AN EMPIRICAL STUDY OF THE BEHAVIOR OF A FACE RECOGNITION SYSTEM BASED ON EIGENFACES AND K-NEAREST NEIGHBORS TECHNIQUES

Fabio Abrantes Diniz[1], Thiago Reis da Silva[2], Francisco Eduardo Silva Alencar[3]

Abstract *— Developing a face recognition computational model is a hard task. Extracting facial features from facial images becomes a hard when the images have different dimensions, especially in the steps of extraction and classification. In this paper, we propose an empirical study of optimization on the rate accuracy results from a facial recognition based on Eigenfaces and K-Nearest Neighbors techniques. It was investigated the following topics: images with three different dimensions, number of features (Eigenfaces), k values from K-Nearest Neighbors technique and three distance measures. Addressing the problems of image dimensionality for facial recognition, understanding which parameters are more relevant from the addressed techniques in order to enhance the accuracies rate of facial recognition were the goals of this study. Following this, it was proved from the experiments that images with 12x9 sizes produce the best facial recognition accuracies rate, using the normalized Euclidean distance and a number of Eigenfaces equals to twenty.*

Index Terms — Empirical Study, Recognition System, Eigenfaces and K-Nearest Neighbors Techniques.

INTRODUCTION

Face recognition is one of the most used identification process by humans, allowing then to quick identify any individual. Although facial recognition is a simple task for humans, it is not trivial to implement this process in a machine. Modeling a face that abstracts features that differentiate one face to another faces is the most difficult step in the implementation, since they have few substantial differences [1]. In addition, images with different dimensions make the recognition process a hard task, especially in the steps of extraction and classification of the facial features.

Many algorithms have been proposed to solve the facial recognition problems [2]. In this paper we present an empirical study that optimizes the accuracy rate for a face recognition system [3] based on Eigenfaces techniques [4] and K-Nearest Neighbors (K-NN) [5]. The covered variable techniques which were analyzed are the following: (a) three-dimensional images; (b) number of features ranging from 15 to 20; (c) the value of k of K-NN technique ranging from 1 to 10 and (d) the use of the three measures distances (*Euclidean*, *Manhattan* and the *Normalized Euclidean*). This study was important in the image dimensionality problems for facial recognition in the following way: verifying which of the dimensions of the images provides the best significant facial features. Furthermore, it provided an analysis of the values of the relevant parameters for the techniques were used.

An experiment was done using a database containing 1280 images from a total of 64 individuals. Each individual of this database was represented by 20 images in five different positions of the individual. According to the main results of experimental tests, it was found that images of different sizes produce different rates accuracies. In addition, the best accuracy in facial recognition system reached was had the following combination of parameters: size 12x9, normalized Euclidean distance, value of k equal to one and characteristic number equal to twenty. Also, the images with the smaller dimension analyzed (12x9) produced the best accuracies rates facial recognition than the others studied.

RELATED WORK

There are two basic approaches for face recognition [6]. The first is based on the extraction of feature vectors of the basic parts of a face, such as: eyes, nose, mouth and chin, the second approach is based on information theory concepts [6].

In the first approach, Shu Liao and Chung [9] proposed a new way to formulate the face recognition problem. Each facial image tested was deformed to a face image training in the fixed space for a predefined deformation model. However, in the second approach, Agarwal et. al. [1] it was used the PCA technique, which extracts the most important features of the face of an individual. These characteristics were used as inputs of a neural network classifier of the face.

Therefore, this work uses the second approach, conducting an empirical study on the results of accuracies rates of a face recognition system based on the technical Eigenfaces and K-NN.

[1] Fábio Abrantes Diniz, Federal University of Campina Grande – UFCG/PPgCC, Brazil, fabio.abrantes.diniz@gmail.com
[2] Thiago Reis da Silva, Federal University of Rio Grande do Norte – UFRN/PPgSC, Brazil, thiagoreis@ppgsc.ufrn.br
[3] Francisco Eduardo da Silva Alencar, Rural Federal University of the Semi-Arid – UFERSA/PPgCC, Brazil, eduardu.dudu@gmail.com

DOI 10.14684/ICECE.9.2015.22-26
© 2015 COPEC

DESIGN OF EXPERIMENTS

The design allowed enhancing the quality of the processes, reducing the variability of results, time analysis and cost. In this section it is detailed the experiments, following the process defined by Wohlin [7].

Research questions and hypotheses

The accuracy is widely used to evaluate the efficacy of a facial recognition system. The efficacy of accuracy of the system is related to the system capacity to correctly recognize individuals by their face. Therefore, the main research goals are compare and determine the best values of the parameters provided by the used in order to improve the accuracy of the system when supplied with facial images with different dimensions. Based on the facial recognition information, other Research Questions (RQ) were formulated:

- ***RQ1*** *– What is the distance that applied to a face image with a specific dimension produces better facial recognition accuracies rates?*
- ***RQ2*** *– What is the feature value (Eigenfaces) that applied to face image with specific dimension produces better facial recognition accuracies rates?*
- ***RQ3*** *- What is the k value of the k-NN classifier applied to facial image with specific dimension produces better facial recognition accuracies rates?*

These questions brought out the following hypotheses that answer it, which are:

- **H1-0:** The accuracy of face recognition system using distinct distances is equal.
- **H1-1:** The accuracy of face recognition system using distinct distances is different.
- **H2-0:** The accuracy of face recognition system using distinct k values is equal.
- **H2-1:** The accuracy of face recognition system using distinct k values is different.
- **H3-0:** The accuracy of face recognition system using distinct Eigenfaces is equals.
- **H3-1:** The accuracy of face recognition system using distinct Eigenfaces is different.

Formally, the three hypotheses described above are defined in Table 1.

TABLE 1.
FORMULATION OF HYPOTHESES

Hypothesis	Null Hypothesis	Alternative Hypothese
H1	$H1_0 = A(Dim_1) = A(Dim_2) = A(Dim_3)$	$H1_1 = A(Dim_1) \neq A(Dim_2) \neq A(Dim_3)$
H2	$H2_0 = A(D_1) = A(D_2) = A(D_3)$	$H2_1 = A(D_1) \neq A(D_2) \neq A(D_3)$
H3	$H3_0 = A(k_1) = A(k_3) = A(k_4) = A(k_5) = A(k_6) = A(k_7) = A(k_8) = A(k_9) = A(k_{10})$	$H3_1 = A(k_1) \neq A(k_3) \neq A(k_4) \neq A(k_5) \neq A(k_6) \neq A(k_7) \neq A(k_8) \neq A(k_9) \neq A(k_{10})$

Independent and dependent variables

The independent variables (also called factors) are all variables that can be manipulated or controlled in the process of experimentation, while the dependent variables are only "dependent" measures of manipulation or experimental conditions [8]. This subsection describes the step that selects independent variables (control factors) with its respective levels variation ranges and dependent variables (variables-responses). The factors levels are defined in accordance with Table 2.

TABLE 2.
FACTORS LEVELS.

Factor	Variable Type	Level
Dimension of facial images	Qualitative	120x90, 32x24, 12x9
Eigenfaces	Quantitative	15, 16, 17, 18, 19, 20
Nearest neighbors (k)	Quantitative	1, 3, 4, 5, 6, 7, 8, 9,10
Distances	Qualitative	Euclidean, Manhattan, Normalized Euclidean

In this study, it was only analyzed the accuracy rate as the dependent variable. The independent variables used in the experiment are:

- **Dimension of the facial images**: each image that represents an individual was transformed in na array of length *w x h*, where *w* and *h* are, respectively, the pixel numbers related to the width and height of the image;
- **Main characteristics (Eigenface)**: a set probability distributions vectors (eigenvectors of the covariance matrix) that generates a mathematical information of a human face [4];
- **Nearest neighbors (k):** the algorithm is based on the demand of k nearest neighbors of the test pattern [5]. The pattern is said to belong to the class that present the highest frequency among the k neighbors used [5];
- **Distances:** the search for neighborhood using k-NN algorithm [5] is made using a distance measure in this search.

Design of the experiment and the experimental units

The experimental planning model used was a complete factorial design with 10 replications. In where, it was elaborated 486 tests, performed 10 times each, reaching 4860 executions as a total. Since each execution has a

minimal cost, there is no problem using this type of design. Table 3 describes each of the treatments.

In this experiment, the experimental units are sets of facial images of each dimension. For each dimension, it will be randomly chosen, facial images for the training set and the test set. Therefore, each training set correspondent to it dimension is a random sampling under which the "treatment" of the experiment (number of factors) is applied to obtain addressed dependent variable. It is by means these units that will be possible to obtain the statistical variation in the analysis of research results. There was a randomized to form the training basis with 75% of facial images of the database used for the training set and 25% of the images for the tests.

TABLE 3.
COMPLETE FACTORIAL DESIGN.

