Jan 19th, 2025

Analysis on Data preprocessing and algorithm description

Main Directory Structure

```
└── CTNNB1 and arm/
    ├── CTNNB1/
    └── ARM/
```

The wings information is embedded within the CTNNB1 and arm folders.

Information domain and data characterization

The original size of an image taken from the dataset was quite variable, with some images much wider than tall, and a resolution over 3K pixels on either rows or columns.

The images were produced by what appears to be single shots stacked together to form bigger images. This is evident by the presence of "missing" tiles, noticeable as white areas in the original image.

Another consideration is the variable of orientation in the way images were taken, making it challenging for an algorithm to automate the cropping process of the data using cv2 or Pillow.

The dataset went through three algorithms to produce a final, standardized dataset:

From utils.py file, located in the path:

└── scripts └── Wings/ ├── utils.py

FD. raw_files_transfer() to produce folder 00_Metadataset

Resizes the images to a standard size of 512 for the biggest dimension (either rows or columns). For example, an image of dimensions:

```
2,000 w x 1,000 h  --> 512 w x 256 h
1,000 w x 2,000 h  --> 256 w x 512 h
```

This resizing protocol ensures the cropping happens proportionally to the direction that better captures the wing.

Manual editing Approx. 3 days to produce folder 01_cleaned_Dataset

Taking as a starting point the resized dataset, Adobe Photoshop CC2018 was used to: - Separate the pixels belonging to the wings from the background. - Scaling the wing to match a template image - Rotating the wing to match a template image - Resizing the image to remove excess empty information

FD. center_dataset() to produce folder 02_stnd_Dataset

Centers the cleaned dataset by finding the biggest w and h and centering the rest of the images to that new scale, while adding a black background to replace the transparency.

Changing the transparency was necessary because of the introduction of artifacts by Pillow.Image.

The program was then executed with a modification to the original version of main for the PCA decomposition of eyes, and modified to the name

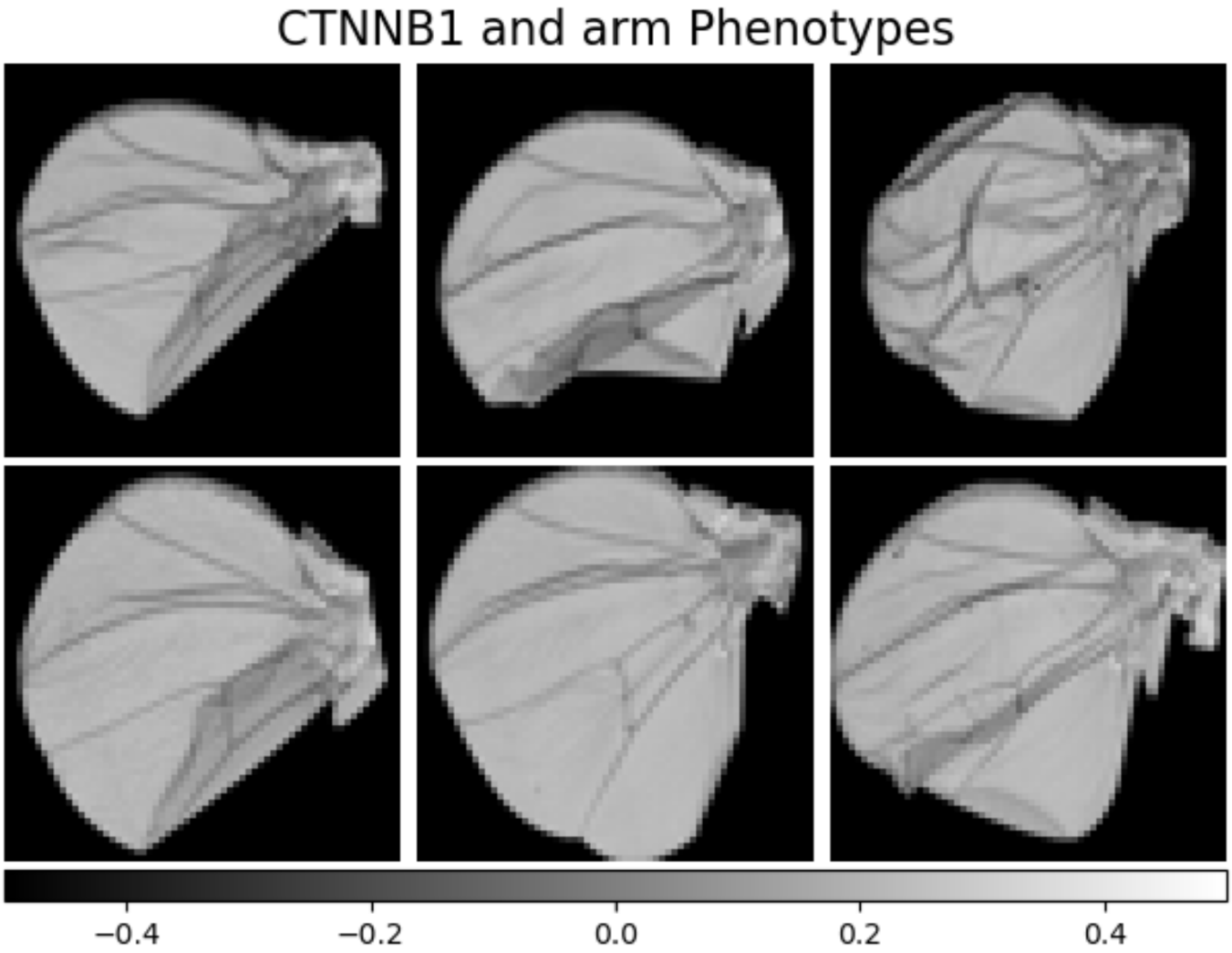
main_CTNNB1_arm.py

. This file contains adjustments in the path to find the images, and avoiding further resizing.

Preliminary results

```
The final dataset
We have standardized image dimensions of: W:512 x H:300
Total images in the dataset is: 133
```

Preview of the dataset, displayed as square images for visualization purposes.



Preview of the PCA analysis with NMF