41	1000
1	0.37	0.39	0.38	1000
2	0.22	0.19	0.20	1000
3	0.21	0.17	0.19	1000
4	0.20	0.18	0.19	1000
5	0.27	0.30	0.29	1000
6	0.27	0.25	0.26	1000
7	0.33	0.36	0.35	1000
8	0.37	0.38	0.38	1000
9	0.32	0.35	0.34	1000
accuracy			0.30	10000
macro avg	0.30	0.30	0.30	10000
weighted avg	0.30	0.30	0.30	10000



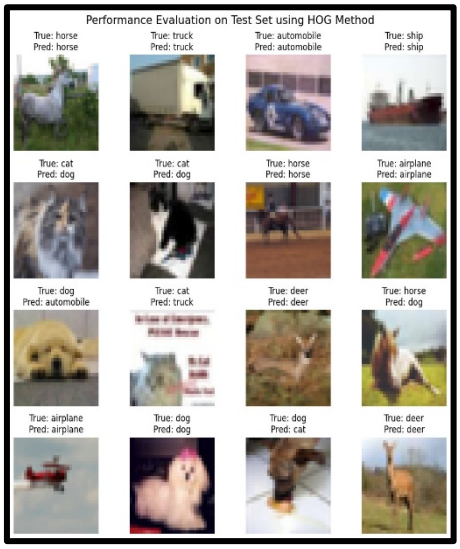
Confusion Matrix Insights (SIFT)

The confusion matrix for SIFT features indicated frequent misclassifications between categories with similar shapes or overlapping features. For example, dogs (class '5') were often confused with cats (class '3'), and trucks (class '9') were occasionally mistaken for automobiles (class '1'). This suggests that while SIFT is robust in capturing key points, its effectiveness is limited in distinguishing between objects with similar outlines or textural properties in a small-scale image setting.



HOG Features

HOG features demonstrated a notable improvement over SIFT, with an overall accuracy of about 62% on both validation and test sets. This method showed a better capability in capturing the essence of object forms and textures, leading to higher precision and recall, particularly for classes like automobiles and trucks, which benefit from the gradient-based feature extraction of HOG. The enhanced performance in these classes can be attributed to the clear structural lines and distinct shapes that are effectively captured by the gradient orientations in the HOG descriptors.



Classification report for HoG features- Validation set:

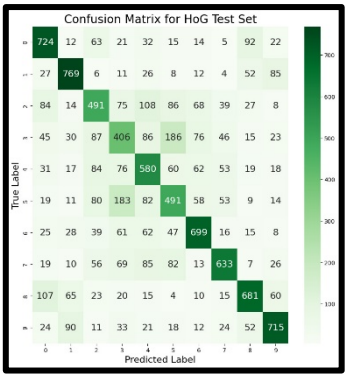
	precision	recall	f1-score	support
0	0.66	0.72	0.69	1000
1	0.74	0.77	0.75	1000
2	0.52	0.49	0.51	1000
3	0.43	0.41	0.42	1000
4	0.53	0.58	0.55	1000
5	0.49	0.49	0.49	1000
6	0.68	0.70	0.69	1000
7	0.71	0.63	0.67	1000
8	0.70	0.68	0.69	1000
9	0.73	0.71	0.72	1000
accuracy			0.62	10000
macro avg	0.62	0.62	0.62	10000
weighted avg	0.62	0.62	0.62	10000

.

.

Confusion Matrix Insights (HOG)

The confusion matrix for HOG features less frequently confused categories compared to SIFT, though challenges remained. There was a noticeable confusion between animals, such as cats and dogs, and between similarly shaped objects like ships and trucks. However, the greater distinction in texture and form processing through HOG led to more accurate classification of objects with unique structural attributes.



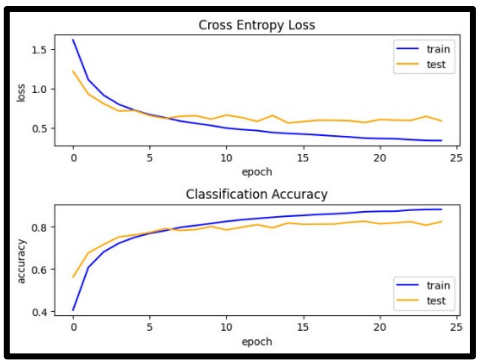
CNN Model

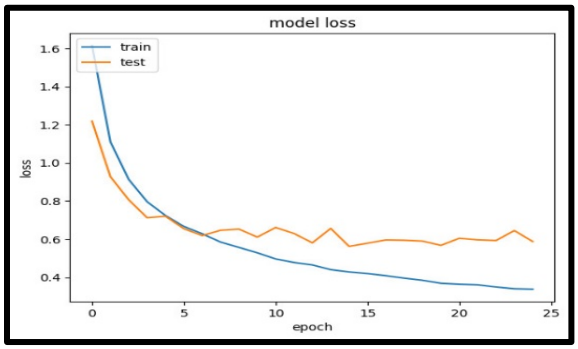
The CNN model exhibited superior performance across all metrics with an overall accuracy of 82% on the test set. The precision, recall, and F1-score significantly outstripped the traditional methods, showcasing the power of deep learning in handling complex image recognition tasks. Classes that traditionally challenge computer vision systems, such as cats and dogs, saw improved results due to the CNN's ability to learn from many trainable parameters and layers specifically designed for image analysis.

