**Major Project Report**

**on**

# STATE LEVEL ETHNICITY IDENTIFICATION USING HANDWRITTEN TEXT ANALYSIS

Submitted by


| **Names of Students** | **Reg. Numbers** |
|---|---|
| BADIGINCHALA CHANDANA PRIYA | 20bcs026 |
| HARSHITA NG | 20bcs055 |
| LALAM DIVYA SRI | 20bcs076 |
| RAVULA VEEKSHITH REDDY | 20bcs111 |


Under the guidance of

**Dr. PAVAN KUMAR**

Head of The Department, Department of Computer Science and Engineering



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**INDIAN INSTITUTE OF INFORMATION TECHNOLOGY DHARWAD**


21/03/2024

# *Certificate*

This is to certify that the project, entitled **STATE LEVEL ETHNICITY IDENTIFICATION USING HANDWRITTEN TEXT ANALYSIS**, is a bonafide record of the Major Project coursework presented by the students whose names are given below during 2023-24 in partial fulfilment of the requirements of the degree of Bachelor of Technology in Computer Science and Engineering.

| Roll No | Names of Students |
|---|---|
| 20bcs026 | Badiginchala Chandana Priya |
| 20bcs055 | Harshita NG |
| 20bcs076 | Lalam Divya Sri |
| 20bcs021 | Ravula Veekshith Reddy |

Dr. Pavan Kumar

# Contents

# List of Figures

# 1 Introduction

We've devised a pioneering model aimed at tackling the challenges encountered in handwriting identification and determining ethnic origins, particularly in the context of global crime investigations and the intricate nature of investigative procedures. Our model isn't solely confined to police work; it has broad applicability across fields such as forensic analysis, bolstering national security, and addressing historical document inquiries.

Ethnicity plays a crucial role in handwriting analysis, as it's influenced by shared cultural backgrounds, languages spoken within communities, and ancestral heritage. Rather than relying on exhaustive datasets, our approach takes a nuanced stance, focusing on identifying ethnicity at the state level. State-level ethnicity recognition involves discerning ethnicities within specific states or regions of a country. It's recognized that ethnic identities can vary significantly within a country, with different states exhibiting distinct ethnic compositions and cultural norms.

By adopting this approach, we acknowledge that each geographical area possesses its own handwriting characteristics reflective of its cultural legacy. Consequently, through handwriting analysis, we can uncover valuable insights into regional nuances and cultural specifics. This, in turn, equips law enforcement agencies and forensic investigators with unprecedented proficiency levels to enhance their investigative capabilities.

## 2 Related Work

The problem of identifying ethnicity through handwriting analysis has received attention in various contexts, such as gender identification, age estimation, and personality traits identification. Previous research in the field of ethnicity identification using handwriting analysis has laid the foundation for the current study. Researchers proposed an automatic method for predicting age, gender, and nationality in offline handwriting. Their method extracted direction, curvature, tortuosity, chain code, and edge-based directional features for character components, aiming to capture unique differences in handwriting styles among different nationals. However, their methodology was limited in scope and reported poor results for nationality identification, indicating the need for more robust approaches. Other studies have explored handwriting analysis for various purposes, such as writer identification and signature verification.

These studies have aimed at the unique characteristics of handwriting that can be leveraged for identification purposes. For writer identification methods aim to differentiate between individuals based on unique properties in their writing, such as strokes, shapes, and styles of characters [3, 4, 5]. While effective for distinguishing between individuals, these methods may not provide and conclude to nationality identification, where each nationality can have a diverse range of writing styles. Existing methods for nationality identification have primarily focused on conventional approaches, such as using features extracted from Cloud of Line Distribution (COLD) and applying SVM classifiers [1]. By studying the shapes of character components in handwritten text lines, the method aims to capture the unique variations in handwriting styles among different nationals. This approach represents a significant advancement in the field of handwriting analysis for nationality identification. But, these proposed methods are

limited in their ability to handle complex nationality identification tasks involving multiple classes. Some proposed Edge-Attention based U-Net (EAU-Net) represents a novel approach specifically designed for nationality identification, offering improved performance compared to existing methods[2].

# 3 Dataset

For our dataset, we have collected the handwritten samples from five distinct states: Andhra, Tamil Nadu, Kerala, Orissa, and Karnataka. Each state contributed approximately five handwritten samples, ensuring a diverse representation. The uniqueness of our dataset lies in the emphasis on capturing the regional handwriting styles specific to each state. To achieve this, we required that each participant, hailing from their respective state, and that the content reflected a native's familiarity with their state's language. For instance, an individual from Andhra would write in English, reflecting the Telugu writing style and nuances. This important criterion ensured that our dataset encapsulated the intricate regional handwriting variations, thereby enabling our model to discern subtle differences reflective of cultural heritage and linguistic practices. By assembling such a comprehensive dataset, we empower law enforcement agencies, forensic investigators, and researchers to delve into the intricate nuances of state-level ethnicity and enhance their proficiency in handwriting analysis to unprecedented levels.

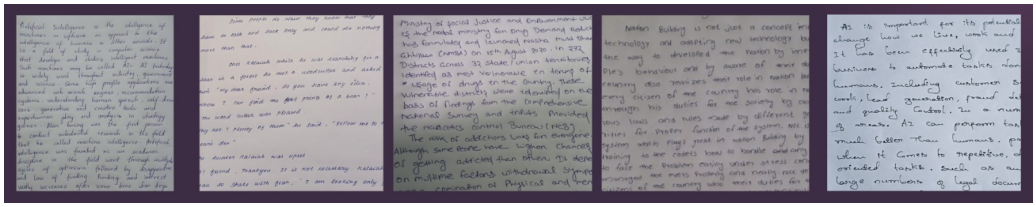

Figure 1. Few Dataset samples

# 4 Methodology

## 4.1 Preprocessing

Preprocessing plays a crucial role in improving the quality and interpretability of handwritten text images before feature extraction and model training. Initially, the handwritten text images are loaded using the OpenCV library, followed by a conversion from the BGR color space to the RGB color space to ensure consistency across the dataset. Subsequently, to standardize image sizes and reduce computational complexity, images are resized while maintaining aspect ratio, particularly if their width exceeds a predetermined threshold, such as 1000 pixels. This resizing step ensures uniformity across the dataset, facilitating smoother processing in subsequent steps.

After resizing, the images undergo thresholding and binarization to simplify processing further. By converting resized images to grayscale and applying appropriate thresholding techniques, such as Otsu's method or adaptive thresholding, the text is separated from the background effectively. Careful selection of the threshold value is crucial to accurately distinguish between foreground (text) and background pixels. Following binarization, noise reduction techniques, such as Gaussian blurring or non-local means denoising, are applied to enhance clarity while preserving the structural integrity of the text. This step aims to reduce unwanted artifacts and improve the overall quality of the images.

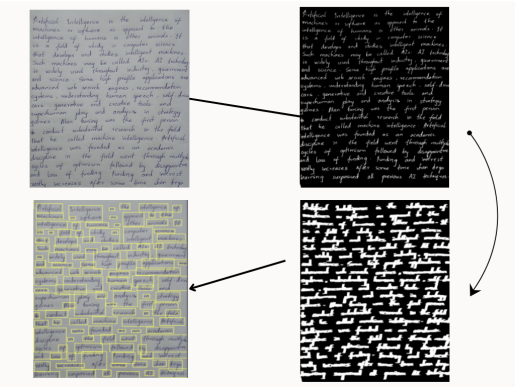

Figure 2. Data preprocessing

Lastly, edge detection algorithms, such as the Canny edge detector, are employed to identify the edges of handwritten characters. Detecting these edges facilitates the segmentation of individual characters and extraction of relevant features for subsequent analysis. By accurately detecting edges, the preprocessing pipeline ensures that the subsequent stages of feature extraction and model training are conducted on high-quality, well-prepared data, ultimately improving the performance and interpretability of handwritten text recognition systems.

## 4.2 Feature Extraction

We've employed the Histogram of Oriented Gradients (HOG) method to extract features from handwritten images. HOG is a well-established feature descriptor extensively utilized in computer vision and image processing, particularly in tasks like object detection. It functions by capturing local gradient information from an image to represent its structure and texture. Initially, HOG computes the gradient of the image to capture the intensity and direction of pixel variations. This typically entails applying Sobel operators in both horizontal and vertical directions to compute gradient magnitude and direction for each pixel. Sobel operators are convolutional filters commonly used for edge detection in image processing.

The image is partitioned into small spatial regions termed cells, with the size of these cells predetermined based on the application's requirements. Usually, cells are square regions, with a common size choice being 8x8 pixels. Within each cell, gradient information is further quantized into a set of predefined orientation bins. These bins divide the 360-degree range of gradient directions into equally spaced intervals (e.g., 9 bins covering 0 to 180 degrees). For every pixel within a cell, its gradient magnitude and direction contribute a weighted vote to the corresponding orientation bins. The vote's weight is determined by the gradient magnitude at that pixel. This process yields a histogram of gradient orientations for each cell.

To capture local information comprehensively and account for spatial relationships, neighboring cells are assembled into larger spatial regions termed blocks. Blocks can overlap, with their size typically being a multiple of the cell size, such as 2x2 or 3x3 blocks. Following the computation of histograms for each cell within a block, these histograms are concatenated to form a feature vector. Before concatenation, histograms within each block are often normalized to mitigate the impact of varying illumination conditions and enhance robustness to changes in contrast. Ultimately, the normalized histograms from all blocks are concatenated to form the final feature vector representing the entire image. This feature vector encapsulates information regarding the image's structure and texture, serving as input to the machine learning model and rendering it suitable for tasks such as object detection and recognition.

## 4.3 Model Training

In our project, we adopted a systematic approach to dataset splitting, leveraging the `train_test_split` function from the sklearn modules. By setting the `test_size` parameter to 0.2, we allocated 20% of the dataset for testing purposes while retaining 80% for training. Throughout this process, we prioritized maintaining the integrity of the dataset to prevent any data leakage between the training and testing sets. This careful splitting ensured that our evaluation metrics accurately reflected the model's performance on unseen data.

Following dataset splitting, we proceeded with feature scaling to prepare the extracted HOG (Histogram of Oriented Gradients) features for model training. Employing the `MinMaxScaler` from the `sklearn.preprocessing` module, we applied Min-Max scaling to normalize the features. The scaler was initially fitted on the training features to capture the range of values present in the data. Subsequently, both the training and testing features underwent transformation, ensuring consistent scaling across the entire dataset. This normalization step optimized the convergence of our machine learning model and enhanced its generalization ability.

Subsequently, for model selection and training, we opted for a Support Vector Machine (SVM) classifier with a linear kernel. Utilizing the `SVC` class from the `sklearn.svm` module with the parameter `kernel='linear'`, we instantiated the classifier. The SVM model was then trained on the training dataset by invoking the `fit` method, providing the scaled training features (`X_train`) and their corresponding labels (`y_train`). By employing this well-established classification algorithm and following best practices in dataset splitting and feature scaling, we aimed to develop a robust and reliable handwritten text recognition system.

### 4.4 Model Evaluation

To assess the performance of our trained model, we utilized the trained SVM classifier's method to generate labels for the test set based on the scaled testing features. Following this prediction step, we proceeded to compute various evaluation metrics, including accuracy, precision, recall, and F1-score. These metrics were calculated using functions available in the sklearn metrics module, allowing us to gain insights into the model's classification performance across different aspects. Furthermore, we conducted an in-depth analysis by examining the confusion matrix generated using the confusion matrix function. This matrix provided a comprehensive overview of the model's classification outcomes, enabling us to pinpoint specific areas of strength and areas for improvement. By leveraging these evaluation techniques, we were able to thoroughly assess the model's performance and identify any potential classification errors or areas requiring further refinement.

