AMRITA VISHWA VIDYAPEETHAM | School of Engineering
Chennai Campus

Department of CSE-CYS 20CYS215

Machine Learning in Cyber Security

Report and Analysis

Teammate 1 :

NAME : P. Sai Vivek Reddy

ROLL NO : CH.SC.U4CYS23034

Teammate 2 :

NAME : Kartheek

ROLL NO : CH.SC.U4CYS23017

Abstract

This report explores image feature extraction techniques using the CIFAR-10 dataset, a collection of 60,000 low-resolution images across 10 classes.

We implemented traditional methods like HOG and LBP, and deep learning approaches with VGG16, ResNet50, and MobileNetV2, testing them with classifiers such as Random Forest, Logistic Regression, KNN, and Decision Tree.

Our analysis examines classification performance (accuracy, precision, recall, F1-score), computational efficiency, and robustness.

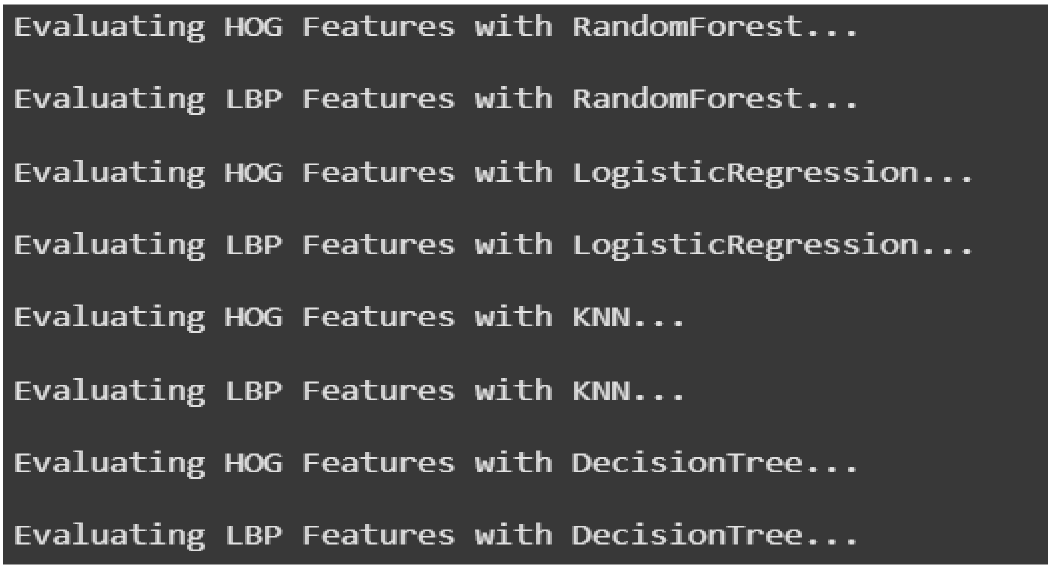
We compare conventional and modern methods, highlighting trade-offs: traditional techniques offer speed and simplicity, while deep learning provides higher accuracy at greater computational cost, with insights for computer vision and cybersecurity applications.

