

Report

Goal: to study and verify the influence of various factors on the accuracy of the program (model)

Experiment №1

Split matrix for training / test in various proportions: 50/50, 80/20, 90/10. The photos were separated by the algorithm “every second”, “every fifth” and “every tenth”, respectively. And various coefficients of primary normalization (standardization).

Tab. 1

	50/50	80/20	90/10
/1 (Without primary normalization)	94,5	97,5	97,5
/ 100 (partial)	94,5	97,5	97,5
/ 255 (full)	94,5	97,5	97,5

Inference:

- 1) Primary normalization (standardization) does not affect the accuracy of data definition. This is due to the “Euclidean distance” recognition algorithm used. The algorithm is based on comparing the arithmetic differences between the data vectors, and they changing the absolute values (because of the multiplier) do not change the relative location of the compared vectors.
- 2) The highest recognition accuracy was achieved when dataset was divided in the 80/20 ratio. For further experimentation, this kind of data separation will be used.

Experiment №2

The sample data sampling algorithm has been changed. Applied random sampling method. More than 70 repetitions were performed. The results are summarized in table 2.

Tab. 2

The number of dropped reps	accuracy, %
45	100,00%
19	98,75
6	97,5
2	96,25
Mean:	99,36

Inference:

- 1) A random sample in most cases gave a 100% result. Apparently because of a more uniform selection of test photos, and not particularly highlighted in the list.
- 2) We also see two cases of low accuracy (96.25%) of recognition.
- 3) Due to the unpredictable result of each specific repetition and in order to identify the refined influence of the studied parameters on the accuracy of determining persons from the photo, we do not use the random sampling method.

Experiment №3

We add data normalization to the base indicator. We use the method of "Subtracting the average" and "Dividing by the standard deviation." We will use them separately for features and samples.

We also use two dataset options - att (data without processing) & ds (which are named and sorted)

The results are summarized in table 3.

Tab. 3

	samples		featers	
	att	ds	att	ds
mean	95,00%	100,00%	97,50%	100,00%
std	95,00%	97,50%	97,50%	97,50%

Inference:

When using basic dataset (att), greater accuracy was shown by applying normalization methods to features. From the optimization method, the accuracy has not changed. But when applying a sorted dataset (ds), the result is 100% with the method “minus average”, and a little lower with the method with “division by standard deviation”

Dataset Optimization

Experiment №4

Method One - PCA Algorithm

The value of `pca_energy` is 0.95; 0.98; 0.99

The results in table 4:

Tab. 4

<i>pca_energy</i>	<i>accuracy</i>	<i>Number of featers</i>	<i>T ?</i>
0,95			
0,98			
0,99			

Inference:

With `pca_energy` = XXXXX, we obtained the optimal combination of accuracy level and program speed.

Experiment №5

Photo optimization method (and therefore the whole Dataset) LBT

The results in table 5

Tab. 5

<i>Method</i>	<i>1</i>	<i>2</i>	<i>3</i>
<i>accuracy</i>	XXX	XXX	XXX

Inference: