

2A. Cropping and Resizing Dog Images from XML Annotations

Using XML Processing and PIL

This code snippet demonstrates how to extract bounding box information from XML annotation files, crop images of dogs from a specified directory using these bounding boxes, and resize each cropped image to 128×128 pixels. The cropped images are saved in a designated output directory.

```
In [2]: import os
import glob
import numpy as np
import matplotlib.pyplot as plt
from skimage import io, color, filters, feature
from sklearn.metrics import pairwise
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer
import json
import os
import numpy as np, pandas as pd
import matplotlib.pyplot as plt
import matplotlib.image as implt
import xml.etree.ElementTree as ET
import glob
from PIL import Image
from pathlib import Path
from PIL import Image, ImageEnhance

In [3]: # Define directories
dog_images_dir = glob.glob('/Users/thota/Documents/Datamining/Datasets/MyDogs/*/*')
annotations_dir = glob.glob('/Users/thota/Documents/Datamining/Datasets/MyAnnos/*/*')
cropped_images_dir = '/Users/thota/Documents/Datamining/Datasets/Cropped/'

# Function to parse bounding boxes from annotation files
def get_bounding_boxes(annot):
    tree = ET.parse(annot)
    root = tree.getroot()
    objects = root.findall('object')
    bbox = []
    for o in objects:
        bndbox = o.find('bndbox')
        xmin = int(bndbox.find('xmin').text)
        ymin = int(bndbox.find('ymin').text)
        xmax = int(bndbox.find('xmax').text)
        ymax = int(bndbox.find('ymax').text)
        bbox.append((xmin, ymin, xmax, ymax))
    return bbox

# Function to get image path based on the annotation file
def get_image(annot):
    img_path = '/Users/thota/Documents/Datamining/Datasets/MyDogs/'
    path_parts = annot.split('\\')
    folder = path_parts[-2]
    filename = f"{path_parts[-1].replace('.xml', '')}.jpg"
    img_filename = os.path.join(img_path, folder, filename)
    return img_filename

# Process the images
print(len(dog_images_dir))
for i in range(len(dog_images_dir)):
    bbox = get_bounding_boxes(annotations_dir[i])
    dog = get_image(annotations_dir[i])
    im = Image.open(dog)
    print(im)

# Crop and resize the bounding boxes
for j in range(len(bbox)):
    im2 = im.crop(bbox[j])
    im2 = im2.resize((128, 128), Image.Resampling.LANCZOS)
    new_path = dog.replace('MyDogs', 'Cropped')
    new_path = new_path.replace('.jpg', '-' + str(j) + '.jpg')

# Create directories if they don't exist
head, tail = os.path.split(new_path)
Path(head).mkdir(parents=True, exist_ok=True)

# Save the cropped image
im2.save(new_path)
```

<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x332	at	0x25911C5C05>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=333x500	at	0x25914F63C5>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=333x500	at	0x25914F63FD0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=300x293	at	0x25914F63950>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=350x300	at	0x25914F62E10>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x332	at	0x25914F63390>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x332	at	0x25914F6CFD0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x335	at	0x25914F6D290>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6D3D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F6D550>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x465	at	0x25914F6D2D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=400x300	at	0x25914F6D850>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F6DB90>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x412	at	0x25914F6DBD0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6E190>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F6E450>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6C550>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F6E6D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x342	at	0x25914F6E910>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x414	at	0x25914F6E950>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6E890>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x347	at	0x25914F6EE50>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=296x360	at	0x25914F6ED90>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6F350>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x334	at	0x25914F6E890>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=333x500	at	0x25914F6F5D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F6F990>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6FC50>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6FE90>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=333x500	at	0x25914F6FF50>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6F890>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=360x256	at	0x25914F6E290>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F6F7D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F6E390>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F74690>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=222x200	at	0x25914F74D90>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=200x164	at	0x25914F74E90>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F743D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x357	at	0x25914F74ED0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=357x500	at	0x25914F74A50>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=250x210	at	0x25914F75110>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F751D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=240x360	at	0x25914F75250>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=250x270	at	0x25914F75290>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=375x500	at	0x25914F742D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F75410>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x332	at	0x25914F75510>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=473x500	at	0x25914F74A10>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x333	at	0x25914F75650>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F74490>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=500x375	at	0x25914F75790>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=320x240	at	0x25914F74810>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=471x500	at	0x25914F75190>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=357x500	at	0x25914F749D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size=166x197	at	0x25914F741D0>
<PIL.JpegImagePlugin.JpegImageFile	image	mode=RGB	size		

file:///C:/Users/thota/Downloads/DMAssignment Sricharan Thota (1).html

file:///C:/Users/thota/Downloads/DMAssignment Sricharan Thota (1).html

<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F83FD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x374 at 0x25914F83D90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914A01E10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x368 at 0x25914F80FD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F80F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=325x500 at 0x25914F83390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=332x500 at 0x25914718510>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x325 at 0x25914B9DF50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x332 at 0x25914F81390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F88E90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F88F50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=375x500 at 0x25914F88F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=324x500 at 0x25914F88FD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25913E6B6D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x225 at 0x25914F88490>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=308x410 at 0x25914F635D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F89210>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x398 at 0x25914F89290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=352x500 at 0x25914F88450>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F887D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F88910>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=288x352 at 0x25914F76A90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25913A30390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F88390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25913A30390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x373 at 0x25914F89010>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F894D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89A10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x332 at 0x25914A846D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=324x500 at 0x25914F88510>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F88050>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=363x500 at 0x25914F88B90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x480 at 0x25914F883D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=368x342 at 0x25914F88C50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x393 at 0x25914B466D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89310>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=334x500 at 0x25914F89390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F89710>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=258x350 at 0x25914F88350>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F890D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=332x500 at 0x25914F76FD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=918x714 at 0x25914F62B50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=618x525 at 0x25914F88910>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89A90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F89B10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=400x500 at 0x25914F80490>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x275 at 0x25914F89090>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x450 at 0x25914F89CD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89D90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F89E90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x400 at 0x25914F89F50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x332 at 0x25914F8A010>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x2590D7A7290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8A150>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=379x500 at 0x25914E11390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=567x377 at 0x25914F8A190>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=679x599 at 0x25914F89310>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8A6D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x358 at 0x25914F6EFD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=268x185 at 0x25914F89C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F89ED0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=443x500 at 0x25914F8AED0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F8A9D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=250x369 at 0x25914F6EFD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=640x480 at 0x25914F88590>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89490>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=320x348 at 0x25914F8A6D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x359 at 0x25914F89610>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=550x412 at 0x25914F8AC90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=570x428 at 0x25914F8A5D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x500 at 0x25914F888D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=550x412 at 0x25914F8AA10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8A5D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F8A110>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F880D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=217x162 at 0x25914F893D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F8A250>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F8B090>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x336 at 0x25914F888D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=640x420 at 0x25914F89050>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=404x500 at 0x259148858D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=187x160 at 0x25914F8ACD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x371 at 0x25914F8AC10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89550>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8AB10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F8B210>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F890D0>

