CorelDB Classification Using K-NN with Different Methods of Feature Extraction

Bianca-Alexandra Ghiorghiu, Maria-Ecaterina Constantinescu

***Abstract*—This study presents a comparison of different methods of feature extraction for classifying the CorelDB dataset using the k-nearest neighbors (K-NN) algorithm and provides insight into the effectiveness of different feature extraction techniques for image classification tasks. The project contains a statistical analysis for the dataset, as well as different methods of feature extraction for the images. In addition, a comparison with other classification algorithms such as support vector machine (SVM), logistic regression and random forest was made.**

***Index Terms*—K-NN, Random Forest, SVM, Logistic Regression, PCA, HOG, Color Histogram, DCT, Co-occurrence Matrix**

I. Introduction

THE success of image classification, which is the task of identifying and classifying objects within an image, is heavily dependent on the way in which the images are represented. This representation, known as the feature representation, is crucial for the classification algorithm to be able to accurately distinguish between different classes of objects. In the case of the dataset used in this study, it was found to be particularly challenging and resulted in poor accuracy when using the k-nearest neighbors (K-NN) algorithm with principal component analysis (PCA) as the feature extraction method.

To improve the results, it is necessary to explore different methods of feature extraction and other classification algorithms. This will involve comparing the performance of various techniques in order to determine which ones are most effective in accurately classifying the objects within the images. The comparison will be made in the following sections of the study and will hopefully lead to a better understanding of the best practices for image classification and improved results.

II. Theoretical Background

A. Statistical Analysis

Statistical moments such as mean and variance can be calculated for an image's color histogram. The mean represents the average color value in the image, while the variance represents the spread of color values around the mean [1].

The Kolmogorov-Smirnov test (K-S test) is a nonparametric statistical test that can be used to compare two probability distributions. It can be used to compare the color histograms of two images and determine if they are significantly different [2].

A color histogram is a graphical representation of the distribution of colors in an image. It plots the number of pixels of each color in an image, providing a rough idea of the color composition of the image. These techniques can be used to analyze image data and extract meaningful information. For example, color histograms can be used to classify images into different categories, while statistical moments and K-S tests can be used to compare images and detect differences or similarities.

B. Feature Extraction

Histogram of Oriented Gradients (HOG) is a feature descriptor used in computer vision for object detection. It works by analyzing the gradients of pixel intensities in an image, and creating a histogram of gradient orientations. HOG is particularly useful for detecting objects in images with complex backgrounds, as it is able to emphasize edges and lines that are indicative of an object [3].

Discrete Cosine Transform (DCT) is a mathematical technique used to transform a signal into a new representation. In image processing, it is used to compress image data by removing redundant information. DCT is used in JPEG image compression and is a lossy compression method [4].

Co-occurrence matrix, also known as a gray-level co-occurrence matrix (GLCM), is a method used to extract texture features from an image [5]. It is a matrix that describes the relationship between the gray level values of pixels in an image. It can be used to identify patterns and textures in an image, and has applications in image segmentation and object recognition.

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a dataset. It works by finding the principal components of the data, which are the directions of maximum variance [6]. In image processing, PCA is often used to extract features from images and represent them in a lower dimensional space. It is also used in image compression and object recognition.

C. Classification Algorithms

SVM (Support Vector Machine) is a supervised learning algorithm that can be used for classification and regression tasks. It works by finding the best boundary (a hyperplane) that separates the different classes in a dataset. The boundary is chosen so that it maximizes the margin, which is the distance between the boundary and the closest data points from each class. SVM can be used for non-linearly separable data by using kernel trick [7].

k-NN (k-Nearest Neighbors) is a simple non-parametric algorithm used for classification and regression. It works by finding the k closest data points to a given test point, and then classifying the test point based on the majority class of the k nearest data points. The value of k is a user-defined parameter, usually set to an odd number to prevent ties.

Random Forest is an ensemble learning method used for both classification and regression. It works by creating multiple decision trees and combining their predictions. Each tree is built using a random subset of the data and a random subset of the features, which helps to reduce overfitting [8]. The final prediction is made by averaging the predictions of all the trees.

Logistic Regression is a statistical method used for classification. It works by fitting a logistic function to the data, which models the probability of a certain class given a set of input features. The logistic function produces a probability value between 0 and 1, which can be thresholded to produce a binary classification. Logistic regression can also be extended for multi-class classification problem [9].

III. About the Dataset

The dataset contains 10,800 images with a resolution of either 80x120 pixels or 120x80 pixels. The images are labeled and have a total of 80 classes.

To test the different methods of classification, the whole dataset was firstly split into 80% training set and 20% test set. Because the dataset is very diverse, this resulted in a very low overall accuracy, so to improve the results, we grouped the 80 classes into 10 classes based on the same type of images.

To gain some insight about how the total accuracy is affected by different classes we also tried to get the accuracy for subclasses for the images of the same type to see how well the methods can differentiate between one another. The classes for This split are as follows: art (with 5 subclasses), buildings (with 4 subclasses), food (2 subclasses), objects (21 subclasses), people (3 subclasses), pet (2 subclasses), plants (3 subclasses), scenery (14 subclasses), texture (6 subclasses), wild life (20 subclasses).

For the third method of analysis we merge the different sublclasses creating only 10 classes as follows: animals (wild life and pets merged), art, buildings, food, objects, people, plants, scenery, texture, vehicles (some of the objects were transferred here to try and balance the classes). The percentage for the splitting remains the same and all the images will be resized to 80x80 pixels for all the three methods.

IV. Implementation

The practical implementation is done in Python using the sklearn library. The first part of the code consists of the statistical moments, tests and representation of the color histogram, while the second part consists of the implementation of the classification algorithms using different types of feature extraction.

The general principle of the code can be seen in the figure 1.

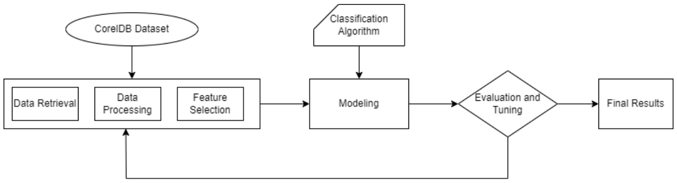

Fig. 1: Flowchart of the Implementation

V. Statistical Analysis

A probability density function (PDF) is a mathematical function that describes the probability of a continuous random variable taking on a particular value. The probability density function assigns a probability to each possible value of the random variable, such that the total probability of the random variable falling within a given range is equal to the integral of the probability density function over that range [10]. In other words, a probability density function gives the probability of a random variable taking on any value within a continuous range of values, rather than a specific value.

For this representation, the fourth image from the class art_1 was chosen:

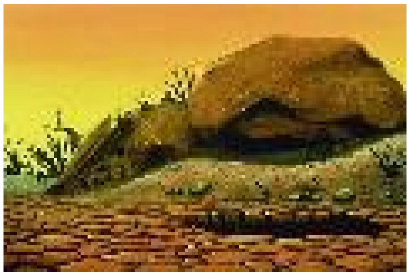

Fig. 2: The 4th Image from the art_1 class

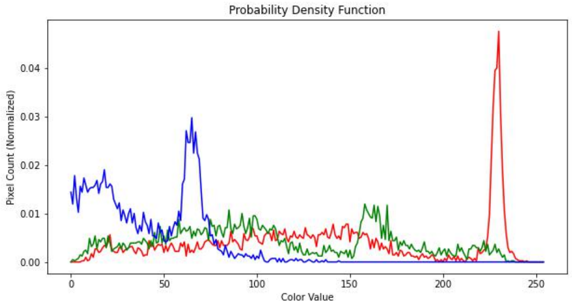

Fig. 3: Probability Density Function

A color histogram is a representation of the distribution of colors in an image. It is a graphical representation of the number of pixels in an image that have colors within various color ranges. The x-axis of the histogram represents the colors in the image, and the y-axis represents the number of pixels that have that color [11]. The colors are grouped into a set number of color channels - red, green, and blue (RGB), and a separate histogram is generated for each channel.

The color histogram was generated for the same image, as seen in 4.

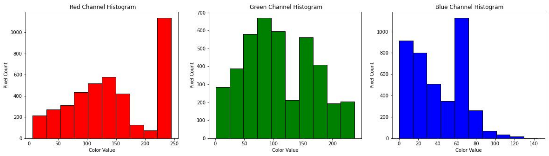

Fig. 4: Color Histogram

The mean and variance per class refer to the average and spread of a particular variable within each category or class in a dataset. The mean represents the central tendency of the data, while the variance measures the spread or dispersion of the data [12].

By interpreting both Figure 4 and Table I, it can be seen that both the Red and Green channel are more spread out and have a higher mean, while the Blue channel has a smaller mean value and variance since it doesn't even reach the maximum value for the pixel intensity interval. By looking at the actual image from Figure 2, it can be seen that the green and red tones are more predominant than the blue tones.

Channel	Mean	Variance
Red	142.13	4600.49
Green	109.20	3446.96
Blue	40.44	723.98
Gray-scale	111.21	3207.78

TABLE I: Mean and Variance for the 4th Art 1 Image

For the graphical representation the merged dataset was used in Figure 5 and 6. The variables for the art, people and food classes are the most spread out from all the classes on al three channels.

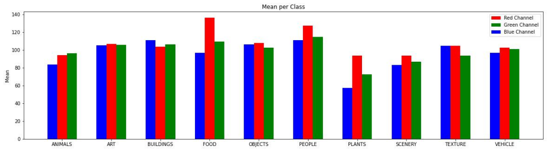

Fig. 5: Mean per Class for the Merged Dataset

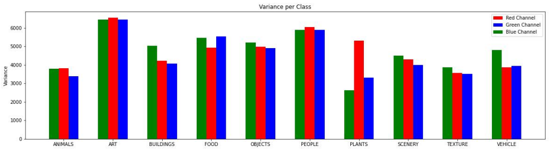

Fig. 6: Variance per Class for the Merged Dataset

The distribution for the gray-scale image was plotted in Figure 7. Gray-scale refers to the range of shades of gray in an image, with black being the darkest and white being the lightest. The distribution of gray-scale values in an image can provide information about the overall brightness and contrast of the image, as well as the presence of certain features or patterns. From the graph it can be seen that it resembles an exponential distribution and has the lowest pvalue for this distribution for the Kolmogorov-Smirnov statistical test in Table II

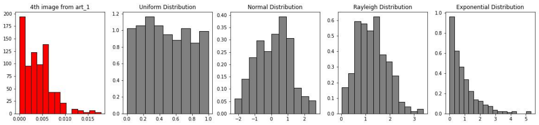

Fig. 7: Distribution for the 4th Image from Art 1

Distribution	K-S Test Statistic	pvalue
Normal	0.80	1.20e-83
Uniform	0.84	8.64e-93
Rayleigh	0.79	8.82e-81
Exponential	0.77	2.70e-75

TABLE II: K-S Test

VI. Feature Extraction and Classification

A. *Classification for the Whole Dataset*

For this section, a evaluation of various parameters was conducted using the grid search method to determine the optimal settings for each classification algorithm. Specifically, the SVM algorithm was optimized by experimenting with different kernel functions and regularization parameters. The random forest algorithm was tuned by varying the number of estimators. The logistic regression algorithm was optimized by adjusting the regularization parameters. Lastly, the K-NN algorithm was fine-tuned by testing different distance metrics and varying the number of neighbors.

The methods that were used for feature extraction are: color histogram, HOG, DCT and co-occurence matrix.

	Train accuracy	Test accuracy
k-NN	34.8%	36.2%
Logistic Regression	30.4%	22.7%
SVM	35.1%	31.2%
Random Forest	100%	38.8%

TABLE III: PCA with 200 componentes on flattened image

	Train accuracy	Test accuracy
k-NN	53%	38.5%
Logistic Regression	46.4%	39.2%
SVM	63.4%	41.7%
Random Forest	100%	49.7%

TABLE IV: HSV Color Histogram, 4 Bins per channel

	Train accuracy	Test accuracy
k-NN	100%	43.1%
Logistic Regression	65.7%	40.7%
SVM	98.2%	43.1%
Random Forest	100%	50.3%

TABLE V: HSV Color Histogram, 10 Bins per channel

	Train accuracy	Test accuracy
k-NN	53.6%	43.5%
Logistic Regression	79.5%	39.4%
SVM	99.6%	40.1%
Random Forest	100%	47.5%

