Visual Analytics Portfolio

Assignment 1: Building a simple image search algorithm

Cultural Data Science, 2023

Author: Aleksander Moeslund Wael
Student no. 202005192

Assignment notes (Ross)

For this assignment, you'll be using `OpenCV` to design a simple image search algorithm. For this exercise, you should write some code which does the following:

- Define a particular image that you want to work with
- For that image
    - Extract the colour histogram using `OpenCV`
- Extract colour histograms for all of the **other* images in the data
- Compare the histogram of our chosen image to all of the other histograms
    - For this, use the `cv2.compareHist()` function with the `cv2.HISTCMP_CHISQR` metric
- Find the five images which are most simlar to the target image
    - Save a CSV file to the folder called `out`, showing the five most similar images and the distance metric:

Filename	**Distance]**
target	0.0
filename1	---
filename2	---

Introduction

This repo contains a Python script, `top_five_similar.py`, which can be used to calculate the colour histogram of an image and compare it to all other images in a folder. It returns a csv file with the top 5 most similar images in the folder (when comparing colour histograms), and the distance metric for these images compared to the target image (as calculated by chi-squared).

Data

The dataset for this project contains 1360 images of flowers. The dataset is a collection of over 1000 images of flowers, sampled from 17 different species. The dataset comes from the Visual Geometry Group at the University of Oxford, and full details of the data can be found here. The data should be stored in a folder called `flowers` within the `data` folder.

Pipeline

The Python script is sectioned as follows:

1. Import the data
2. Calculate color histograms for all images in the target folder
3. Compare target image to all other images
4. Print and save top-5 most similar images dataframe to `out` folder

Requirements

The code is written for Python 3.11.2. Futhermore, if your OS is not unix-based, a bash-compatible terminal is required for running shell scripts (such as Git Bash for Windows).

How to run

NOTE: *Depending on your OS, run either `WIN_*` (on Windows) or `MACL_*` (on MacOS or Linux).*

1. Clone repository to desired directory

```
git clone https://github.com/AU-CDS/assignment1-simple-image-search-alekswael
cd assignment1-simple-image-search-alekswael
```

2. Run shell script

The `*run.sh` script does the following:

1. Creates a virtual environment for the project
2. Activates the virtual environment
3. Installs the correct versions of the packages required
4. Runs `top_five_similar.py` located in the `src` folder
5. Deactivates the virtual environment

```
bash WIN_run.sh
```

Specify target data and image

You can pass arguments through `argparse` to use other image datasets, or to compare other flower images from the project dataset. To do this, manually activate the virtual environment, and then run the Python script with arguments.

```
source ./top_five_similar_venv/Scripts/activate # WINDOWS
source ./top_five_similar_venv/bin/activate # MAC OR LINUX
```

```
top_five_similar.py [-h] [--folder FOLDER] [--image IMAGE]

options:
  -h, --help       show this help message and exit
  --folder FOLDER  Name of image-containing folder located in the data folder
  --image IMAGE    Name of the image you want to compare to the rest of the images
in the folder
```

When running the script, the default argument values are `--folder flowers --image image_0001.jpg`.

Example output from running script

```
Input image: image_0007.jpg

            Filename  Distance
6             target      0.00
1202  image_1203.jpg    133.13
772   image_0773.jpg    133.50
594   image_0595.jpg    136.10
725   image_0726.jpg    136.62
237   image_0238.jpg    136.78
```

Repository structure

In your working directory, you should have two folders: data and out. The data-folder should contain a subfolder with the images. The out-folder is the save location for the output .csv file.

```
│   .gitignore
│   README.md
│   requirements.txt
│   run.sh
│
├───.github
│       .keep
│
├───data
│   └───flowers
│           image_1360.jpg   # Make sure to download the data
├───out
│       .gitkeep
│
├───src
│       top_five_similar.py
│
├───utils
│       imutils.py
│       __init__.py
```