was prepared by **Ivan Luchko**

Digital Camera Autodetect: Day or Night

Project consist of 4 modules:

RGBhistExtraction.py
colorHistExtraction.py
imageClassification.py
testYourImageDemo.py

First, two modules serve for data extraction from the images dataset stored in ***"/images"*** folder. Extracted data is saved in pickle files which are stored in ***"/jobs"*** folder. "***imageClassification.py"*** module further processes this data and useses it for training and testing different classification models.

"/models" folder contains the binary pickle files of pretrained classification models.

"/px_samples" folder contains pixels sampling files over image dataset (Npx pixels per image).

"testYourImageDemo.py" module provides classification demo on the images dataset store in ***"/images/test"*** folder .

Detailed description of modules and functions can be found in doc-comments in corresponding files.

In order to classify Day/Night images three models of different complexity were tested:

Model	Accuracy score achieved
'gray' (black/white) model	0.927 +- 0.023
RGB histogram model	0.940 +- 0.022
Colors histogram model (70 colors)	0.962 +- 0.015

In order to train and test different models 723 labeled images dataset was used stored in **"/images"**.

1. 'gray' (black/white) model

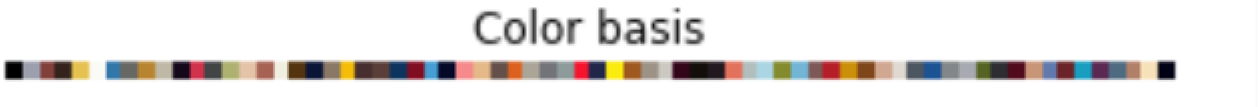
The most simple one. First, for each image, mean of the pixels intensity over image and three channels is calculated. Then this one number feature is used for images classification.

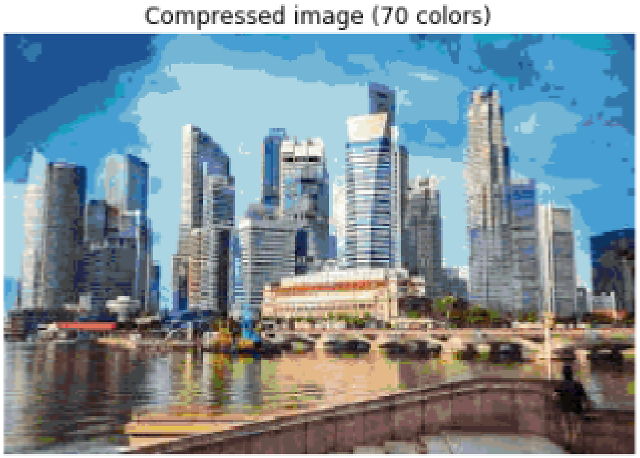
2. RGB histogram model

Similarly to the previous model for each image, mean of the pixels intensity of the image along three channel is calculated. Thus, RGB color histogram is obtained and it is further used for images classification using SVC (feature vector contains 3 elements).

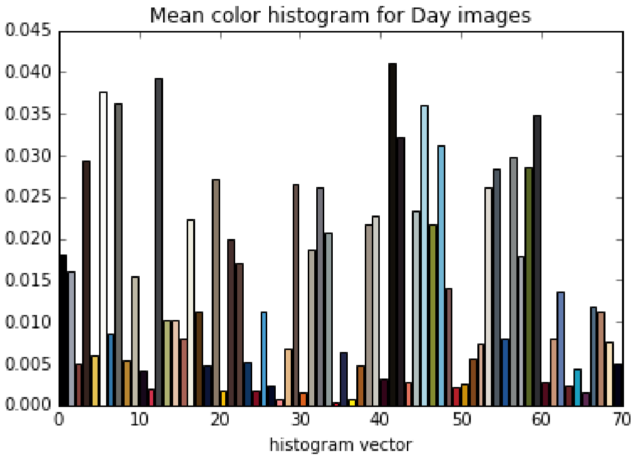
3. Colors histogram model (70 colors – optimal for given images dataset size)

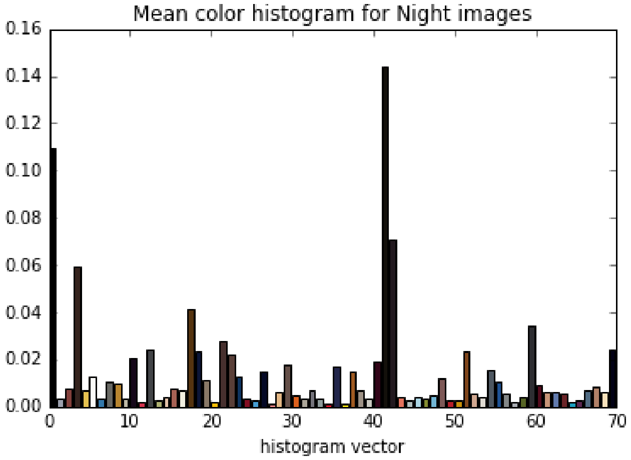
First, k-means classifier is applied to image dataset for building the color basis (via color quantization) of the given ***n_colors.***