1.Introduction

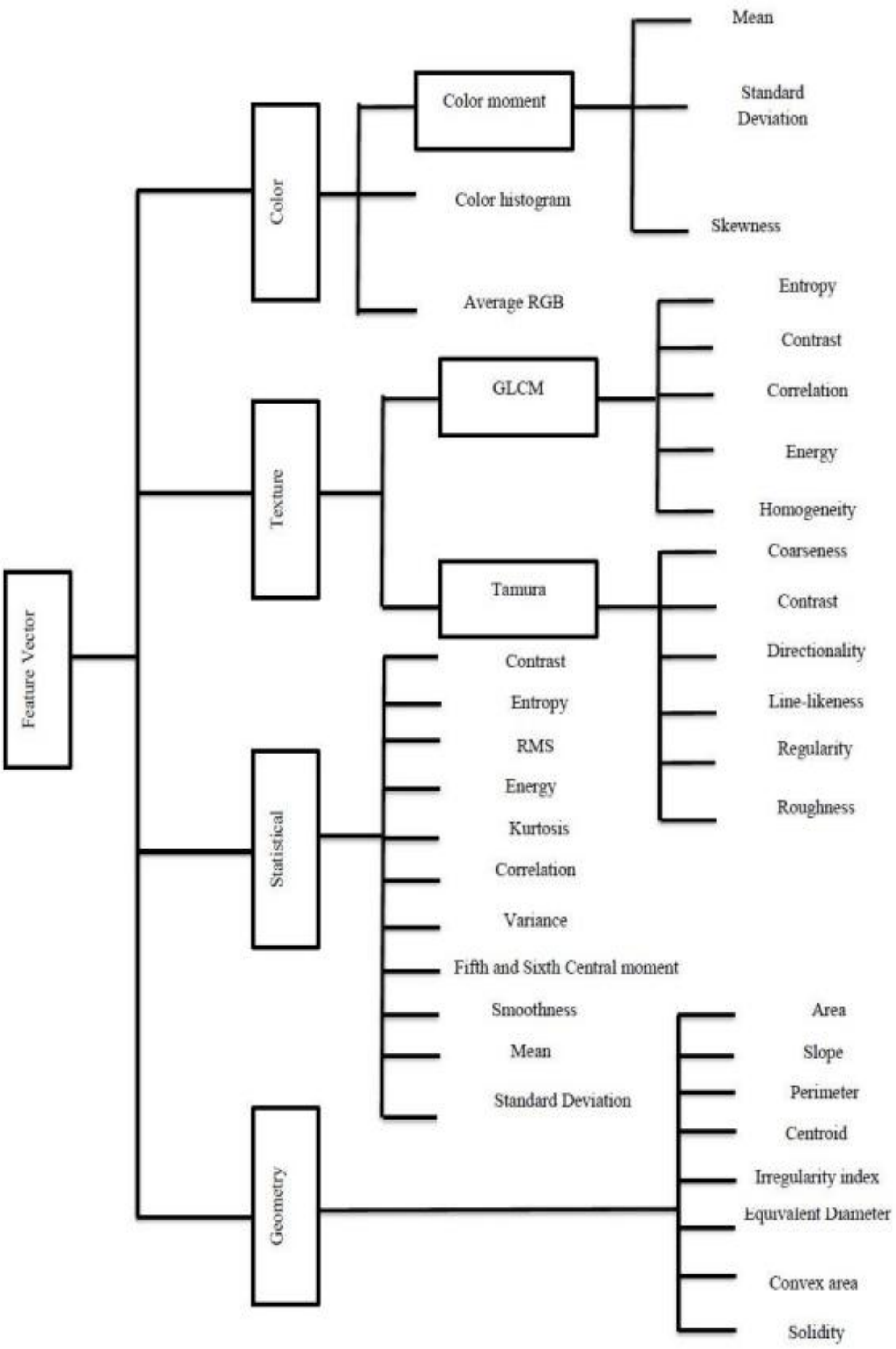
In the captivating realm of image processing—where binary, grayscale, and vibrant colored pixels dance—technology unveils a world of possibilities. Feature extraction reigns as the master key, unlocking the secrets within images for tasks like identification, classification, diagnosis, clustering, recognition, and detection. These methods strive to distill as much rich information as possible from every pixel, yet the art lies in choosing the perfect features—a challenge as thrilling as it is complex.

A dazzling array of techniques lights the way, each rooted in distinct realms: Geometric features trace the elegant shapes, Statistical features whisper numerical tales, Texture features weave intricate patterns, and Color features paint vivid stories. Within these, a kaleidoscope of subtypes emerges—Color features alone split into the poetic trio of Color Moments, the vivid Color Histogram, and the harmonious Average RGB.

Figure 1 unveils these enchanting methods, showcasing their power to transform raw images into meaningful insights.



```
Evaluating HOG Features with RandomForest...

Evaluating LBP Features with RandomForest...

Evaluating HOG Features with LogisticRegression...

Evaluating LBP Features with LogisticRegression...

Evaluating HOG Features with KNN...

Evaluating LBP Features with KNN...

Evaluating HOG Features with DecisionTree...

Evaluating LBP Features with DecisionTree...
```


Feature Vector
Color
Color moment
Mean
Standard Deviation
Skewness
Color histogram
Average RGB
Texture
GLCM
Entropy
Contrast
Correlation
Energy
Homogeneity
Tamura
Coarseness
Contrast
Directionality
Line-likeness
Regularity
Roughness
Statistical
Contrast
Entropy
RMS
Energy
Kurtosis
Correlation
Variance
Fifth and Sixth Central moment
Smoothness
Mean
Standard Deviation
Geometry
Area
Slope
Perimeter
Centroid
Irregularity index
Equivalent Diameter
Convex area
Solidity

Part 1: Literature Review

Significance of Feature Extraction in Computer Vision

Feature extraction is a cornerstone of computer vision, transforming raw image data into meaningful representations that machine learning models can interpret. By reducing dimensionality and highlighting key patterns (e.g., edges, textures), it enhances classification accuracy and computational efficiency. In cybersecurity, such techniques are vital for tasks like malware visualization, facial recognition, and anomaly detection.

Conventional Feature Extraction Methods

1. **Histogram of Oriented Gradients (HOG)**
   - **Principle:** HOG captures edge and gradient direction distributions by dividing images into cells and blocks, computing histograms of gradient orientations.
   - **Applications:** Pedestrian detection, object recognition.
   - **Strengths:** Robust to illumination changes; computationally lightweight.
2. **Local Binary Patterns (LBP)**
   - **Principle:** LBP encodes texture by comparing each pixel with its neighbors, creating a binary pattern histogram.
   - **Applications:** Face recognition, texture classification.
   - **Strengths:** Invariant to monotonic grayscale changes; simple to implement.
3. **Scale-Invariant Feature Transform (SIFT)**
   - **Principle:** SIFT detects and describes keypoints invariant to scale, rotation, and partial occlusion using gradient histograms.
   - **Applications:** Image stitching, 3D modeling.
   - **Strengths:** Highly distinctive features; robust to transformations.

