OPTICAL CHARACTER RECOGNITION

Faculty: Dr. Vinayak Abrol

Machine Learning Project

By: Prem Kamal Jain

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

PROBLEM STATEMENT

- **OCR's GOAL:** Convert Text Images to Text Data.
- **KEY USES :** Archive digitization, automated data entry.
- **MAIN CHALLENGE:** Achieving consistent accuracy due to:
  - Complex text layouts.
  - Diverse font styles and sizes.
  - Varied text orientations and backgrounds.
- **PROJECT FOCUS:**
  - Overcome these accuracy limitations in OCR.
  - Approach: Implement different machine learning models.
- **GOAL:** Identify the model with the highest accuracy for OCR tasks.

PROPOSED SOLUTION

- Develop OCR system without template matching.
- Testing various machine learning models.
- Models include:
  - K-Means: Clustering and pattern recognition.
  - SVM: Classification tasks in complex text.
  - KNN: Effective pattern recognition.
  - GMM: Categorizing text styles/backgrounds.
  - Random Forest: Boosting predictive accuracy.
- Test and compare models for best OCR results
- Selecting the model with top accuracy and efficiency.

DATASET DESCRIPTION

- The dataset is divided into two primary sets: "data" and "data2", each containing 'training_data' and 'testing_data' folders.
- These images, capturing varied text forms and fonts, are used for training the machine learning model and evaluating its performance against contemporary standards
- Each folder has 573 image.
- Images have class name from 0- 9 and A-Z.

METHODOLOGY

- **FOCUS:** Create OCR system for diverse data, no template matching.
- **DATA PROCESSING:** Enhance OCR data clarity; adapt to various text and images.
- **MODEL APPLICATION:** Implement and customize multiple ML models for text.
- **PERFORMANCE ANALYSIS:** Evaluate accuracy across text types and environments.
- **OPTIMIZATION:** Fine-tune for accuracy; adapt to OCR challenges.
- **TESTING:** Conduct rigorous tests; benchmark against standards.
- **IMPROVEMENT:** Refine continually based on results and feedback.

RESULT

Performance Comparison of Machine Learning Algorithms

Support Vector Machine (SVM)

- Accuracy: 98.1%
- Highlight: Exceptionally effective for high-dimensional data spaces.

K-Nearest Neighbors (KNN)

- Accuracy: 95.83%
- Highlight: Ideal for scenarios where data interpretation is straightforward.

K-Means Clustering

- Accuracy: 85%
- Highlight: Best suited for quick exploratory data analysis.

Random Forest Algorithm

- Accuracy: 98.71%
- Highlight: Provides high accuracy through decision tree ensemble.

COMPARISON WITH EXISTING ANALYSIS

Model Used	Accuracy According to Existing Analysis	Accuracy according to Models created by us
K Means	81.6%	85%
KNN	93.96%	95.83%
SVM	92%	98.1%
Random Forest	91%	98.71%

CONCLUSION

- Successfully applied machine learning models in OCR.
- Achieved high accuracy in English languages and Numbers with different handwriting styles.
- Demonstrated effectiveness of K-Means, SVM, KNN, and Random Forest in OCR.
- Showcased OCR's potential beyond traditional methods.
- Set the stage for future innovations in text recognition and digitization.
- Achieved highest accuracy in Random Forest that is 98.71%.