Comparative Analysis of Classical, Geometric, and Modern Approaches for Image Classification and Detection on MNIST

Sushovit Nanda (21283)
Department of Electrical Engineering and Computer Science
Indian Institute of Science Education and Research, Bhopal
Email: sushovit21@iiserb.ac.in

***Abstract*—This project presents a comparative analysis of classical handcrafted, geometric, and modern deep learning-based approaches for image classification and detection using the MNIST dataset. The study evaluates traditional feature extraction techniques (SIFT with Bag-of-Words, HOG, and LBP) combined with SVM and KNN classifiers, geometric algorithms including the Hough Transform and RANSAC, and pretrained deep models such as ResNet-18 and OpenAI's CLIP. A unified experimental pipeline with structured augmentations across training and test sets ensures fair evaluation of robustness, accuracy, and computational efficiency. The findings highlight trade-offs among interpretability, robustness, and semantic generalization across these paradigms.**

***Index Terms*—MNIST, SIFT, HOG, LBP, Bag-of-Words, RANSAC, Hough Transform, CLIP, ResNet-18, image classification**

I. Introduction and Motivation

Image classification has transitioned from handcrafted feature descriptors and geometric shape models to powerful deep and multimodal architectures. While deep learning models such as ResNet-18 and Contrastive Language-Image Pretraining (CLIP) achieve exceptional performance, they often require large-scale data and compute resources. Conversely, handcrafted and geometric approaches retain advantages in interpretability, efficiency, and stability under distortion. This project provides a unified framework comparing these paradigms-classical, geometric, and modern pretrained methods-using the MNIST dataset as a controlled dataset. The objective is to understand how each approach differs in robustness, generalization, and computational trade-offs under controlled augmentations involving rotation, scaling, occlusion, and noise.

II. Methodology

The experimental workflow integrates dataset augmentation, feature extraction, and evaluation under identical test conditions. The MNIST dataset was normalized and resized to 28×28 grayscale format. Four augmentations were applied-rotation ($15°$), scaling ($0.8\times$), occlusion (25% region masking), and Gaussian noise addition. Three training configurations were prepared: (1) *original* (no augmentation), (2) *mixed-augmented* (equal distribution of all four augmentations applied randomly across the dataset), and (3) *combined-augmented* (all augmentations applied sequentially to every sample). Six test sets were designed: *original*, *rotation*, *scaling*, *noise*, *occlusion*, and *mixed* (containing all augmentations). This enabled a 3×6 study of training–testing combinations, yielding 18 experiments per method. For CLIP, all six test sets were analyzed using four prompt template sets to explore linguistic sensitivity in zero-shot prediction.

A. Classical Feature-Based Methods

Three handcrafted pipelines were implemented for the classical feature-based study. The SIFT-based method employed a Bag-of-Words (BoW) representation, where local descriptors were extracted and clustered using MiniBatchKMeans to form a visual vocabulary. Each image was represented by a histogram of visual word occurrences, later classified using SVM and KNN. HOG and LBP were used as direct feature extractors-HOG captured gradient orientation histograms encoding edge and contour structure, while LBP described local binary intensity variations. The extracted features were flattened and standardized before being fed to the classifiers. All models were evaluated across training–testing combinations, allowing comparative study of handcrafted descriptor stability under rotation, occlusion, and noise. These classical models are interpretable, computationally efficient, and rely on explicit feature engineering rather than data-driven learning.

B. Geometric Detection Methods

The geometric framework, developed explored low-level structure detection through the Hough Transform and RANSAC algorithms. The Hough Transform mapped parametric shapes such as lines and circles to capture digit strokes, loops, and geometric alignments. RANSAC estimated geometric models in the presence of noise by iteratively selecting subsets of edge points, rejecting outliers, and refitting consistent models. These approaches are non-learning-based and provide insight into digit morphology, offering geometric robustness even when texture or gradient information is corrupted. Performance evaluation focused on their ability to preserve structural integrity across augmented images, providing a complementary robustness analysis to classification metrics.

C. Modern Deep and Zero-Shot Approaches

Modern learning-based evaluation used pretrained networks to study semantic generalization. ResNet-18 model, initialized

with ImageNet weights *"IMAGENET1K_V1"*, was fine-tuned on MNIST to serve as a supervised deep baseline. It utilized convolutional blocks with residual connections to learn hierarchical features from pixel-level inputs. For zero-shot evaluation, the CLIP model *"openai/clip-vit-large-patch14"* was employed, which aligns visual and textual embeddings in a shared semantic space. CLIP was evaluated using four prompt template sets: *basic* ("a photo of the number {}.", "the digit {}."), *descriptive* ("a handwritten digit {}.", "a photo of a handwritten number {}.", etc.), *contextual* ("a black and white image of number {}.", "a grayscale photo of the number {}."), and *minimal* ("{}."). Each prompt group was tested across all six augmented test sets to analyze zero-shot robustness and prompt dependency.

III. Results and Discussion

The 18 training–testing evaluations revealed distinct performance differences across paradigms, assessed through F1 scores. Among handcrafted methods, SIFT with Bag-of-Words achieved the strongest performance, with SVM-RBF yielding an average F1 of 0.69 (max: 0.81, min: 0.51). Other classifiers showed comparable results: Logistic Regression (0.68), MLP (0.67), and LightGBM (0.62). The gradient-based SIFT descriptors provided robustness to scale changes and moderate rotations.

LBP features performed poorly, with Random Forest achieving the highest average F1 of only 0.15 (max: 0.36), followed by XGBoost (0.15) and LightGBM (0.14). The texture-based encoding lacked the structural information essential for digit recognition, degrading rapidly under distortion and rotation.

HOG with SVM achieved superior F1 scores of 0.96–0.97, demonstrating robustness through gradient-based edge orientation encoding. Geometric algorithms (RANSAC, Hough Transform) showed resilience under noise and occlusion but were unsuitable for classification.

Deep pretrained models outperformed all handcrafted approaches. ResNet-18 achieved an F1 of approximately 0.98 with strong generalization, while CLIP reached 0.93–0.95 in zero-shot settings. Descriptive and contextual prompts yielded the highest CLIP performance. The performance hierarchy (ResNet-18 $>$ CLIP $>$ HOG $>$ SIFT BoW $>>$ LBP) demonstrates that gradient-based representations are essential for digit recognition, with pretrained models providing superior discrimination and robustness across augmentations.

IV. Conclusion

This work establishes a unified and reproducible framework for comparing handcrafted, geometric, and deep pretrained paradigms for image classification on MNIST. Through standardized augmentation and consistent F1 score evaluation across the 18 experiments, it demonstrates that pretrained models such as ResNet-18 (F1: 0.98) and CLIP (F1: 0.93–0.95) achieve superior generalization and semantic reasoning. Among handcrafted methods, gradient-based approaches prove essential: HOG attains F1 scores of 0.96–0.97, while SIFT BoW achieves 0.69, both significantly outperforming texture-based LBP (F1: 0.06–0.15). The performance hierarchy reveals that gradient information is critical for digit recognition, whereas texture-based descriptors fail to capture discriminative structural patterns. Classical methods retain computational efficiency and interpretability, while deep models provide robustness under diverse perturbations. This study highlights the complementary strengths of these paradigms and advocates the integration of geometric priors and gradient-based classical descriptors with pretrained embeddings to develop hybrid, explainable, and resilient vision systems for resource-constrained and real-world scenarios.