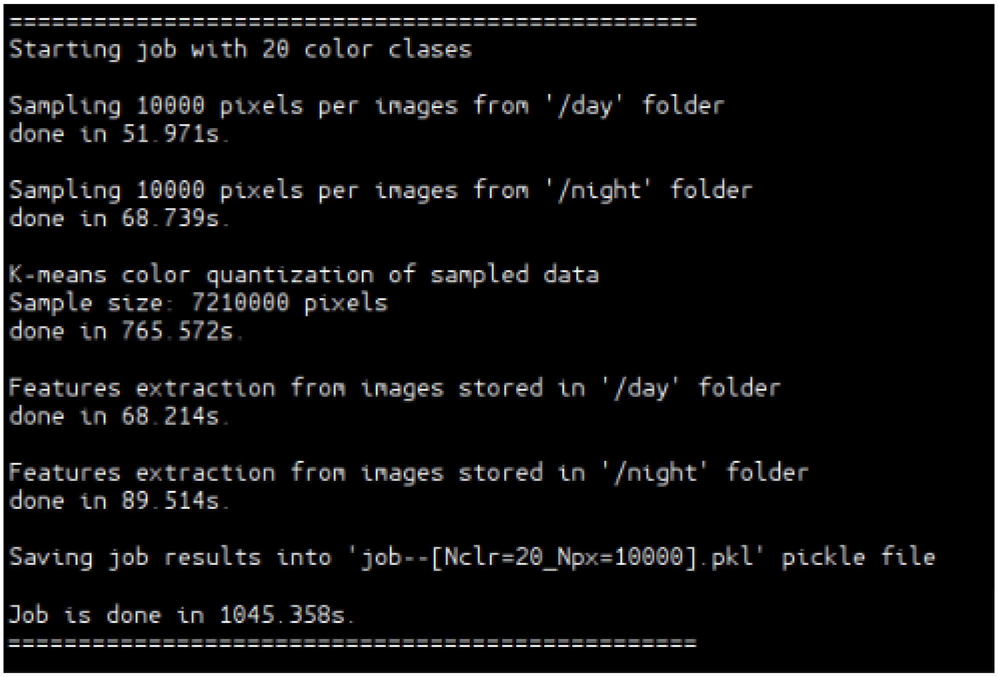


Then, density color histograms of each image are calculated according to this basis. Further, this histograms are used for feature matrix calculation and image classification. Mean Day and Night images color histograms are demonstrated below



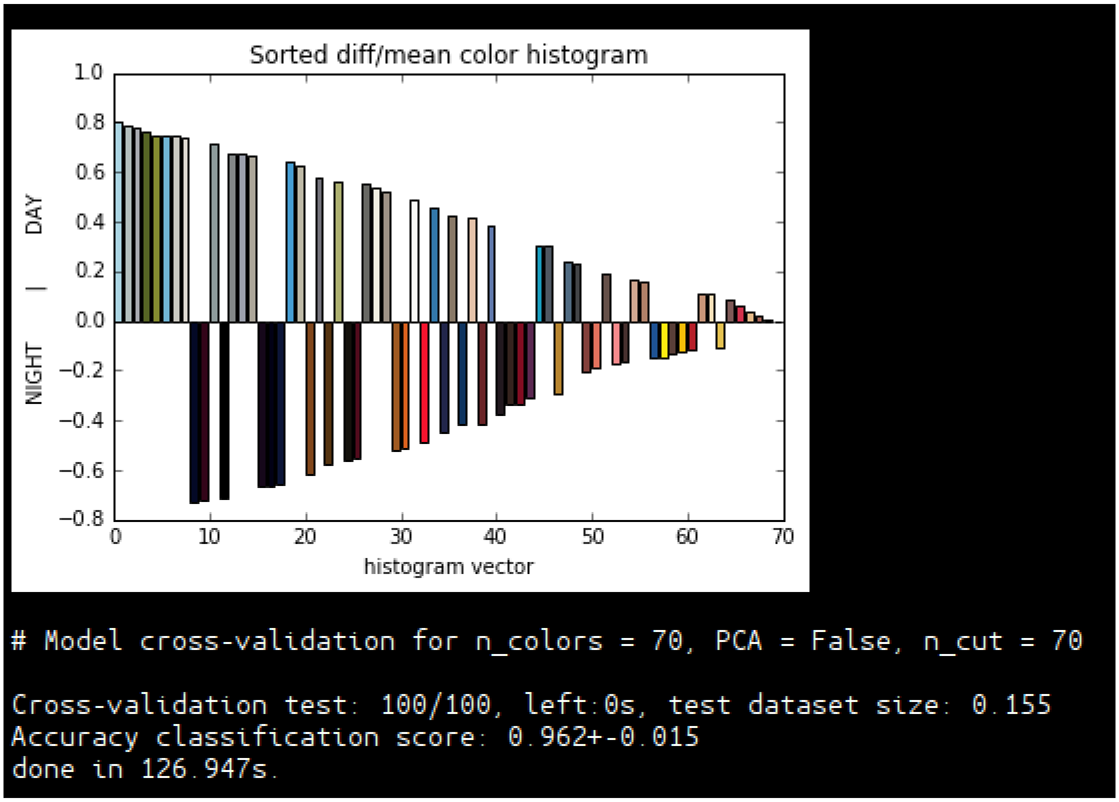


Some steps are represented in console during the calculation process.