[illegible]

<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x334 at 0x25914F82510>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8A410>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F6C090>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F8BAD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x360 at 0x25914F83910>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F8A5D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F83910>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F8BC90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914A020D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x334 at 0x25914F8BA90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F6E890>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=373x500 at 0x25914F89010>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F81B50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8BD50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=400x481 at 0x2590D7A7290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x334 at 0x25914F82410>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914A01F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x259149E3ED0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=375x500 at 0x25914F8BF90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8AC10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x333 at 0x25914F88ED0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x334 at 0x25914F74050>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=375x500 at 0x25914F8B910>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=452x500 at 0x25914F8A850>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=534x372 at 0x25914F8AC10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x259145D3BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F89FD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x363 at 0x25914F89C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=375x500 at 0x25914F83BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8BC90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=402x500 at 0x25914F83F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89DD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x329 at 0x25914F83ED0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F8BE50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x332 at 0x25914F83F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=340x313 at 0x25914F6C090>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=349x350 at 0x25914F81F10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=285x430 at 0x25914F89890>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x280 at 0x25914F81F10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=279x500 at 0x25914F758D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=390x255 at 0x25914F88DD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=366x500 at 0x25914662F10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x328 at 0x25914F89DD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=332x500 at 0x25914F81C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=330x500 at 0x25914F89DD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=200x280 at 0x25914F88DD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x383 at 0x25914F8AF10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=375x479 at 0x25914F81C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=358x480 at 0x25914F8BC50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=369x500 at 0x25914F826D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=376x479 at 0x25914F82C10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=250x250 at 0x25914F88BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=351x500 at 0x25914F757D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=400x281 at 0x25914F82B90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F757D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=241x160 at 0x25914886890>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=600x603 at 0x25914F62B90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=250x207 at 0x25914886890>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=499x470 at 0x25914F757D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x400 at 0x25914B5D8D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914D76110>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=177x246 at 0x25914F8BED0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=400x300 at 0x25914B5D8D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=250x322 at 0x25914F82E10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=360x286 at 0x25914F6E3D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=394x396 at 0x25914F8AE10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=800x600 at 0x25914F763D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=450x392 at 0x25914F98C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F885D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=340x287 at 0x25914E108D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=230x172 at 0x25914F826D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=379x379 at 0x25914F763D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x358 at 0x25914F82C10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=298x451 at 0x25914F98510>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=639x600 at 0x25914F8A990>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=324x253 at 0x25914F6E690>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x278 at 0x25914F8B110>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=640x427 at 0x25914F98A90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914EFE690>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=800x877 at 0x25914F99150>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=420x500 at 0x25914F8B810>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=200x133 at 0x25914F6E690>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=179x200 at 0x25914F98490>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=370x421 at 0x25914F98210>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=375x500 at 0x25914F89390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=331x500 at 0x25914F77410>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x380 at 0x25914F89DD0>

<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=358x500 at 0x25914F89390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=299x217 at 0x25914F82E50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=200x303 at 0x25914F812D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=338x500 at 0x25914F77290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=216x172 at 0x25914F6E650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=800x600 at 0x25910BE1290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=343x500 at 0x25914F82E50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x378 at 0x25914F893D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=211x311 at 0x25914F629D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=229x200 at 0x25914F89DD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x363 at 0x25914F80FD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=301x270 at 0x25914E11290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=403x500 at 0x25914F8BCD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F893D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=447x500 at 0x25914F8AE50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x334 at 0x25914F89290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=499x470 at 0x25914F98710>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=200x150 at 0x25914F89290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=560x806 at 0x25914F8AE50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=365x500 at 0x25914F89BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=340x256 at 0x25914F812D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x391 at 0x259145D3BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F89C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=225x300 at 0x25914F6E3D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x316 at 0x259117AB650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=340x352 at 0x25914A59C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F98BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x290 at 0x25914F766D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=225x209 at 0x25914A59C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x376 at 0x25914F8B650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=221x300 at 0x25914F98290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=316x428 at 0x25914F89C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x259117AB650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=481x481 at 0x25914A59C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=160x170 at 0x25914F766D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=215x241 at 0x25914A59C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F62990>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=250x332 at 0x25914F99350>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x332 at 0x25914F98850>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=253x263 at 0x25914F98810>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=205x310 at 0x25914F826D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F98510>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=180x135 at 0x25914F993D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=245x237 at 0x25914F98ED0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=332x500 at 0x25914A73CD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=424x426 at 0x25914F98E10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F81310>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=326x350 at 0x25914ECA990>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=160x170 at 0x25914F991D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=229x315 at 0x25914F98F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=252x255 at 0x25914F98DD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x281 at 0x25914F99210>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=349x350 at 0x25914F980D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=800x782 at 0x25914F63F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=600x449 at 0x25914F98310>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x263 at 0x25914F988D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=480x453 at 0x25914F99650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F98CD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=677x421 at 0x25914F98B50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=450x375 at 0x25914F8B150>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x262 at 0x25914EFC090>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x311 at 0x25914F8B150>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x252 at 0x25914F997D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x299 at 0x25911CC3150>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=420x500 at 0x259145D3BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x316 at 0x25914F994D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=320x236 at 0x25911946650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=600x400 at 0x25914F6C090>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x263 at 0x25911946650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x282 at 0x25914F81710>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x316 at 0x25914F83FD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=393x224 at 0x25914F89290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=331x500 at 0x25914F83BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=340x360 at 0x25914E11450>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=290x360 at 0x25914F986D0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=350x263 at 0x25914F83BD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x421 at 0x25914F98CD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=450x305 at 0x25914F98C50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F99490>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=340x350 at 0x25914F99810>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=344x500 at 0x25914F98310>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=300x316 at 0x25914F98F90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=333x500 at 0x25914F8BA90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=334x500 at 0x25914F80390>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=241x336 at 0x25914F80450>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=450x351 at 0x25914BDEDD0>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=391x500 at 0x25914F81B90>


```
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=250x200 at 0x25914F80450>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=356x500 at 0x25914F81A90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x481 at 0x25914F99890>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=330x500 at 0x25914F98290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=358x500 at 0x25914F98D10>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=288x286 at 0x25914F99890>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914F98C50>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=500x375 at 0x25914B46650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=640x634 at 0x25914F88290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=160x200 at 0x25914B46650>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=200x210 at 0x25914F99C90>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=288x216 at 0x25914E11290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=360x301 at 0x2590D7A7290>
<PIL.JpegImagePlugin.JpegImageFile image mode=RGB size=600x603 at 0x25914F99790>
```

2B. Feature Extraction: Edge Histogram and Similarity Measurements

This section describes how to extract features from cropped dog images and measure their similarity.

- 1. **Select Images:** Choose one image from each class.
- 2. **Convert to Grayscale:** Change the color images to grayscale.
- 3. **Calculate Edge Angles:** Use Sobel filters to find the edge direction for each pixel:
- 4. **Compare Histograms:** Measure similarity between the first two images using Euclidean, Manhattan, and Cosine distances.

```
In [4]: def process_images_and_histograms(cropped_images_dir):
# Step 1: Get all directories inside the cropped images directory
directory_paths = glob.glob(os.path.join(cropped_images_dir, '*'))

# Step 2: Choose one image from each directory
selected_images = []
class_names = []

for directory in directory_paths:
# Get the first image in each directory
image_paths = glob.glob(os.path.join(directory, '*'))
if image_paths:
selected_images.append(image_paths[0])
class_name = os.path.basename(directory)
class_names.append(class_name)

# Function to process images
def process_image(image_path):
# Load the image
image = io.imread(image_path)

# Convert to grayscale
gray_image = color.rgb2gray(image)

# Calculate angles using Sobel filters
angle_sobel = np.mod(np.arctan2(filters.sobel_v(gray_image), filters.sobel_h(gray_image)), np.pi)

# Obtain histogram with 36 bins
hist, bin_edges = np.histogram(gray_image, bins=36, range=(0, 1))

return image, gray_image, hist, bin_edges, angle_sobel

# Step 3: Process selected images
results = [process_image(img_path) for img_path in selected_images]
original_images, gray_images, histograms, bin_edges_list, angles = zip(*results)

# Step 4: Plot original images, grayscale images, and their corresponding histograms
fig, axs = plt.subplots(len(gray_images), 3, figsize=(15, 8))
for i in range(len(gray_images)):
# Plot original image
axs[i, 0].imshow(original_images[i])
axs[i, 0].axis('off')
axs[i, 0].set_title(f'Class Image {i + 1}')

# Plot grayscale image
axs[i, 1].imshow(gray_images[i], cmap='gray')
axs[i, 1].axis('off')
axs[i, 1].set_title(f'Grayscale Image {i + 1}')

# Plot histogram
axs[i, 2].bar(bin_edges_list[i][:-1], histograms[i], width=np.diff(bin_edges_list[i]), edgecolor='black')
axs[i, 2].set_title(f'Histogram for Image {i + 1}')
axs[i, 2].set_xlabel("Bins")
axs[i, 2].set_ylabel("Pixel Count")

plt.tight_layout()
plt.show()
```

```
# Step 5: Histogram comparison between the first two images
if len(histograms) >= 2:
    hist1 = histograms[0]
    hist2 = histograms[1]

    # Normalize histograms
    hist1_normalized = hist1 / np.sum(hist1)
    hist2_normalized = hist2 / np.sum(hist2)

    # Calculate distances
    euclidean_distance = np.linalg.norm(hist1_normalized - hist2_normalized)
    manhattan_distance = np.sum(np.abs(hist1_normalized - hist2_normalized))
    cosine_distance = pairwise.cosine_distances([hist1_normalized], [hist2_normalized])[0][0]

    print(f'Euclidean Distance: {euclidean_distance}')
    print(f'Manhattan Distance: {manhattan_distance}')
    print(f'Cosine Distance: {cosine_distance}')
else:
    print("Not enough images to compare histograms.")

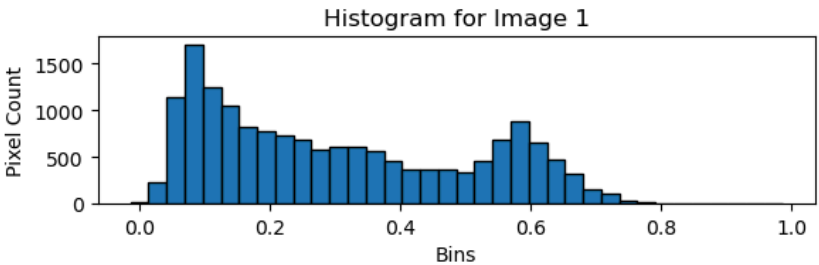
return histograms, selected_images

cropped_images_dir = '/Users/thota/Documents/Datamining/Datasets/Cropped'
histograms, selected_images = process_images_and_histograms(cropped_images_dir)
```

Class Image 1



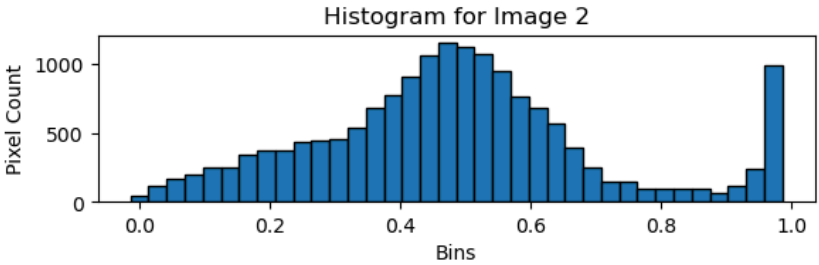
Grayscale Image 1



Class Image 2



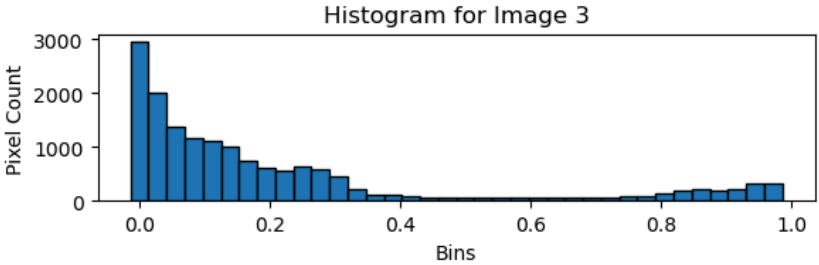
Grayscale Image 2



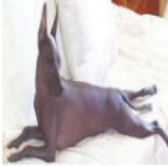
Class Image 3



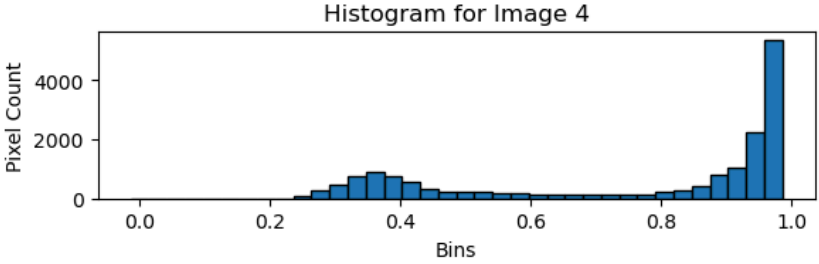
Grayscale Image 3



Class Image 4



Grayscale Image 4



Euclidean Distance: 0.18533730969634396
Manhattan Distance: 0.777587890625
Cosine Distance: 0.3626026514519938

2C. Histogram of Oriented Gradient (HOG) Feature Descriptor

In this section, we computed and visualized the HOG descriptors for a selected image.

Steps:

1. **Select an Image:** Choose one image from your dataset.
2. **Compute HOG Descriptors:**
 - Use a function to load the image and convert it to grayscale.
 - Compute the HOG features and generate the HOG image.
 - Visualize both the original grayscale image and the HOG descriptor side by side.
3. **Visualization:** The original image will be displayed alongside its HOG descriptor, allowing for a visual comparison of the features captured by the HOG method.

For more information about HOG, refer to the [Wikipedia page on HOG](#) and the [scikit-image documentation](#).

```
In [6]: def compute_and_visualize_hog(image_path):
        # Load the image
        image = io.imread(image_path)
```

```
# Convert to grayscale
gray_image = color.rgb2gray(image)

# Compute HOG features and HOG image
hog_features, hog_image = feature.hog(gray_image, visualize=True, block_norm='L2-Hys')

# Plot original image and HOG image
fig, axs = plt.subplots(1, 2, figsize=(12, 6))

axs[0].imshow(gray_image, cmap='gray')
axs[0].set_title('Class Image')
axs[0].axis('off')

axs[1].imshow(hog_image, cmap='gray')
axs[1].set_title('HOG Descriptor')
axs[1].axis('off')

plt.show()

for image_path in selected_images:
    compute_and_visualize_hog(image_path)
```

Class Image



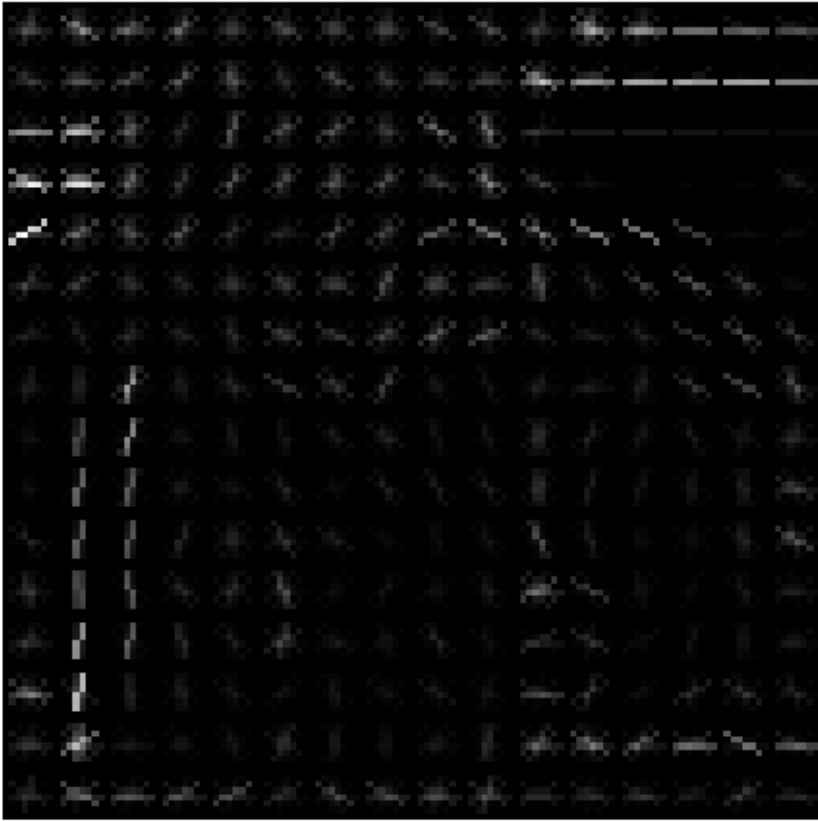
HOG Descriptor

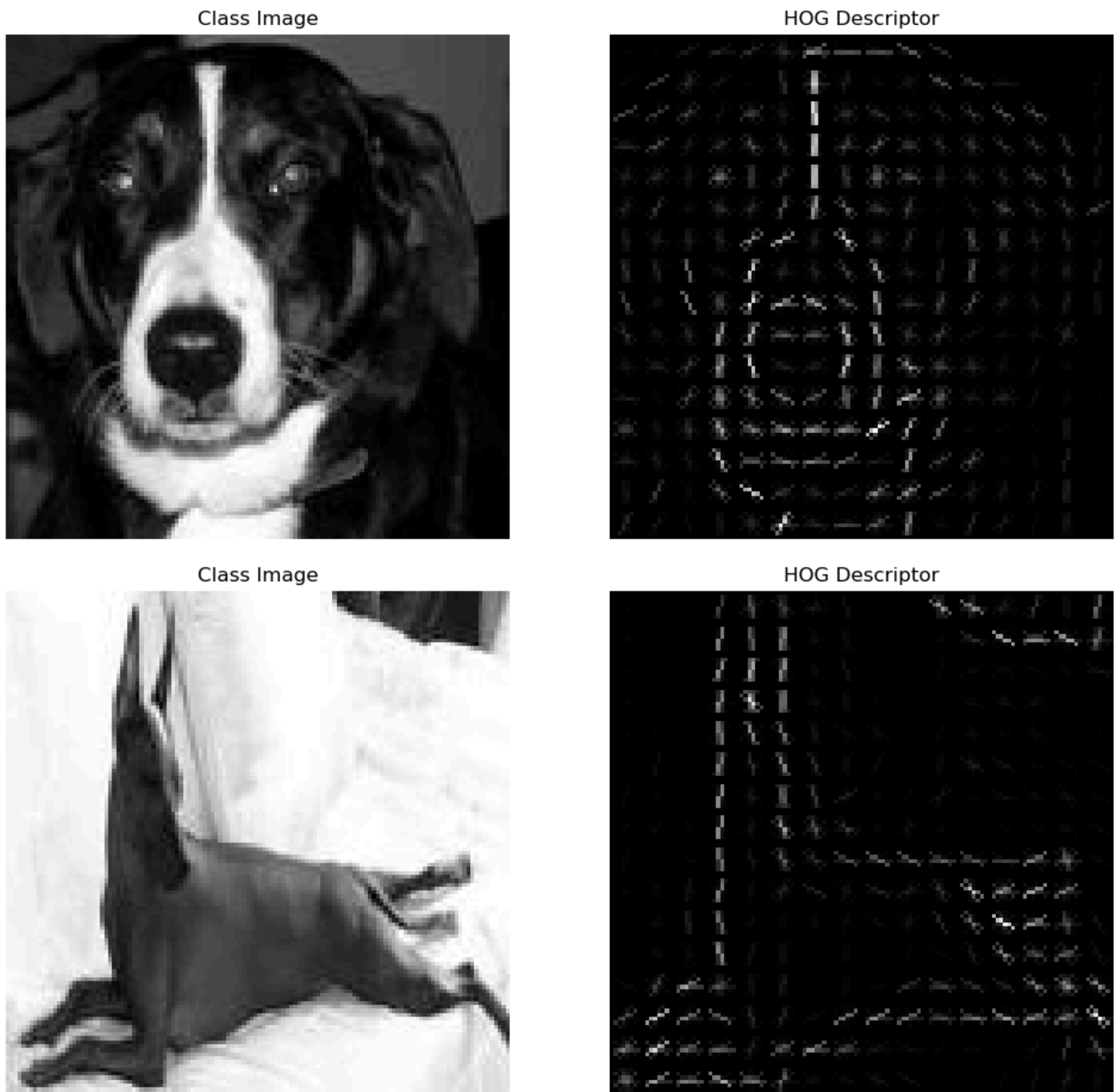


Class Image



HOG Descriptor





2D. Dimensionality Reduction using Principal Component Analysis (PCA)

In this section, we performed dimensionality reduction on edge histograms obtained from images of four different classes using Principal Component Analysis (PCA).

Steps:

1. **Select Images:** Use images from all four classes in your dataset.
2. **Convert to Edge Histograms:**
 - Convert all images from the four classes into edge histograms, which capture the distribution of edges in the images.
3. **Perform PCA:**
 - Apply PCA to the set of histograms to reduce their dimensionality from 36 dimensions to 2 dimensions.
 - Note that during this step, class labels will not be used.
4. **Plot Results:**
 - Visualize the reduced 2D points on a scatter plot.
 - Use different colors to represent the different classes, allowing for a visual assessment of class separability in the reduced feature space.

