CSD3185 Team Project
Group 5

Chiok Wei Wen Gabriel
Jed Goh Yujie
Seow Kai Jun
Yin Shengkai
Tay Khai Tjong Dennys
Huang Wei Jhin

Background

The proliferation of digital images necessitates effective organization and retrieval systems. Manual categorization is time-consuming and prone to inconsistency, highlighting the need for automated image classification systems. Machine learning offers robust solutions by enabling computers to recognize patterns and categorize images based on their features. This project aims to leverage machine learning to automate the organization of images into predefined categories, enhancing accessibility and manageability of digital assets.

Application Description

The application developed in this project will automatically sort images into user-specified folders based on the content recognized within each image. Users will define categories, and the application will employ an image recognition model to assign labels to each image. These labels will then be used to sort the images into the corresponding folders.

Workflow:

Data Collection and Labeling: Images will be gathered, and labels will be generated either manually or through a pre-trained image recognition model. The generated labels should accurately reflect the content of the images to ensure effective categorization.

Model Training:

The corresponding labels will be used to train three different machine learning models: KNN, Random Forest, and SVM. Each model will learn to associate the features of the images with the appropriate categories.

Classification and Sorting:

Once trained, the models will classify new images based on their features and sort them into the predefined folders according to their predicted labels.

Dataset

For the data, any dataset that contains many images to train will suffice. An example would be this link.

https://www.kaggle.com/datasets/meherunnesashraboni/multi-label-image-classification-dataset

The link above features a dataset of images already categorized to their labels. This would help in training to ensure that the classification algorithm is accurate. There are many other datasets on Kaggle that could be used as well but primarily we will use the link until we need more data for training or testing.

Algorithms

K-Nearest Neighbors (KNN):

This algorithm will classify images based on the similarity of their features to those of training samples. KNN is intuitive and simple yet effective for image classification, especially when the dataset is not too large. However, its performance may degrade with very large datasets or high-dimensional data due to the curse of dimensionality.

Random Forest:

This ensemble learning method uses multiple decision trees to make predictions, reducing the risk of overfitting associated with single decision trees. It is robust and capable of handling large datasets with higher dimensionality. For image classification, Random Forest can effectively handle the complexity and variance in image data.

Support Vector Machine (SVM):

SVM is particularly suited for binary classification tasks. It works by finding the hyperplane that best separates different categories in the feature space. For multi-class image categorization, one-vs-all or one-vs-one strategies can be employed. SVM is known for its effectiveness in high-dimensional spaces, making it suitable for image data where each pixel may represent a dimension.

Timeline

Week 8	Project Planning, Creation of the Project Proposal
Week 9	Start collecting or accessing the necessary image datasets. Find a suitable image recognition tool or model.
Week 10	Clean and preprocess image data, perform feature extraction, and begin initial exploratory data analysis to understand the dataset's characteristics and prepare it for modeling.
Week 11	Implement the KNN, Random Forest, and SVM algorithms, train the models using the training dataset, and conduct preliminary testing and performance evaluation on the development set.
Week 12	Refine the models based on initial results, tune parameters, and perform extensive testing. Compare the algorithms' performance meticulously to identify the best-performing model on the test dataset.
Week 13	Creation of Presentation Slides, Recording of Presentation Video, Bug Fixing and Submission of Machine Learning Project