Introduction

This is an analysis of the performance (runtime and accuracy) of the Duplicate Image Detector program.

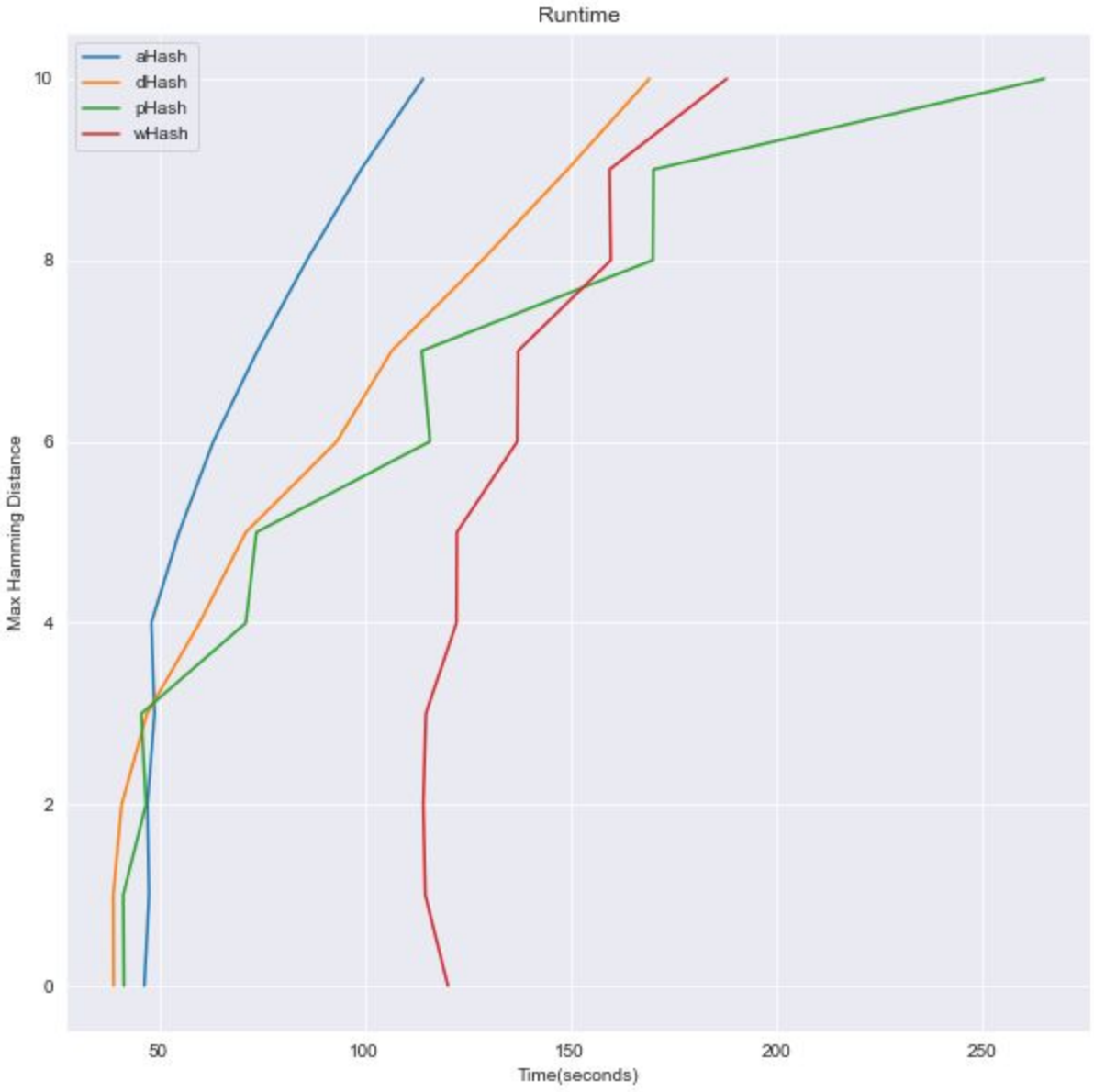
Methodology

The dataset that was used for this test was the Dogs vs. Cats dataset found on Kaggle. I combined the training and test sets (for both the cats and the dogs) in the dataset for a total image set of 10,000 images. I then randomly selected 10% of the images to modify in the following manner:

- One third of selected images were saved in an alternate file format (converted from .jpg to .png)
- One third of selected images were converted to greyscale
- One third of selected images were altered by changing the color of 50 random pixels to black.

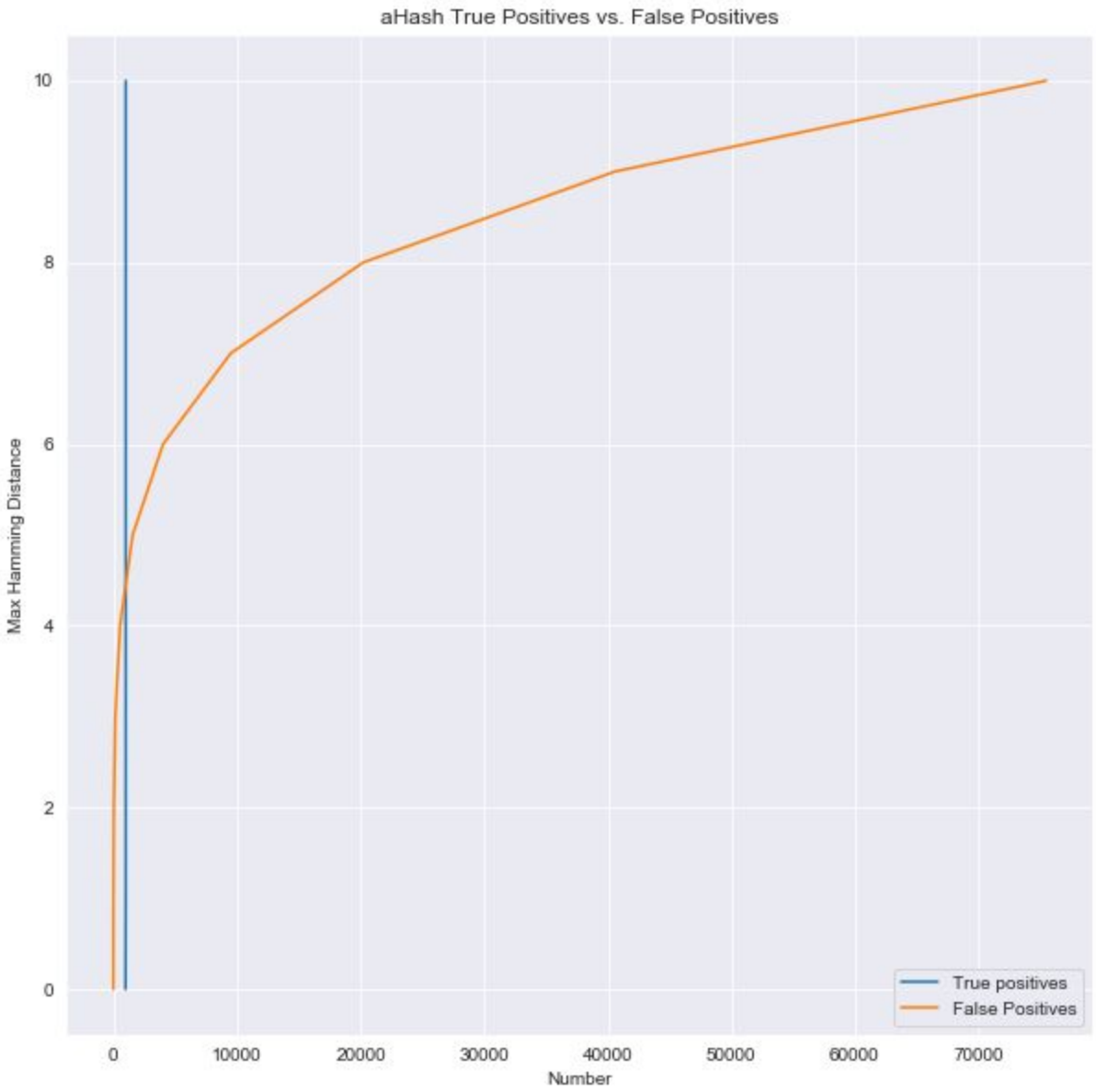
The code that was used to modify the images can be found in the CreateTestPictures notebook in the test directory. The altered images were then added to the dataset, giving a total of 11,002 images (1,002 modified images). The tests were performed using a modified version of dupe_checker.py, which can also be found in the tests directory. An analysis of the performance was done by the Test notebook in the test directory. The results can be found below.