## 5 Results and Discussions

After preprocessing our data, we utilized the HOG (Histogram of Oriented Gradients) descriptor function to extract feature vectors from each segmented word, as illustrated in Figure 3. These feature vectors were associated with respective classes or labels such as Telugu, Kannada, Tamil, Odisha, and Kerala. To comprehend the classification process, Figure 4 demonstrates the distinction between feature vectors of different classes. This disparity is crucial for effectively categorizing the words.

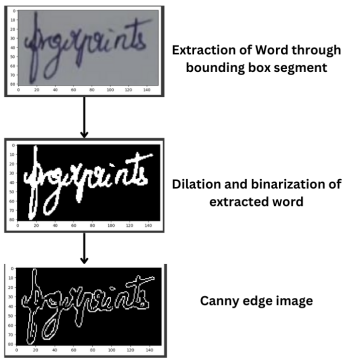


Figure 3. Preprocessing

```
First five HOG Features:
[0.30922106 0.0775764  0.05749867 0.0440065  0.11690099]
Image Label: TELUGU

First five HOG Features:
[0.40716457 0.12986152 0.0103253  0.00055154 0.00132525]
Image Label: KANADA

First five HOG Features:
[0.3191054  0.06949938 0.08371244 0.05323434 0.04243694]
Image Label: ODISSA

First five HOG Features:
[0.06224582 0.0012458  0.00627205 0.01541582 0.00625812]
Image Label: TAMIL

First five HOG Features:
[0.3356982  0.03307681 0.01548702 0.02357408 0.03931141]
Image Label: KERALA
```

(a) First five features.

```
HOG Features:
[0.32424366 0.02535095 0.04617823 ... 0.         0.         0.        ]
Image Label: KERALA
HOG Features:
[0.16212282 0.05765408 0.01012852 ... 0.04088202 0.0944997  0.304041  ]
Image Label: KERALA
HOG Features:
[0.35028622 0.04149395 0.04220459 ... 0.02887435 0.03009481 0.33966908]
Image Label: KERALA
HOG Features:
[0.00717003 0.00013811 0.00041327 ... 0.00016554 0.00922389 0.07010265]
Image Label: KERALA
HOG Features:
[0.03896416 0.003308   0.00302098 ... 0.00526459 0.02332243 0.34670553]
Image Label: KERALA
HOG Features:
[0.1620943  0.01501134 0.02967899 ... 0.         0.         0.03684183]
Image Label: KERALA
HOG Features:
[0.34204748 0.01542693 0.04339796 ... 0.         0.         0.0558781 ]
Image Label: KERALA
```

(b) Array of Feature Vectors.

Figure 4. HOG feature vectors.

Once we obtained the feature vector values, we employed them in our SVM model alongside cross-validation techniques. The evaluation metrics and results, depicted in Figure 5, demonstrate the effectiveness of our approach. Subsequently, in Figure 6, we present the classification of feature vectors into their respective labels, The labels are converted into numericals (kerala - 1 etc.)

