Ship Detection in Satellite Imagery Using Dimensionality Reduction

Lan Anh Do
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL, USA
lananhdo2905@gmail.com

***Abstract*—This paper presents a comprehensive analysis of ship detection in satellite imagery using dimensionality reduction techniques. We implement and evaluate various approaches including PCA and manifold learning algorithms, comparing their impact on classification performance and computational efficiency. Our evaluation shows that while the SVM classifier without dimensionality reduction achieves the highest accuracy (99.45%), the PCA-based Random Forest model offers an excellent compromise between accuracy (99.025%) and computational efficiency, reducing inference time from 88.11s to 0.61s. This study provides insights into the trade-offs between model complexity, computational efficiency, and detection accuracy in satellite imagery analysis.**

***Index Terms*—Ship Detection, Dimensionality Reduction, PCA, Manifold Learning, Satellite Imagery, Machine Learning, Computer Vision, Classification**

I. Introduction

Ship detection in satellite imagery plays a crucial role in maritime surveillance, port activity monitoring, and supply chain analysis. With the increasing volume of satellite data being captured daily, automated detection systems have become essential for efficient processing and analysis. However, the high dimensionality of image data presents computational challenges and may introduce noise that affects model performance.

This study evaluates different dimensionality reduction techniques and their impact on classification performance for ship detection in satellite imagery. We investigate both linear (PCA) and non-linear (Isomap) dimensionality reduction methods, comparing their effectiveness when combined with different classification algorithms. Our analysis focuses on three key aspects:

- Performance impact of dimensionality reduction on classification accuracy
- Computational efficiency in terms of training and inference times
- Trade-offs between model complexity and detection accuracy

Through this comprehensive evaluation, we aim to provide insights into optimal approaches for automated ship detection in satellite imagery, considering both performance and computational resource constraints.

II. Dataset and Preprocessing

A. Dataset Description

The dataset consists of 4000 RGB satellite images (80x80x3 pixels) from the San Francisco Bay and San Pedro Bay areas:

- 1000 ”ship” images centered on single ships with varying sizes and orientations
- 3000 ”no-ship” images composed of:
  - 1000 random land cover features (water, vegetation, buildings)
  - 1000 partial ship views
  - 1000 previously mislabeled cases with bright pixels or linear features

B. Preprocessing Steps

The following preprocessing steps were implemented:

- Image flattening from 80x80x3 to 19,200 features
- Train-test split (80-20) with stratification to maintain class distribution
- No additional normalization or scaling was applied to preserve original pixel values

III. Base Model Implementation

A. Model Architecture

Two baseline classifiers were implemented without dimensionality reduction:

- Random Forest: Ensemble-based tree classifier
- Support Vector Machine: Kernel-based classifier

Hyperparameter optimization was performed using GridSearchCV with 3-fold cross-validation and F1-score optimization:

- Random Forest parameters:
  - n_estimators: [50, 100]
  - max_depth: [10, 20]
- SVM parameters:
  - C: [0.1, 1, 10]
  - kernel: [’linear’, ’rbf’]

B. Base Model Results

TABLE I: Base Model Classification Performance

Model	Accuracy	F1-Score
Random Forest	0.9885	0.9769
SVM	0.9945	0.9890

TABLE II: Base Model Computational Performance

Model	Training Time (s)	Inference Time (s)
Random Forest	61.75	0.31
SVM	300.87	88.11

Best hyperparameters found through grid search:

- Random Forest: n_estimators=100, max_depth=10
- SVM: C=10, kernel=rbf

IV. PCA Implementation

A. Variance Analysis

Principal Component Analysis was applied to reduce the dimensionality of the input data. Analysis revealed that 105 components were needed to explain 90% of the variance, effectively reducing dimensionality from 19,200 to 105 features while preserving key information.