Runtime Chart



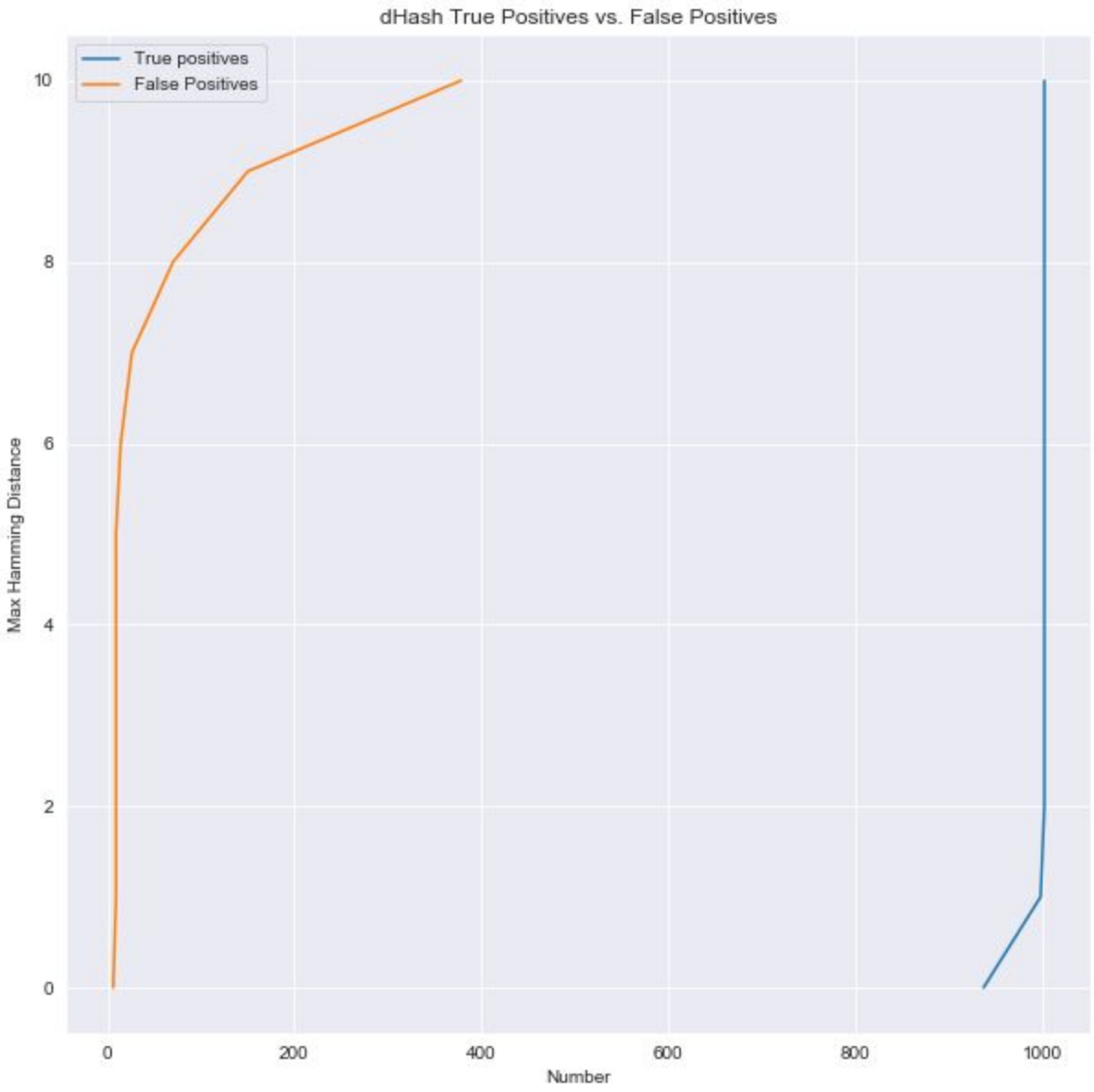
aHash

Max Hamming Distance	Runtime (in seconds)	True Positives	False Positives
0	46.45665712356568	982	4
1	47.56952085494995	1000	12
2	47.184698820114136	1001	43
3	48.961600399017335	1001	150
4	48.167155027389526	1001	540
5	54.92994179725647	1002	1557
6	63.25501799583435	1002	4021
7	73.96183037757874	1002	9496
8	85.9985188484192	1002	20215
9	99.19467363357543	1002	40572
10	114.24873461723328	1002	75480


aHash True Positives vs. False Positives
Max Hamming Distance
Number
0
2
4
6
8
10
0
10000
20000
30000
40000
50000
60000
70000
True positives
False Positives