TABLE VI: HSV Color Histogram, 16 Bins per channel

	Train accuracy	Test accuracy
k-NN	30.8%	22.7%
Logistic Regression	55.2%	19.8%
SVM	97.3%	36.3%
Random Forest	100%	33.1%

TABLE VII: Histogram of Oriented Gradients

	Train accuracy	Test accuracy
k-NN	38.2%	26.7%
Logistic Regression	31.9%	20.4%
SVM	99.9%	30.6%
Random Forest	100%	30.6%

TABLE VIII: Discrete Cosine Transform, 100 coefficients

	Train accuracy	Test accuracy
k-NN	36.5%	25%
Logistic Regression	78.7%	11.6%
SVM	98.2%	30.7%
Random Forest	100%	26.2%

TABLE IX: Discrete Cosine Transform, 500 coefficients

	Train accuracy	Test accuracy
k-NN	43.7%	31.3%
Random Forest	100%	38.5%

TABLE X: Co-occurance Matrix

B. Classification for the Subclasses

The 80 classes were merged into 10 bigger classes as mentioned in the ”About the Dataset” section. The feature extraction method used was the HSV color histogram with 10 bins for each channel, as it provided the best results. Then, for each of the ten obtained classes, a grid search was performed to identify the best hyperparameters for the four implemented algorithms, and the accuracy on the test set was computed using the models with the best found parameters.

As it can be seen from the tables, the accuracy for the texture, scenery, and building classes is lower compared to the other classes. This indicates that these classes may be contributing to the overall low accuracy of the dataset. It may be worth further investigating the performance on these classes and addressing any issues that may be causing the low accuracy.

	Train accuracy	Test accuracy
k-NN	94.9%	95%
Logistic Regression	100%	83.3%
SVM	96.25%	90%
Random Forest	100%	91.6%

TABLE XI: Art

	Train accuracy	Test accuracy
k-NN	68.2%	55.8%
Logistic Regression	69.3%	56.5%
SVM	73.1%	50.6%
Random Forest	100%	54.5%

TABLE XII: Buildings

	Train accuracy	Test accuracy
k-NN	100%	85.7%
Logistic Regression	95.7%	76.5%
SVM	100%	83.3%
Random Forest	100%	83.5%

TABLE XIII: Food

	Train accuracy	Test accuracy
k-NN	100%	70.5%
Logistic Regression	91.9%	60.3%
SVM	99.6%	68.1%
Random Forest	100%	72.9%

TABLE XIV: Objects

	Train accuracy	Test accuracy
k-NN	100%	97.1%
Logistic Regression	99.4%	97.6%
SVM	99.8%	95.9%
Random Forest	100%	97.1%

TABLE XV: People

	Train accuracy	Test accuracy
k-NN	89.3%	85.6%
Logistic Regression	90.7%	84.8%
SVM	89.4%	84.8%
Random Forest	100%	88%

TABLE XVI: Pets

	Train accuracy	Test accuracy
k-NN	81.1%	73.8%
Logistic Regression	86.7%	75.7%
SVM	84.5%	73.8%
Random Forest	100%	79.2%

TABLE XVII: Plants

	Train accuracy	Test accuracy
k-NN	68.2%	58.3%
Logistic Regression	75.7%	58.9%
SVM	72.7%	58.2%
Random Forest	100%	66.2%

TABLE XVIII: Scenery

	Train accuracy	Test accuracy
k-NN	63.9%	54.1%
Logistic Regression	79.9%	58.2%
SVM	70.5%	57%
Random Forest	100%	63.9%

TABLE XIX: Texture

	Train accuracy	Test accuracy
k-NN	100%	48.3%
Logistic Regression	71.6%	47.3%
SVM	91.2%	49%
Random Forest	100	55.7%

TABLE XX: Wildlife

The results presented in tables XIV, XVII, XVIII, XIX and XX may be attributed to the inconsistencies present in the images utilized for classification. For example, the 'molecule' class exhibits a diverse array of patterns that may pose a challenge for accurate detection and classification. This variability in the visual characteristics of the molecular class may have contributed to the observed results.