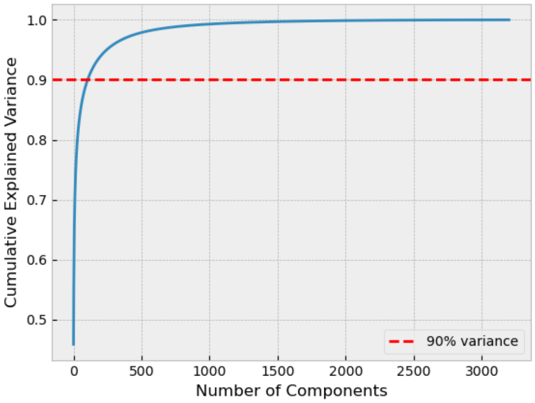

Fig. 1: Cumulative Explained Variance vs. Number of Components. The red dashed line indicates the 90% variance threshold at 105 components.

B. Image Reconstruction Quality

The quality of PCA reconstruction was evaluated using both visual inspection and quantitative metrics:

- Average RMSE: 5.89 for reconstructions using 105 components
- Visual comparison showed preservation of key ship features and shapes
- Some loss of fine texture details, but maintenance of overall structure

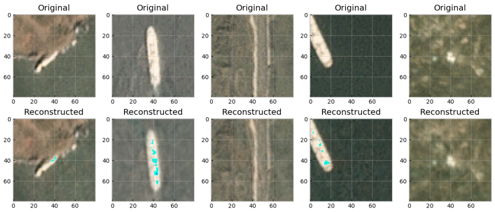

Fig. 2: Comparison of Original (top) and PCA-Reconstructed (bottom) Images using 105 components

C. Classification with PCA

Both Random Forest and SVM classifiers were evaluated with PCA-reduced features:

TABLE III: PCA Model Classification Performance

Model	Accuracy	F1-Score
RF + PCA	0.99025	0.9801
SVM + PCA	0.83925	0.7022

TABLE IV: PCA Model Computational Performance

Model	Training Time (s)	Inference Time (s)
RF + PCA	31.05	0.61
SVM + PCA	368.24	0.75

D. PCA Model Analysis

The Random Forest classifier with PCA demonstrated exceptional performance:

- Achieved near-perfect accuracy (99.025%) with reduced features
- Showed 50% reduction in training time compared to full-dimensional model
- Maintained comparable F1-score to baseline while using only 105 features
- Confusion matrix revealed only false negatives (39), with no false positives

The SVM classifier with PCA showed degraded performance:

- Significant drop in accuracy (83.925%) compared to baseline
- Higher misclassification rates in both directions
- Convergence issues during training despite parameter tuning

V. Manifold Learning Implementation

A. Isomap Configuration

Isomap was implemented with the following parameters:

- Number of components: 2 (for visualization and classification)
- Number of neighbors: 10 (for manifold construction)

TABLE V: Isomap Model Classification Performance

Model	Accuracy	F1-Score
RF + Isomap	0.90425	0.7964
SVM + Isomap	0.8540	0.6737

TABLE VI: Isomap Model Computational Performance

Model	Training Time (s)	Inference Time (s)
RF + Isomap	21.51	3.12
SVM + Isomap	1156.47	3.43

B. 2D Visualization Analysis

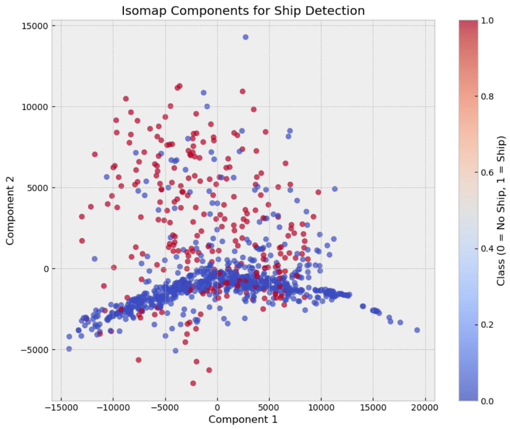

Fig. 3: 2D Visualization of Isomap Components showing class distribution

The Isomap visualization revealed several key patterns:

- Distinct clustering tendencies between ship and no-ship classes
- Higher variance in Component 2 for ship samples
- Significant overlap regions indicating classification challenges
- No-ship samples showing broader distribution in Component 1