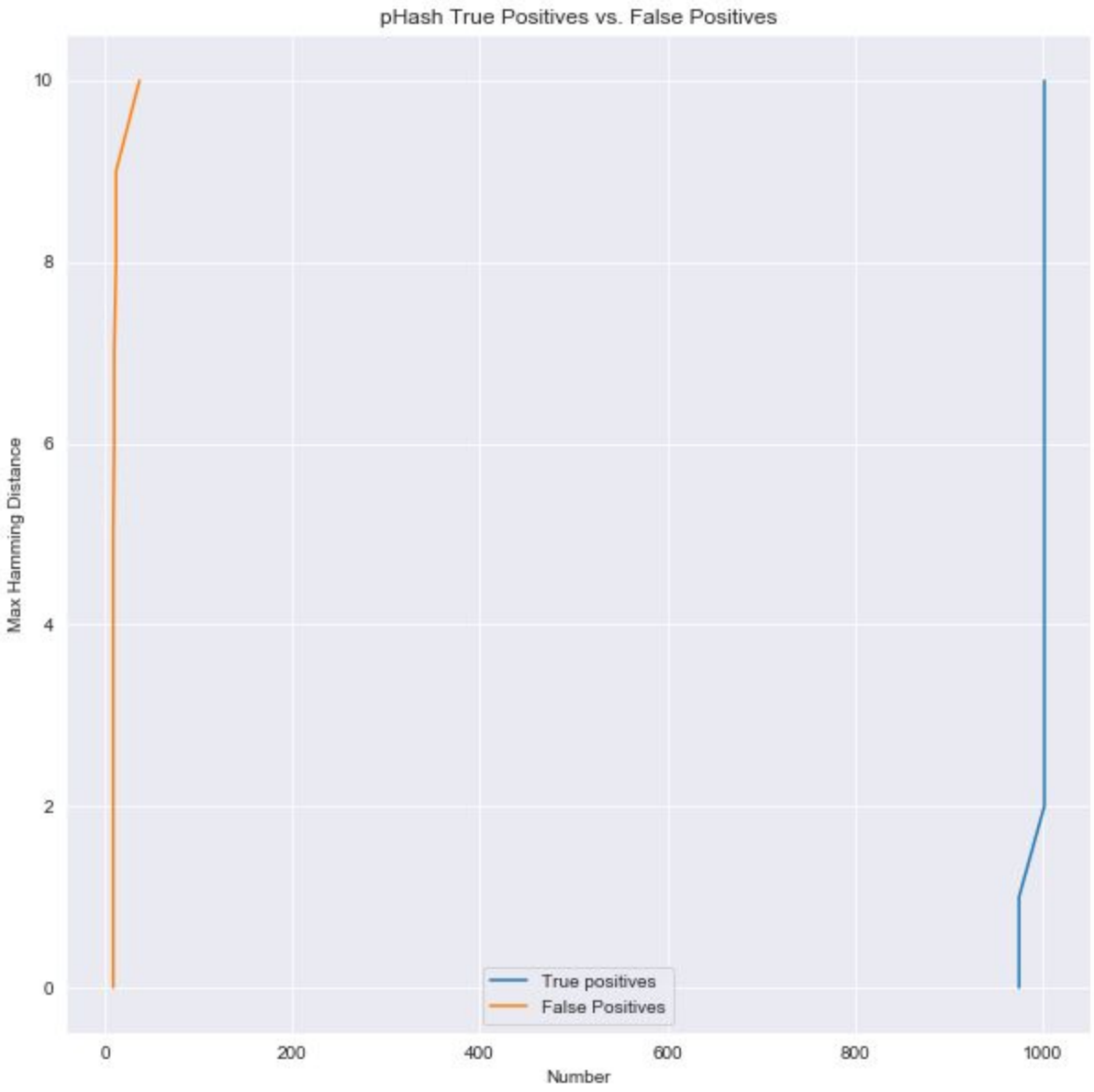
dHash

Max Hamming Distance	Runtime (in seconds)	True Positives	False Positives
0	38.97362914085388	937	6
1	38.89662475585938	998	9
2	40.96754322052002	1002	9
3	47.103694152832034	1002	9
4	59.873224544525144	1002	9
5	71.16007018089294	1002	9
6	93.24793343544006	1002	14
7	106.64049949645997	1002	26
8	128.65555863380433	1002	70
9	149.4123459339142	1002	150
10	169.27968225479125	1002	378


dHash True Positives vs. False Positives
True positives
False Positives
Max Hamming Distance
Number
0
2
4
6
8
10
0
200
400
600
800
1000

