Homework 1

Statistics W4240: Data Mining (Section 001) Columbia University

Due Friday, February 5

Zhang Yunyan UNI: yz2861

Problem 1

a.

The class of face_01 is "pixmapGrey"

The size of the original image in pixels is 192x168.

b.

The maximum value that a pixel can take for this type of file is 1.

The maximum value in the faces_matrix_01 is 0.007843137, so the minimum value a pixel can take for this type of file is supposed to be 0.

Because the value for grey(1) is “#FFFFFF”, 1 is white, while grey(0)’s value is ”#000000”, 0 is black.

c.

The length of dir_list_1 is 38.

The length of dir_list_2 is 2547.

The elements contained in dir_list_1 are names of sub-folders under file “CropYale”, such as "yaleB01" "yaleB02" "yaleB03" "yaleB04" "yaleB05" "yaleB06".

The elements contained in dir_list_2 are names of files under file “CropYale”, such as

[1] "yaleB01/DEADJOE"
[2] "yaleB01/WS_FTP.LOG"
[3] "yaleB01/yaleB01_P00_Ambient.pgm"
[4] "yaleB01/yaleB01_P00.info"
[5] "yaleB01/yaleB01_P00A-005E-10.pgm"
[6] "yaleB01/yaleB01_P00A-005E+10.pgm"

d.

hw01_01d: 3x3 grid of faces

Problem 2

a.

The size of the matrix is 152 x 32256.

b.

The length of this list is 2452. The 10^{th} element is "yaleB01/yaleB01_P00A-025E+00.pgm".

c.

The names of the black images are:

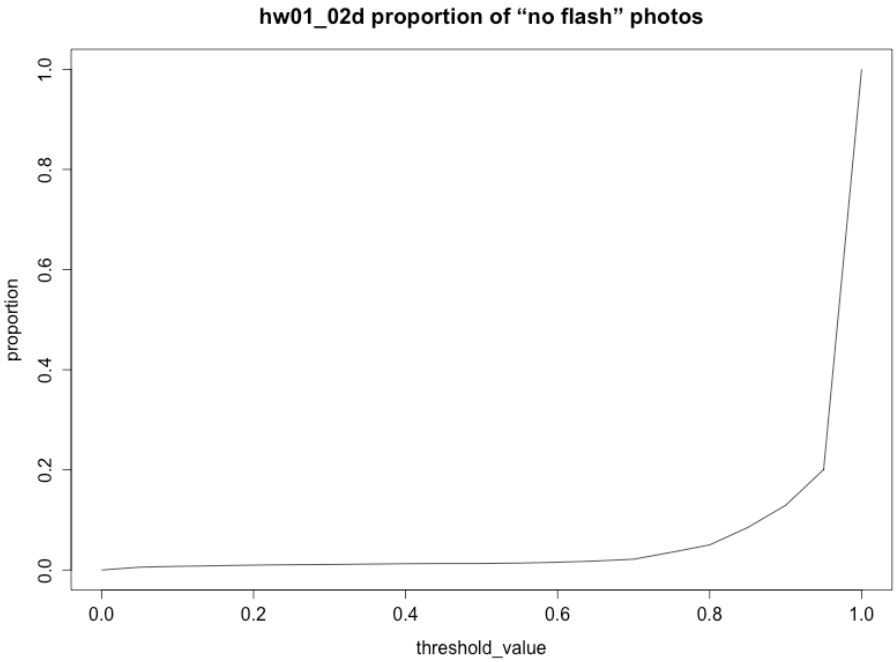
[1] "yaleB01/yaleB01_P00_Ambient.pgm"
[2] "yaleB02/yaleB02_P00_Ambient.pgm"
[3] "yaleB02/yaleB02_P00A+095E+00.pgm"
[4] "yaleB04/yaleB04_P00_Ambient.pgm"
[5] "yaleB05/yaleB05_P00_Ambient.pgm"
[6] "yaleB06/yaleB06_P00_Ambient.pgm"
[7] "yaleB07/yaleB07_P00_Ambient.pgm"
[8] "yaleB08/yaleB08_P00_Ambient.pgm"
[9] "yaleB34/yaleB34_P00A+095E+00.pgm"
[10] "yaleB35/yaleB35_P00A+095E+00.pgm"
[11] "yaleB36/yaleB36_P00A+095E+00.pgm"
[12] "yaleB37/yaleB37_P00A+095E+00.pgm"
[13] "yaleB38/yaleB38_P00A+095E+00.pgm"
[14] "yaleB39/yaleB39_P00A+095E+00.pgm"

When the threshold is set to 0.05, the minimum value for all the bad images' pixel is 0, and the maximum value for all the bad images' pixel is 0.04705882, and the grey level for the maximum value is "#0A0A0A", which is quite close to black.

d.

When the threshold is set from 0 to 1, and step length is set as 0.05, the proportions of bad images are as follows:

[1] 0.000000000 0.005709625 0.007340946 0.008564437 0.009787928
[6] 0.010603589 0.011011419 0.011827080 0.012642741 0.013050571
[11] 0.013050571 0.013866232 0.015497553 0.017944535 0.021615008
[16] 0.035481240 0.050163132 0.084828711 0.129282219 0.200652529
[21] 1.000000000



I think the best threshold should be 0.6 when the proportion of bad images (38) is 1.55%. After I examine some photos of threshold value 0.6, all the photos are “no flash” or “bad lightening”. When the threshold value is bigger than 0.6, for example, when is set to be 0.65, there is some good photo be chosen, such as ”CroppedYale/yaleB04/yaleB04_P00A+020E-40.pgm” (shown as below). Also, from the plot above, after 0.6, the slope of proportion is getting larger and larger.