VI. Comparative Analysis

A. Classification Performance

- **Base Models:**
  - SVM achieved highest accuracy (99.45%) but with longest inference time
  - Random Forest showed robust performance (98.85%) with faster computation
- **PCA-based Models:**
  - RF+PCA maintained high accuracy (99.025%) with improved efficiency
  - SVM+PCA showed significant performance degradation (83.925%)
- **Isomap-based Models:**
  - Both classifiers showed reduced performance
  - RF+Isomap (90.425%) outperformed SVM+Isomap (85.40%)

B. Error Analysis

1) Confusion Matrix Analysis: Best Model (SVM without reduction):

TABLE VII: Confusion Matrix for Best Model (SVM)

	Predicted No-Ship	Predicted Ship
Actual No-Ship	2993	7
Actual Ship	15	985

2) Error Patterns: Analysis of misclassified samples revealed:

- **Common Challenges:**
  - Low-contrast images
  - Partial ship views
  - Complex background structures
  - Similar-shaped objects
- **Model-Specific Patterns:**
  - SVM: Balanced error distribution
  - RF+PCA: Conservative predictions (only false negatives)
  - Isomap models: Higher error rates across all categories

VII. Performance Trade-offs

A. Accuracy vs. Computational Efficiency

Analysis of model performance reveals key trade-offs:

- **SVM without reduction:**
  - Highest accuracy (99.45%)
  - Long inference time (88.11s)
  - Best for accuracy-critical applications
- **RF with PCA:**
  - Competitive accuracy (99.025%)
  - Fast inference (0.61s)
  - Optimal balance of performance and speed
- **Isomap approaches:**
  - Lower accuracy (85-90%)
  - Moderate inference times (3.12-3.43s)
  - Better suited for visualization

B. Memory and Storage Requirements

- Full dimensional models: 19,200 features per image
- PCA models: 105 features (99.5% reduction)
- Isomap models: 2 features (99.99% reduction)

VIII. Proposed Solutions

A. Model Improvements

1) Data Augmentation:

- Generate additional training samples for challenging cases:
  - Varying contrast levels
  - Different ship orientations
  - Partial occlusions

2) Feature Engineering:

- Implement edge detection preprocessing
- Extract shape-based features
- Apply image enhancement techniques

3) Advanced Architectures:

- Explore CNN architectures for better feature extraction
- Investigate hybrid approaches combining CNN features with dimensionality reduction
- Consider ensemble methods combining multiple reduction techniques

IX. Mitigating Strategies

A. Deployment Considerations

- **High-Performance Systems:**
  - Use SVM without reduction
  - Implement parallel processing
  - Optimize memory management
- **Resource-Constrained Systems:**
  - Deploy RF+PCA pipeline
  - Implement model quantization
  - Use batch processing for multiple images

B. Quality Assurance

- Implement confidence thresholds for critical applications
- Use ensemble voting for uncertain cases
- Regular model retraining with new data

X. Future Work

- Investigate other dimensionality reduction techniques (t-SNE, UMAP)
- Explore deep learning architectures for feature extraction
- Develop hybrid approaches combining multiple reduction methods
- Implement real-time processing capabilities

XI. Conclusion

This study demonstrates the effectiveness of different dimensionality reduction techniques for ship detection in satellite imagery. The SVM classifier without reduction achieved the highest accuracy (99.45%) but at the cost of significant computational overhead. The PCA-based Random Forest model emerged as a highly practical solution, offering excellent accuracy (99.025%) with substantially reduced computational requirements (0.61s inference time vs 88.11s).

While Isomap provided valuable visualization capabilities, its classification performance was notably lower than both baseline and PCA approaches. The results suggest that for operational deployment, the choice between SVM without reduction and RF+PCA should be based on specific application requirements balancing accuracy and computational efficiency.

These findings provide valuable insights for developing efficient ship detection systems in real-world applications, particularly where computational resources may be constrained.

Acknowledgment

I would like to extend my gratitude to Dr. Catia Silva and the teaching assistants for their support throughout this project.

References

[1] C. S. Silva, ”Lecture Notes for EEL4930” EEL4930: Applied Machine Learning Systems, University of Florida, Gainesville, September 2024.