On the other hand, the imbalance in the distribution of data within the scenery classes poses a significant challenge for the model, as it increases the likelihood of overfitting. This is particularly concerning given the limited number of images available for each class. Furthermore, this issue of imbalanced data may also extend to the subclasses within the scenery category, potentially resulting in suboptimal performance. To reduce these problems, it may be necessary to employ data augmentation techniques in order to balance the distribution of images across classes and prevent overfitting.

C. Classification for the Merged Classes

The training and testing was done using only the color histogram for feature extraction. This decision was made because in previous tests, it was found that using the color histogram as a feature had the best overall results.

	Train accuracy	Test accuracy
k-NN	65.5%	56.3%
Logistic Regression	44%	42.5%
SVM	45.2%	43.5%
Random Forest	100%	59.5%

TABLE XXI: HSV Color Histogram, 10 Bins per channel

The results of the training and testing process are presented in Table XXI. As can be observed from the table, the accuracy of all the algorithms used in the model improved compared to the one for the whole dataset. Among all the algorithms, Random Forest was found to be the most accurate. As previously discussed, the imbalanced nature of the data within the previously mentioned classes results in a decrease in overall accuracy. Additionally, even with data augmentation techniques, some images may still be difficult to classify due to factors such as poor lighting, angles, or image quality.

VII. Conclusions

Feature extraction is a crucial step in the machine learning algorithms that can greatly impact the performance of a model. There are various methods of feature extraction that can be used, each with their own strengths and weaknesses and are a powerful tool that can greatly enhance the performance of a model by reducing the dimensionality of the data, improving the interpretability of the features, and providing a more compact representation of the data.

However, it is important to note that even with the correct feature extraction method, dealing with imbalanced data can still be a challenging task. Imbalanced classes occur when the classes in a dataset are not equally represented and can lead to poor performance and bias in predictive models. This is because most machine learning algorithms are designed to maximize overall accuracy, which can lead to a bias towards the majority class. As a result, it is crucial to not only consider the feature extraction method but also to address the class imbalance problem in order to achieve optimal performance and unbiased result

References

[1] Li, Yanmei, and Sumei Zhang. *"Statistical Analysis."* In Applied Research Methods in Urban and Regional Planning, pp. 109-148. Springer, Cham, 2022.

[2] Berger, Vance W., and YanYan Zhou. *"Kolmogorov–smirnov test: Overview."* Wiley statsref: Statistics reference online (2014).

[3] Zhou, Wei, Shengyu Gao, Ling Zhang, and Xin Lou. *"Histogram of oriented gradients feature extraction from raw Bayer pattern images."* IEEE Transactions on Circuits and Systems II: Express Briefs 67, no. 5 (2020): 946-950.

[4] Hossain, Md Tahmid, Shyh Wei Teng, Dengsheng Zhang, Suryani Lim, and Guojun Lu. *"Distortion robust image classification using deep convolutional neural network with discrete cosine transform."* In 2019 IEEE International Conference on Image Processing (ICIP), pp. 659-663. IEEE, 2019.

[5] Garg, Meenakshi, and Gaurav Dhiman. *"A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants."* Neural Computing and Applications 33, no. 4 (2021): 1311-1328.

[6] Kherif, Ferath, and Adeliya Latypova. *"Principal component analysis."* In Machine Learning, pp. 209-225. Academic Press, 2020.

[7] Cervantes, Jair, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua, and Asdrubal Lopez. *"A comprehensive survey on support vector machine classification: Applications, challenges and trends."* Neurocomputing 408 (2020): 189-215.

[8] Genuer, Robin, Jean-Michel Poggi, Robin Genuer, and Jean-Michel Poggi. *"Random forests."* Springer International Publishing, 2020.

[9] LaValley, Michael P. *"Logistic regression."* Circulation 117, no. 18 (2008): 2395-2399.

[10] Lee, Yong-Gu, and Sam-Yong Kim. *"Introduction to statistics."* Yulgokbooks, Korea (2008): 342-351.

[11] Chakravarti, Rishav, and Xiannong Meng. *"A study of color histogram based image retrieval."* In 2009 Sixth International Conference on Information Technology: New Generations, pp. 1323-1328. IEEE, 2009.

[12] Witte, Robert S., and John S. Witte. *Statistics.* John Wiley & Sons, 2017.