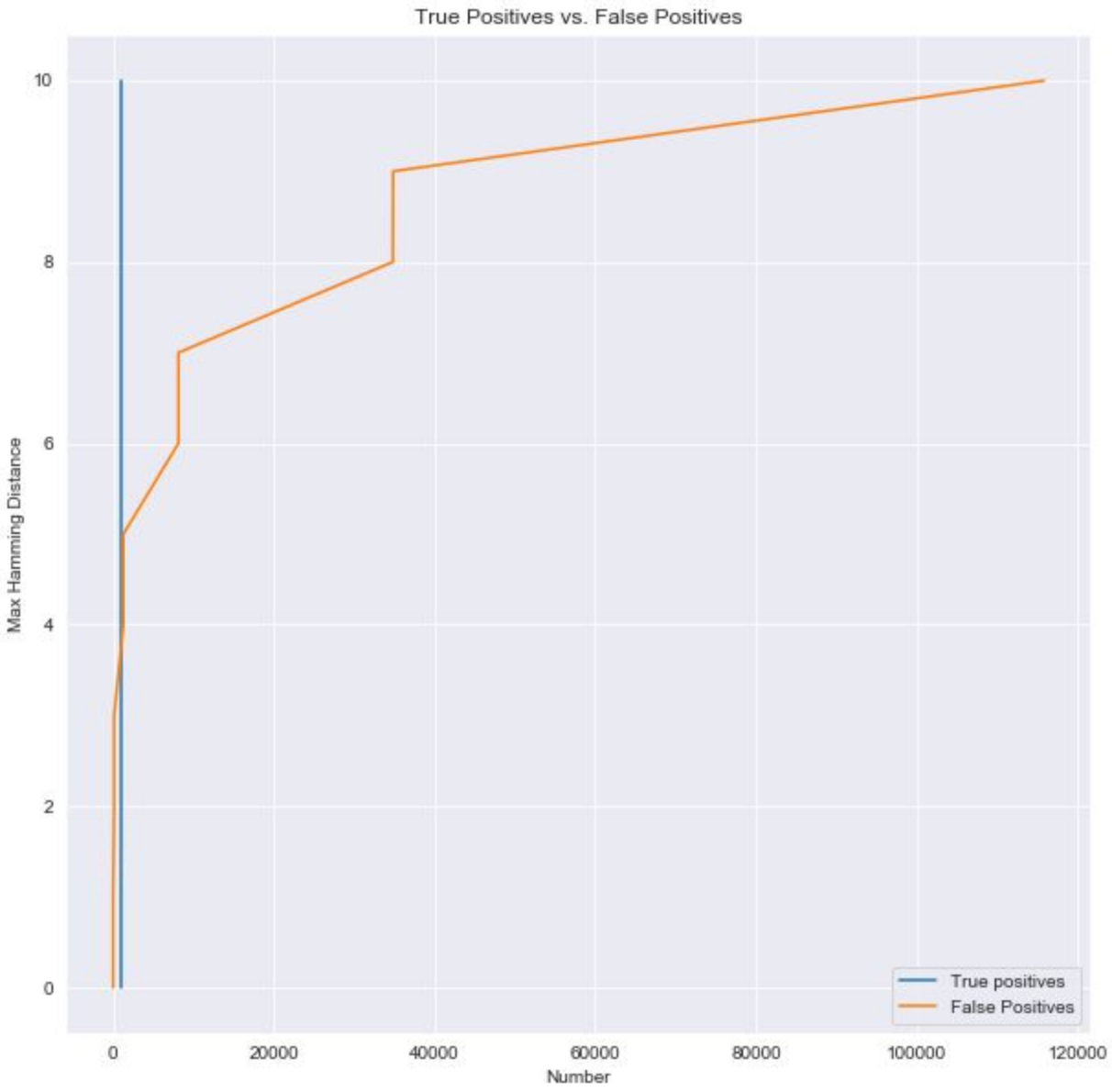
pHash

Max Hamming Distance	Runtime (in seconds)	True Positives	False Positives
0	41.49457335472107	975	9
1	41.32996392250061	975	9
2	46.857480096817014	1002	9
3	45.729815578460695	1002	9
4	71.14286913871766	1002	9
5	73.7318172454834	1002	9
6	115.92383046150208	1002	10
7	113.93391661643982	1002	10
8	170.1307309627533	1002	12
9	170.29474024772645	1002	12
10	265.3555775165558	1002	37


pHash True Positives vs. False Positives
Max Hamming Distance
Number
True positives
False Positives
0
2
4
6
8
10
0
200
400
600
800
1000

wHash

Max Hamming Distance	Runtime (in seconds)	True Positives	False Positives
0	120.29228034019471	995	14
1	114.78176517486573	995	14
2	114.30633792877197	1002	126
3	114.92257323265076	1002	126
4	122.37619953155517	1002	1283
5	122.4960063457489	1002	1283
6	137.16044511795045	1002	8134
7	137.38585801124572	1002	8135
8	159.9345477581024	1002	34821
9	159.60692896842957	1002	34839
10	188.0999587059021	1002	115837


True Positives vs. False Positives
Max Hamming Distance
Number
True positives
False Positives
0
2
4
6
8
10
0
20000
40000
60000
80000
100000
120000