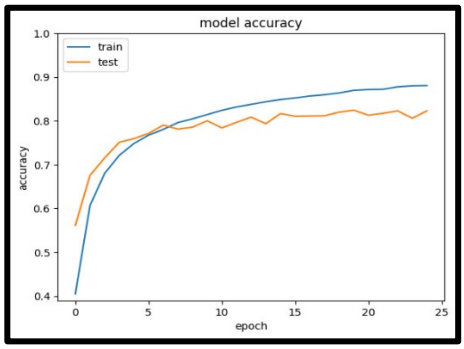
```
...   Classification Report
                precision    recall  f1-score   support

      airplane       0.86      0.85      0.85      1000
    automobile       0.90      0.93      0.92      1000
          bird       0.75      0.74      0.75      1000
           cat       0.65      0.72      0.68      1000
          deer       0.80      0.81      0.80      1000
           dog       0.83      0.66      0.74      1000
          frog       0.78      0.91      0.84      1000
         horse       0.89      0.82      0.86      1000
          ship       0.89      0.91      0.90      1000
         truck       0.91      0.87      0.89      1000

      accuracy                           0.82     10000
     macro avg       0.83      0.82      0.82     10000
  weighted avg       0.83      0.82      0.82     10000
```

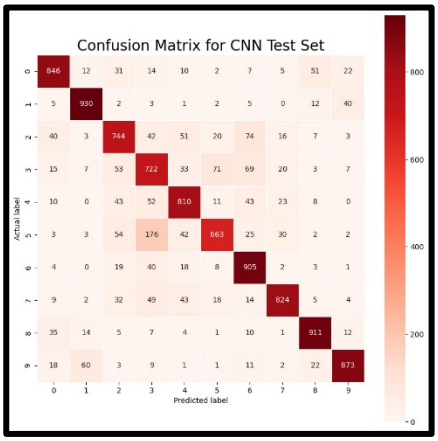






Confusion Matrix Insights (CNN)

The confusion matrix for the CNN model reflected high accuracy rates, with fewer misclassifications compared to SIFT and HOG. Notably, the model excelled in differentiating between visually similar classes, benefiting from the depth and complexity of learned features. Misclassifications that occurred were mostly between categories with close resemblance in color and shape, such as birds and airplanes or trucks and automobiles, which can be attributed to shared attributes in the limited resolution of CIFAR-10 images.



Comparative Analysis:

The comparative analysis of SIFT, HOG, and CNN methods provides valuable insights into the effectiveness of traditional and modern approaches in the field of object recognition. Here, we evaluate these methods based on their performance metrics from the CIFAR-10 dataset and discuss the strengths and weaknesses each method exhibits.

Performance Comparison

Method	Accuracy	Precision	Recall	F1-Score
SIFT	30%	30%	30%	30%
HOG	62%	62%	62%	62%
CNN	82%	83%	82%	82%

From the table, it's evident that the CNN model outperforms the traditional methods in all key metrics—accuracy, precision, recall, and F1-score. The CNN's ability to learn hierarchical features automatically from the data makes it exceptionally well-suited for complex image recognition tasks, as opposed to SIFT and HOG, which rely on predefined features extraction techniques.

Strengths and Weaknesses

SIFT:
Strengths: Good at detecting and describing local features that are invariant to image scale and rotation. Effective in scenarios where object scale and rotation vary.
Weaknesses: Poor performance in recognizing objects in cluttered backgrounds or when objects have varying illumination. Struggles with recognizing objects that do not have well-defined edges and corners.

HOG:
Strengths: Better at capturing edge or gradient structures that are characteristic of objects, leading to improved performance over SIFT in many scenarios.
Weaknesses: Still limited by the need for clear textural or edge information; performance drops when such features are not prominent or are masked by noise.

CNN:
Strengths: Superior in learning complex patterns directly from the data, leading to high accuracy. It is capable of handling variations in images with its deep learning architecture, making it robust against a variety of image distortions.
Weaknesses: Requires significant computational resources and data to train effectively. The model's performance is heavily dependent on the quality and quantity of the training data.