For more details on PCA, you can refer to the [scikit-learn PCA documentation](#) and a code example using the Iris dataset can be found [here](#).

```
In [7]: def process_images_and_histograms(cropped_images_dir):  
        """Process images in the specified directory and compute histograms."""
```

```

# Step 1: Get all directories inside the cropped images directory
directory_paths = glob.glob(os.path.join(cropped_images_dir, '*'))

histograms = []
class_names = []

# Step 2: Iterate through directories and process images
for directory in directory_paths:
    class_name = os.path.basename(directory)
    image_paths = glob.glob(os.path.join(directory, '*'))

    for image_path in image_paths:
        # Load the image
        image = io.imread(image_path)

        # Convert to grayscale
        gray_image = color.rgb2gray(image)

        # Obtain histogram with 36 bins
        hist, _ = np.histogram(gray_image, bins=36, range=(0, 1))

        histograms.append(hist) # Store the histogram
        class_names.append(class_name)

    return np.array(histograms), class_names

def perform_pca(histograms, n_components=2):
    """Perform PCA on histograms and return the principal components."""
    # Standardize the histograms
    scaler = StandardScaler()
    histograms_scaled = scaler.fit_transform(histograms)

    # Perform PCA
    pca = PCA(n_components=n_components)
    principal_components = pca.fit_transform(histograms_scaled)

    return principal_components

def plot_pca(principal_components, class_names):
    """Plot PCA results with labels."""
    unique_classes = list(set(class_names))
    colors = plt.cm.get_cmap('tab10', len(unique_classes))

    plt.figure(figsize=(10, 6))
    for i, class_name in enumerate(unique_classes):
        class_indices = [j for j, name in enumerate(class_names) if name == class_name]
        plt.scatter(principal_components[class_indices, 0], principal_components[class_indices, 1],
                    label=class_name, color=colors(i), alpha=0.6)

    plt.title('PCA of Histograms')
    plt.xlabel('Principal Component 1')
    plt.ylabel('Principal Component 2')
    plt.legend()
    plt.grid()
    plt.show()

# Main Execution
# Specify the directory containing cropped images
cropped_images_dir = '/Users/thota/Documents/Datamining/Datasets/Cropped'

# Step 1: Process images and compute histograms
histograms, class_names = process_images_and_histograms(cropped_images_dir)

# Step 2: Perform PCA on the histograms
principal_components = perform_pca(histograms)

# Step 3: Plot the PCA results
plot_pca(principal_components, class_names)

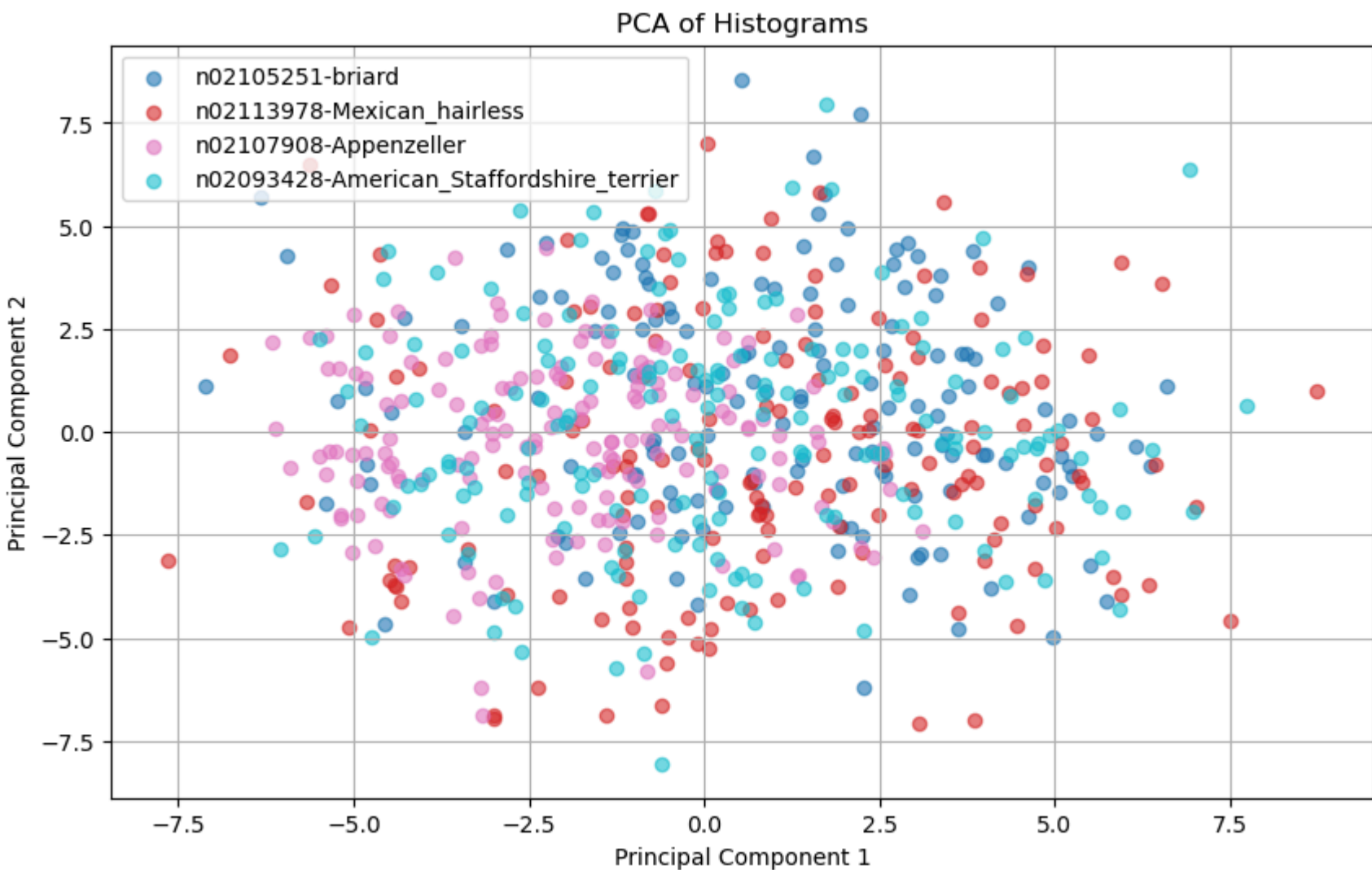
```

C:\Users\thota\AppData\Local\Temp\ipykernel_5220\3092872568.py:44: MatplotlibDeprecationWarning: The get_cmap function was deprecated in Matplotlib 3.7 and will be removed two minor releases later. Use ``matplotlib.colormaps[name]`` or ``matplotlib.colormaps.get_cmap(obj)`` instead.

```

    colors = plt.cm.get_cmap('tab10', len(unique_classes))

```

4 Feature Extraction from Tweets

We will use `CountVectorizer` and `TfidfVectorizer` from Scikit-learn to analyze the tweet dataset.

