Project Report

Linear Algebra-II

Submitted By:

Khushal Khan – 2021261

Abdullah Khan- 2021035

Behram Khan-2021127

Noshair Imtiaz-2021429

How to Handle Large Datasets? PCA Comes to Rescue

Khushal Khan 2021261 Behram Khan 2021127

Abdullah Khan 2021035 Noshair Imtiaz 2021429

Computer Engineering
Ghulam Ishaq Khan Institute of Engineering Sciences and Technology
Topi, Swabi, Pakistan

***Abstract*—A multispectral image captures information across various spectral bands, providing valuable insights beyond the visible spectrum. This paper explores the application of Principal Component Analysis (PCA) to reduce the dimensionality of multi-spectral images and includes a comprehensive error analysis of the transformed images. The experiment reveals the effectivenessof PCA in handling large datasets and emphasizes its significancein remote sensing applications.**

***Index Terms*—Multispectral Imaging, Principal Component Analysis, Dimensionality Reduction, Error Analysis, Remote Sensing.**

I. Introduction

Multispectral imaging goes beyond the limitations of conventional images by employing more than three spectral filters to capture information. This technology enables the simul- timeous analysis of data across multiple bands, allowing for the extraction of valuable insights that remain hidden in standard images. Multispectral imaging finds diverse applications globally, including areas such as land mine detection, weather forecasting, space-based imaging, ballistic missile detection, document, and artwork analysis, as well as military target tracking.

In this project, we delve into the realm of multispectral imaging using datasets obtained from the Landsat program. Specifically, we have acquired two datasets one containing the multispectral image and the other consisting of images from individual bands of the Dera Ghazi Khan region. These datasets, obtainedfrom https://earthexplorer.usgs.gov, serve as the foundation forour exploration.

Principal Component Analysis (PCA) stands as a key statistical technique employed in this study to address the challenges posed by high-dimensional datasets. PCA facilitates dimensionality reduction by linearly transforming the data into a new coordinate system. This new representation captures (most of) the variation in the data using fewer dimensions than the original dataset. PCA has proven applications in various fields, including population genetics, microbiome studies, and atmospheric science.

Group No: 21

II. Methodology

A. Basic Operations on Multispectral Images

The initial phase involves a series of operations to preprocess and prepare the multispectral dataset for PCA.

1) Visualization of Bands: In our dataset, we worked with multiple raster graphic images representing different bands. Utilizing the Geospatial Data Abstraction Library (GDAL) in Python, we read the raster images and converted them into arrays. This conversion enabled us to programmatically access and manipulate pixel values within the images. Following the data transformation, the Matplotlib Python library played a crucial role in visualizing the information.

2) Cropping the Size: The Python Imaging Library was employed for a four-dimensional cropping strategy, considering left, top, right, and bottom dimensions. This cropping process refined the region of interest within the multispectral images.

3) Concatenation of Bands: During this phase, we scaled the images representing different bands of our multispectral dataset. The NumPy library facilitated the concatenation of these scaled bands into a unified multispectral image. The concatenated image was then visualized using Matplotlib.

TABLE I
Operations on Multispectral Images

Operation	Tool/Library
Visualization of Bands	Matplotlib, GDAL
Cropping the Size	Python Imaging Library
Concatenation of Bands	NumPy

B. Applying PCA to Multi-Dimensional Image

The core of the project involves applying PCA to the preprocessed multispectral image. The following steps detail the methodology:

1) Importing Required Libraries: We utilized essential libraries, including Matplotlib for visualization, scikit-learn for PCA implementation, and NumPy for array operations.

2) Reading the Multidimensional Image: Using Matplotlib's imread function, we loaded the preprocessed multispectral image.

3) Preprocessing Step: The preprocessing involved converting the data type of the image array to uint8 for ease of handling, normalizing the array, and converting the image array to 2D to prepare it for PCA.

4) Creating PCA Function: A custom PCA function was developed using scikit-learn's PCA module. This involved finding the principal components, applying these components to the image, and assessing projections of principal components on the image.

TABLE II
PRINCIPAL COMPONENT ANALYSIS STEPS

VI.

Step	Action
Importing Required Libraries	Matplotlib, scikit-learn, NumPy
Reading the Multidimensional Image	Matplotlib
Preprocessing Step	Data type conversion, normalization, 2D conversion
Creating PCA Function	scikit-learn PCA module

C. Error Analysis on the Use of PCA

An in-depth error analysis was conducted to assess the quality of transformed images resulting from PCA. Different numbers of principal components (2-10) were considered, and Mean Squared Error (MSE) was employed to quantify the dissimilarity between the original and transformedimages.

TABLE III
ERROR ANALYSIS OF PCA

Number of Principal Components	Mean Squared Error (MSE)
2	9.36
3	9.31
4	9.29
6	9.29

III. RESULTS

The experiment yielded insightful results, demonstrating that an increase in the percentage of PCA components correlated with a reduction in Mean Squared Error (MSE).

IV. DISCUSSION

The successful application of PCA in reducing dimensionality opens avenues for enhanced multispectral image analysis. The choice of the optimal percentage of principal components is crucial and depends on the specific requirements of the remote sensing application. Further refinements in the PCA algorithm and exploration of advanced techniques could con- tribute to even more precise results.

V. TASK DISTRIBUTION

The distribution of tasks among team members is as follows:

- Khushal Khan: Visualization of Bands
- Behram Khan: Cropping the Size
- Abdullah Khan: Concatenation of Bands
- Noshair Imtiaz: PCA Implementation and Error Analysis

CONCLUSION

In conclusion, this project illustrates the significance of Principal Component Analysis in handling large multispectral datasets. The combination of advanced visualization, preprocessing, and PCA application provides a robust framework for effective dimensionality reduction. The error analysis under- scores the importance of selecting an appropriate percentage of principal components for optimal results in remote sensing applications.

REFERENCES

[1] Wikipedia," Earth observation satellite," 21 Jan 2017. [Online]. Available:https://en.wikipedia.org/wiki/Earth observation satellite