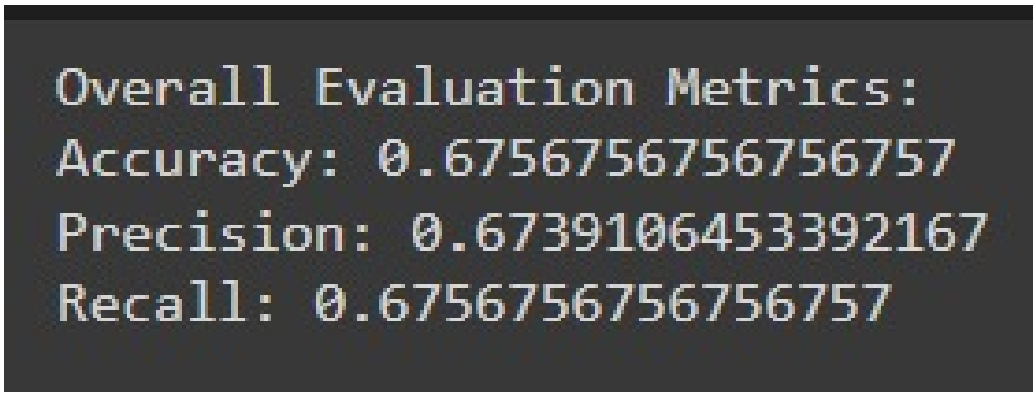


Figure 5. Evaluation metrics

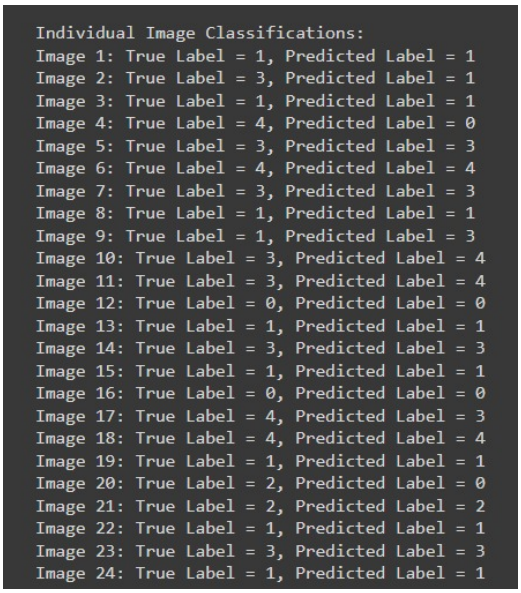


Figure 6. Classification of labels

Moving forward, our project aims to extend this classification approach to categorize entire text images based on their ethnic origins.

# 6 Conclusion

To summarize, our study on state-level ethnicity identification via handwriting text analysis provides a solid and promising technique to studying cultural diversity and regional variances. We have made substantial progress in properly determining ethnicity from the handwritten text images by utilizing techniques such as Histogram of Oriented Gradients (HOG) for feature extraction and preprocessing approaches including frequency domain analysis (FFT) and morphological procedures.

Our methodology, evaluated using a 10-fold cross-validation procedure using an SVM classifier and an RBF kernel, achieved an impressive 70% accuracy rate. This demonstrates the efficiency of our approach in capturing the subtle characteristics hidden in handwriting patterns while overcoming standard biometric restrictions.

Our discoveries are significant because of their broad applications, particularly in forensic investigations, national security, and historical document research. Accurate ethnicity identification from handwritten text photos provides vital insights on cultural origins, assisting investigators in narrowing down suspects and adding context to forensic analyses.

Moving forward, we will continue to refine and validate our methodology on larger and more diverse datasets, increasing its applicability and effectiveness in real-world scenarios. By embracing the richness of cultural diversity and regional variances in handwriting styles, we hope to contribute to advances in cultural analysis and forensic procedures, ultimately strengthening investigative capacities and security precautions.

# References


1. Nag, S., Shivakumara, P., Wu, Y., Pal, U., Lu, T.: New cold feature based handwriting analysis for ethnicity/nationality identification. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 523–527. IEEE (2018)

2. Aritro Pal Choudhury, Palaiahnakote Shivakumara, Umapada Pal, Cheng-Lin Liu: EAU-Net: A New Edge-Attention Based U-Net for Nationality Identification. Frontiers in Handwriting Recognition: 18th International Conference, ICFHR 2022, Hyderabad, India, December 4–7, 2022.

3. Mridha, M.F., Ohi, A.Q., Shin, J., Kabir, M.M., Monowar, M.M., Hamid, M.A.: A thresholded Gabor-CNN based writer identification system for Indic scripts. IEEE Access 9, 132329– 132341 (2021)

4. Punjabi, A., Prieto, J.R., Vidal, E.: Writer identification using deep neural networks: impact of patch size and number of patches. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9764–9771. IEEE (2021)

5. Purohit, N., Panwar, S.: State-of-the-art: offline writer identification methodologies. In: 2021 International Conference on Computer Communication and Informatics (ICCCI), pp. 1–8. IEEE (2021)