What We Will Do

- 1. **CountVectorizer**: Count how many times each word appears in the tweets.
- 2. **TfidfVectorizer**: Calculate the importance of each word in the tweets.

Results

We will check the size of the two feature sets we created.

```
In [11]: import json
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

def load_and_extract(file_path):
    """Load tweet data and extract texts."""
    with open(file_path, 'r') as file:
        return [json.loads(line.strip())['Tweet'] for line in file]

def compute_vectorizations(texts):
    """Compute Count and TF-IDF Vectorizations."""
    count_vectorizer = CountVectorizer().fit_transform(texts)
    tfidf_vectorizer = TfidfVectorizer().fit_transform(texts)
    return count_vectorizer, tfidf_vectorizer

def print_dimensionalities(X_counts, X_tfidf):
    """Print dimensionalities of vectorized representations."""
    print(f'Dimensionality of Count Vectorizer: {X_counts.shape[1]}')
    print(f'Dimensionality of TF-IDF Vectorizer: {X_tfidf.shape[1]}')

file_path = r'C:\Users\thota\Documents\Datamining\Datasets\student_29\train.json'
texts = load_and_extract(file_path)
X_counts, X_tfidf = compute_vectorizations(texts)
print_dimensionalities(X_counts, X_tfidf)
```

Dimensionality of Count Vectorizer: 9613
Dimensionality of TF-IDF Vectorizer: 9613

5 Dimensionality Reduction and Visualization of Tweet Classes

Selected Classes

For this analysis, we will pick the following four classes that are expected to be separable:

- 1. Anger
- 2. Joy

3. Fear
4. Trust

Dimensionality Reduction

We will perform dimensionality reduction on both the token count features and TF-IDF features, reducing them to 2 dimensions.

Visualization

We will create two separate plots:

1. **Token Count Features:** Plot the 2D points using four different colors for the selected classes.
2. **TF-IDF Features:** Plot the 2D points using four different colors for the selected classes.

Observations

After plotting, we will analyze how many classes are visually separable (i.e., non-overlapping) in both plots.

```
In [10]: import json
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer

def extract_labels(tweets, selected_classes):
    """Extract binary labels for selected classes."""
    return np.array([[int(tweet[class_name]) for class_name in selected_classes] for tweet in tweets if 'Tweet' in tweet])

def perform_pca_and_plot(X, labels, classes, feature_type='Count'):
    """Perform PCA and plot the results."""
    X_reduced = PCA(n_components=2).fit_transform(X)
    plt.figure(figsize=(10, 6))
    colors = ['red', 'blue', 'green', 'orange']

    for i, class_name in enumerate(classes):
        plt.scatter(X_reduced[labels[:, i] == 1, 0], X_reduced[labels[:, i] == 1, 1],
                    label=class_name, color=colors[i], alpha=0.6)

    plt.title(f'2D PCA Plot of Tweets - {feature_type} Vectorizer')
    plt.xlabel('Principal Component 1')
    plt.ylabel('Principal Component 2')
    plt.legend()
    plt.grid()
    plt.tight_layout()
    plt.show()

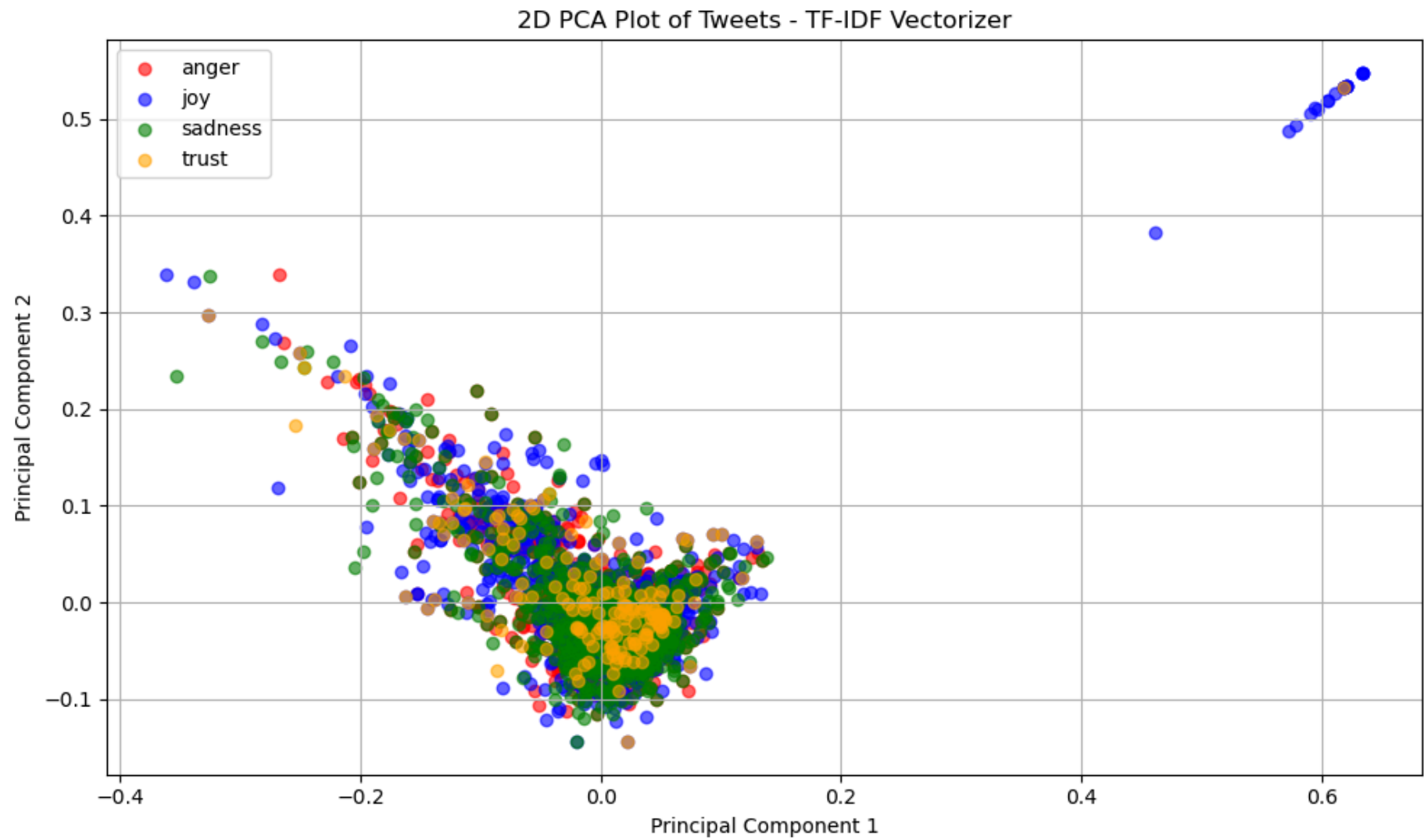
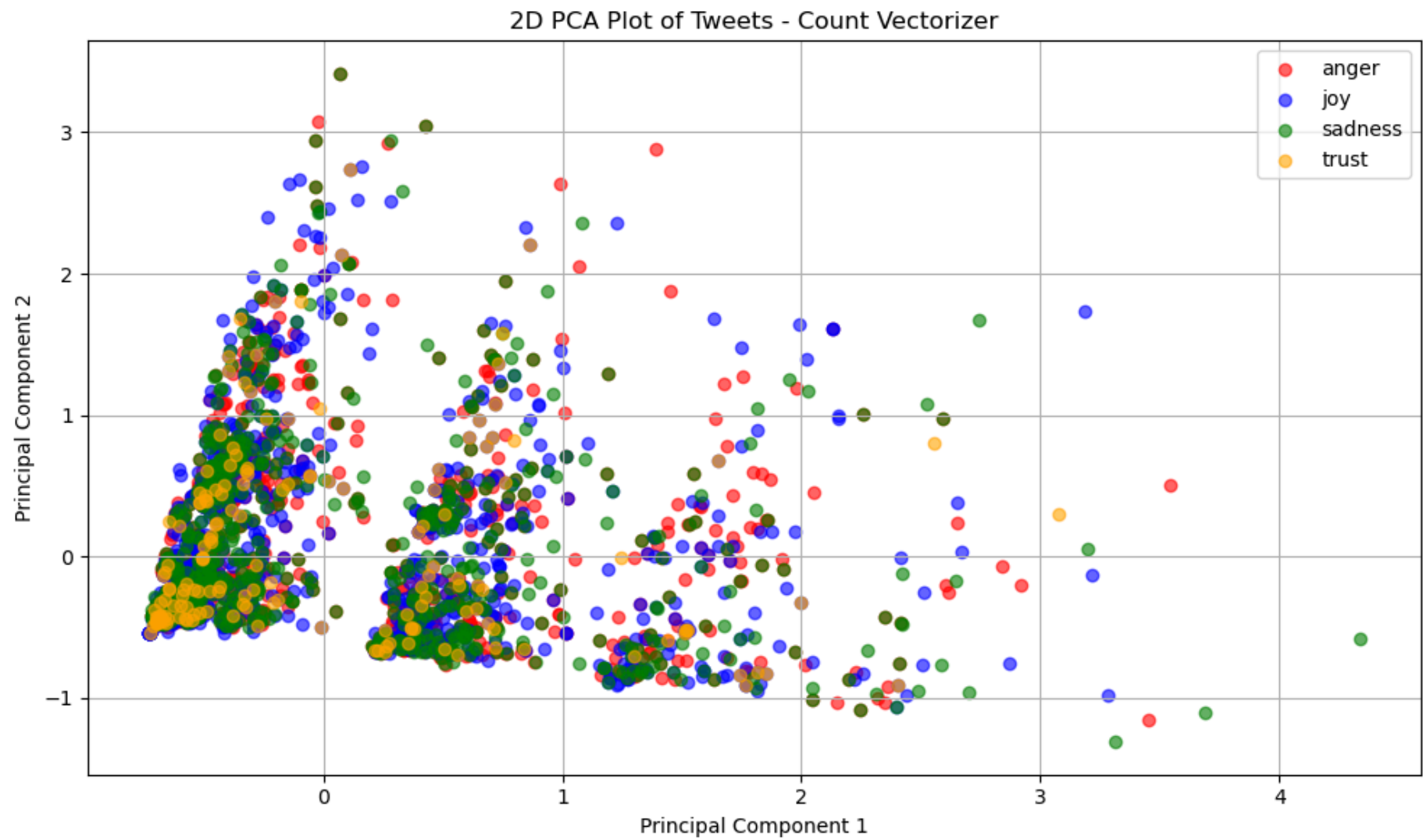
# Load tweet texts
file_path = r'C:\Users\thota\Documents\Datamining\student_29\train.json'
texts = load_and_extract(file_path)

# Define classes to analyze
selected_classes = ['anger', 'joy', 'sadness', 'trust']

# Extract the labels from the original data
with open(file_path, 'r') as file:
    tweets = [json.loads(line.strip()) for line in file]
labels = extract_labels(tweets, selected_classes)

# Vectorization
X_counts = CountVectorizer().fit_transform(texts).toarray()
X_tfidf = TfidfVectorizer().fit_transform(texts).toarray()

# Perform PCA and plot for both vectorizers
for X, feature_type in zip([X_counts, X_tfidf], ['Count', 'TF-IDF']):
    perform_pca_and_plot(X, labels, selected_classes, feature_type)
```



The classes(anger, joy, sadness, and trust) are not separable in either the Count Vectorizer or TF-IDF Vectorizer plots, as there is significant overlap between the points.

```
In [ ]:
```