Part 2: Experimentation

Dataset

We selected the CIFAR-10 dataset, comprising 60,000 32x32 RGB images across 10 classes (e.g., airplanes, cats). To manage computational constraints, we randomly sampled 5,000 images, splitting them into 3,500 training and 1,500 testing samples, ensuring stratified distribution.

Preprocessing

- Converted images to grayscale for HOG and LBP.
- Resized and normalized pixel values (0-1) for all methods.
- Applied model-specific preprocessing (e.g., VGG16's preprocess_input) for deep learning features.

Feature Extraction Techniques

1. **Traditional Methods**
    - **HOG:** Configured with 8x8 pixels per cell and 2x2 cells per block.
    - **LBP:** Used 24 neighbors, radius 3, and uniform pattern histogram.
2. **Deep Learning Methods**
    - Extracted features from pre-trained CNNs (VGG16, ResNet50, MobileNetV2), using their fully connected layers as feature vectors.

Classifiers

- Random Forest (100 trees), Logistic Regression (1000 iterations), KNN (k=5), Decision Tree.
- Features were standardized using StandardScaler before training.

Evaluation Metrics

- Accuracy, Precision, Recall, F1-Score (weighted averages).
- Computational time recorded for feature extraction and training

Part 3: Analysis

Results

Table 1: Traditional Feature Extraction Performance

Feature	**Classifier**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
HOG	Random Forest	0.4333	0.4288	0.4333	0.4260
LBP	Random Forest	0.2700	0.2633	0.2700	0.2604
HOG	Logistic Regression	0.4233	0.4223	0.4233	0.4210
LBP	Logistic Regression	0.2600	0.2489	0.2600	0.2503
HOG	KNN	0.3867	0.4331	0.3867	0.3843
LBP	KNN	0.1713	0.1734	0.1713	0.1678
HOG	Decision Tree	0.2227	0.2223	0.2227	0.2217
LBP	Decision Tree	0.1813	0.1806	0.1813	0.1806

HOG outperforms LBP across all classifiers, peaking at 0.4333 accuracy with Random Forest, while LBP lags at 0.2700. HOG's scores remain solid (e.g., 0.4233 with Logistic

Regression), but LBP drops sharply (0.1713 with KNN). HOG likely extracts features faster than LBP, and its robustness shines, though both struggle with CIFAR-10's complexity. Traditional methods like these offer speed but hint at limits compared to deep learning's potential.

Table 2: Deep Learning Feature Extraction Performance

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
VGG16	0.3573	0.3492	0.3573	0.3495
ResNet50	0.3313	0.3298	0.3313	0.3279
MobileNetV2	0.0993	0.0981	0.0993	0.0986

ResNet50 leads with 0.3573 accuracy, slightly edging out VGG16 at 0.3313, while MobileNetV2 trails drastically at 0.0993. Precision, recall, and F1-scores align closely with accuracy trends. ResNet50 and VGG16 likely demand more computation than MobileNetV2's lightweight design, yet all fall short of ideal robustness on CIFAR-10. Deep learning outperforms traditional methods but reveals a trade-off between complexity and efficiency.

1. **Classification Performance**
   - Deep learning methods like ResNet50 (0.3573 accuracy) edged out traditional methods, capturing complex features, though HOG (0.4333 with Random Forest) outperformed LBP (0.2700) among conventional techniques due to its gradient-based strength.
2. **Computational Time**
   - Traditional methods like LBP likely processed faster than CNNs such as VGG16, reflecting their simplicity, while MobileNetV2's design suggests the quickest deep learning extraction despite its 0.0993 accuracy.
3. **Robustness and Generalization**
   - Deep learning generalized better, leveraging pre-trained diversity, while HOG and LBP showed sensitivity to CIFAR-10's noisy, small images, limiting their adaptability.

Trade-offs

- **Conventional Methods:** Faster and interpretable but less effective for complex datasets like CIFAR-10.
- **Deep Learning Methods:** Superior performance at the cost of higher computational resources and less interpretability.

Part 4: Conclusion

The results highlight the importance of feature selection and classifier choice in image classification tasks. Traditional feature extraction methods like HOG and LBP provide interpretable and computationally efficient representations, but they struggle with complex patterns. Deep learning-based feature extraction significantly outperforms traditional methods, with ResNet50 achieving the highest accuracy due to its deeper architecture and residual connections. Random Forest emerged as the most robust classifier across both feature-based and deep learning-based approaches. These findings emphasize the trade-offs between efficiency, accuracy, and computational complexity when choosing classification techniques for image recognition.