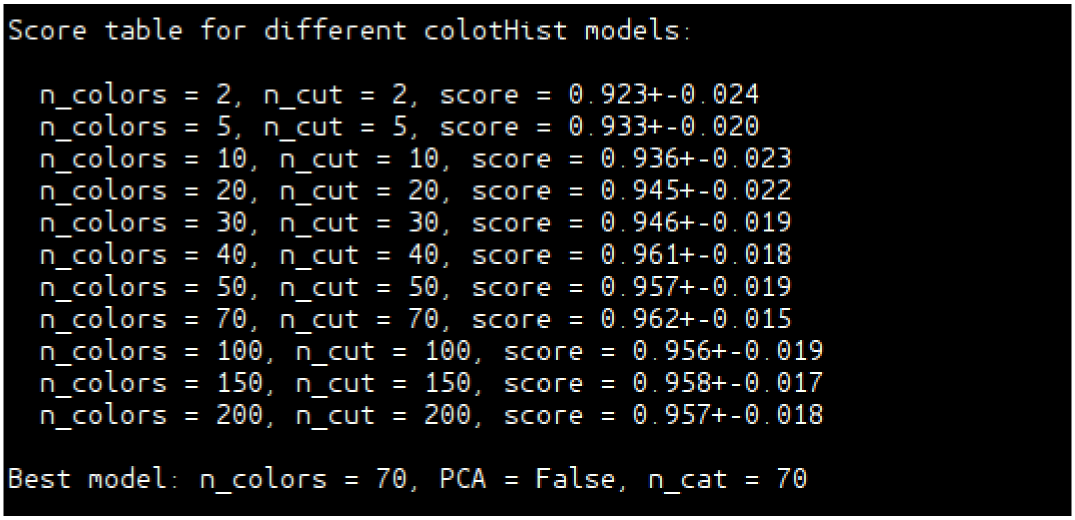
```
================================================
Starting job with 20 color clases

Sampling 10000 pixels per images from '/day' folder
done in 51.971s.

Sampling 10000 pixels per images from '/night' folder
done in 68.739s.

K-means color quantization of sampled data
Sample size: 7210000 pixels
done in 765.572s.

Features extraction from images stored in '/day' folder
done in 68.214s.

Features extraction from images stored in '/night' folder
done in 89.514s.

Saving job results into 'job--[Nclr=20_Npx=10000].pkl' pickle file

Job is done in 1045.358s.
================================================
```

Quasi-PCA can be also applied, neglecting the colors which appear equally in both classes, and thus, are worse discriminators. It is demonstrated on difference/mean color histogram below. As it might be expected light colors characterize Day images and dark colors Night images respectively.



```
# Model cross-validation for n_colors = 70, PCA = False, n_cut = 70

Cross-validation test: 100/100, left:0s, test dataset size: 0.155
Accuracy classification score: 0.962+-0.015
done in 126.947s.
```

The larger training image dataset the better accuracy can be achieved exploiting larger color basis (more complex model). For the given dataset different color basis size (***n_colors)*** were tested exploring high bias-variance trade-off. In order to pick the optimal model, mean scores during 100 cross-validation tests with the test size 0.15 were calculated.



```
Score table for different colotHist models:

  n_colors = 2, n_cut = 2, score = 0.923+-0.024
  n_colors = 5, n_cut = 5, score = 0.933+-0.020
  n_colors = 10, n_cut = 10, score = 0.936+-0.023
  n_colors = 20, n_cut = 20, score = 0.945+-0.022
  n_colors = 30, n_cut = 30, score = 0.946+-0.019
  n_colors = 40, n_cut = 40, score = 0.961+-0.018
  n_colors = 50, n_cut = 50, score = 0.957+-0.019
  n_colors = 70, n_cut = 70, score = 0.962+-0.015
  n_colors = 100, n_cut = 100, score = 0.956+-0.019
  n_colors = 150, n_cut = 150, score = 0.958+-0.017
  n_colors = 200, n_cut = 200, score = 0.957+-0.018

Best model: n_colors = 70, PCA = False, n_cat = 70
```

Having larger image dataset would allow to use larger color basis and achieve even better performance. Color histogram model approach can be used for multi class classification (sunset, park area, city etc.) providing much more robust classification comparing to ‘gray’ and RGB histogram models.

Classification test results

Day image

Night image

Day image

Night image

Day image

Night image

Day image

Night image

Day image

Night image

Day image

Night image