Comparative Insights
Robustness: CNNs demonstrate robustness against a broader range of object variations (such as pose, lighting, and background clutter) compared to SIFT and HOG.
Resource Efficiency: While CNNs require considerable computational resources for training, traditional methods like SIFT and HOG are computationally less demanding but offer reduced performance.
Applicability: Each method has its niche; SIFT and HOG may still be preferred in constrained environments where computational resources are limited, or real-time processing is crucial. In contrast, CNNs are suitable for applications where high accuracy is paramount, and computational resources are available.

Discussion:

The application of SIFT, HOG, and CNN techniques to object recognition in the CIFAR-10 dataset provides a comprehensive overview of how traditional and deep learning approaches can be leveraged in robotics and computer vision.

Integration into Robotics
Traditional Methods (SIFT and HOG): While they may not achieve the high accuracy rates of deep learning models, traditional techniques like SIFT and HOG are valuable in robotic applications where real-time processing is critical and computational resources are limited. For instance, in industrial robotics or simple autonomous drones, these methods can efficiently handle basic object detection tasks, aiding in navigation and manipulation tasks without the overhead of deep learning models.

Deep Learning (CNN): CNNs, with their robust performance, are ideal for advanced robotics applications such as autonomous vehicles, high-precision manufacturing, and service robots interacting in dynamic environments. The ability of CNNs to interpret complex scenes and learn from vast amounts of data allows robots to adapt to new tasks with minimal human intervention, paving the way for more autonomous and intelligent systems.

State of the art in Computer Vision for Robotics:
Deep learning has revolutionized computer vision in robotics, enabling robots to perceive and interact with their environment more effectively. Convolutional Neural Networks (CNNs) have been widely adopted for visual perception tasks such as object detection, segmentation, and pose estimation. Lenz et al. employed CNNs for robotic grasping, demonstrating their effectiveness in detecting graspable objects and predicting grasp configurations.

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants, have been utilized for sequential data processing and control tasks. Finn and Levine proposed an LSTM-based approach for learning robotic manipulation skills from demonstrations, enabling robots to imitate complex tasks.

Reinforcement Learning (RL) algorithms, such as Deep Q-Networks (DQN), have enabled robots to learn control policies through trial-and-error interactions with their environment. Levine et al. employed deep RL for robotic manipulation tasks, demonstrating its effectiveness in learning complex skills.

Autoencoders and Generative Adversarial Networks (GANs) have also been applied in robotics. Gupta et al. employed autoencoders for learning compact representations of robotic grasping configurations, while Bousmalis et al. proposed a GAN-based approach for sim-to-real transfer, enabling robots to leverage simulated data for real-world tasks.

Despite these advancements, challenges remain, including data scarcity, real-time performance, safety and robustness, and interpretability and explainability of deep models. Future research directions include developing more data-efficient learning algorithms, deploying deep models on resource-constrained robotic platforms, improving safety and robustness through techniques like sim-to-real transfer and adversarial training, and enhancing interpretability through explainable AI methods.

Conclusion:

This report embarked on a comprehensive exploration of traditional and deep learning methods for object recognition, applying SIFT, HOG, and CNN techniques to the CIFAR-10 dataset. Through systematic analysis and evaluation, we have gained valuable insights into the strengths and limitations of each method within the context of robotics and computer vision.

The comparative analysis revealed that while traditional methods like SIFT and HOG continue to provide value in specific scenarios requiring low computational overhead and real-time processing, deep learning techniques, exemplified by CNNs, significantly outperform in terms of accuracy and adaptability. CNNs demonstrate remarkable capabilities in complex image recognition tasks, making them highly suited for advanced robotics applications where high precision and adaptability are paramount.

Moreover, the discussion highlighted the critical role of these computer vision techniques in the ongoing advancement of robotics. The integration of these methods into robotic systems holds the potential to enhance autonomy and efficiency, paving the way for more intelligent and adaptable robotic solutions. However, the deployment of these technologies also brings challenges, particularly in terms of computational demands and the need for robust, interpretable models that ensure safety and reliability.

Looking forward, the field of robotics will benefit from the development of hybrid approaches that combine the efficiency of traditional methods with the power of deep learning. Further research into reducing the resource requirements of CNNs and enhancing the interpretability of deep learning models will also be crucial in making these advanced technologies more accessible and trustworthy.

References:

1. Lenz, I., Lee, H., & Saxena, A. (2015). Deep learning for detecting robotic grasps. The International Journal of Robotics Research, 34(4-5), 705-724.
2. Finn, C., & Levine, S. (2017). Deep visual foresight for planning robot motion. In 2017 IEEE International Conference on Robotics and Automation (ICRA) (pp. 2786-2793). IEEE.
3. Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 17(1), 1334-1373.
4. Gupta, A., Satkin, S., Efros, A. A., & Hebert, M. (2011). Data-driven grasping with partial sensor data. In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (pp. 1624-1630). IEEE.
5. Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., ... & Vanhoucke, V. (2018). Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4243-4250). IEEE.