Dimensions = 120x90, 12x9 e 32x24										
Distance	***Eigenface***	**K1**	**K3**	**K4**	**K5**	**K6**	**K7**	**K8**	**K9**	**K10**
Euclidian	15	1	3	4	5	6	7	8	9	10
	16	1	3	4	5	6	7	8	9	10
	17	1	3	4	5	6	7	8	9	10
	18	1	3	4	5	6	7	8	9	10
	19	1	3	4	5	6	7	8	9	10
	20	1	3	4	5	6	7	8	9	10
Normalized euclidian	15	1	3	4	5	6	7	8	9	10
	16	1	3	4	5	6	7	8	9	10
	17	1	3	4	5	6	7	8	9	10
	18	1	3	4	5	6	7	8	9	10
	19	1	3	4	5	6	7	8	9	10
	20	1	3	4	5	6	7	8	9	10
Manhattan	15	1	3	4	5	6	7	8	9	10
	16	1	3	4	5	6	7	8	9	10
	17	1	3	4	5	6	7	8	9	10
	18	1	3	4	5	6	7	8	9	10
	19	1	3	4	5	6	7	8	9	10
	20	1	3	4	5	6	7	8	9	10

RESULTS OF THE EXPERIMENTS

This section presents the analysis of the data collected during the execution of the experiment. Our results are packaged the relevant link[4]. For each factor and interaction between the factors in the metric accuracy of the experiment, the following hypothesis were tested:

- **H0:** There are not differences between the effects of the factors of the independent variable.
  H1: There are differences between the effects of the factors of the independent variable.

TABLE 4.
SIGNIFICANCE OF THE EFFECTS FOR ACCURACY RESPONSE-VARIABLE.

	α	β	γ	λ	αβ	αγ	αλ
SS	0.021	0.837	0.092	0.0083	0.0009	0.0133	0.0016
DF	1	1	2	1	1	2	1
MS	0.022	0.837	0.046	0.008	0.0009	0,0066	0.0016
F	227.3	8660.67	477.4	85.948	191.015	69.107	17.152
	βγ	βλ	γλ	αβγ	βγλ	αβγλ	ϵ
SS	0.036	0.001	0.002	0.0012	0.00073	0.0007	0.4674
DF	2	1	2	2	2	2	4836
MS	0.018	0.001	0.001	0.0006	0.0003	0.0003	0.00009
F	191.0	15.18	11.26	6.475	3.954	3.914	-

Legend: α = dimension; β = k-neighbor; y = distance; λ = eigenfaces; ϵ = effect of the experimental error; DF =Degree Freedom; SS = Sum Squares; MS = Mean Squares.

Conducting an F-test, considering a 5% significance level, almost all factors rejected the null hypothesis. Thus, it can be said that most of the factors have statistical significance in relation to the change of accuracy metric. Only the effects of the interaction between the dimension, distance and Eigenface and between the dimension factors , k and Eigenface did not reject the null hypothesis. That happened because their F values are smaller than the value of F table. Table 4 shows the results F-test of the factors that rejected the null hypothesis.

Verification of Hypotheses

Once identified that all the main factors are statistically significant in the response variable accuracy, experiments were performed to evaluate if the accuracy rate is statistically significant compared with the parameters of all the factors.

To answer the hypotheses **H1-0**, **H2-0**, **H3-0**, it was first done a visual analysis of the confidence intervals of each factor separately present in images (Figures 1, 2 and 3). This is illustrated in Figure 1, the analysis for the confidence intervals comparing the distance factor applied to images with dimensions 120x90, 32x24 and 12x9.

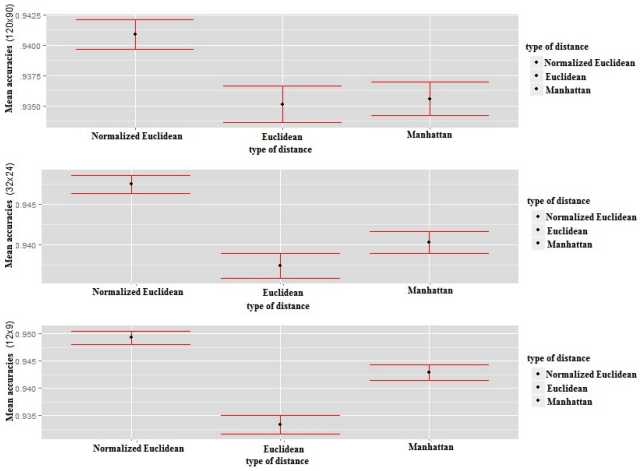

FIGURE 1.
INTERVAL OF CONFIDENCE OF FACTOR DISTANCE.

It can be seen from Figure 1 that at the size 12x9, the best distance is the normalized Euclidean distance, followed by Manhattan and last Euclidean. That is observed for 5% significance level and independent of the completion of the F test. For the dimensions 32x24 and 120x90, it is noticed that the standard Euclidean distance has higher accuracy

[4] https://docs.google.com/file/d/0B4d9yu3dP4GJdU5GNUY1S2R5ak0/edit

than other distances. However, we cannot tell which one is greater or lower between Manhattan and Euclidean distances because of the overlap between the ranges. To be sure who is the greatest, it was done a statistical comparison test (F-test) detailed in Table 5.

TABLE 5.
RESULT OF F-TEST.Y.

	Distance	*k*	*Eigenface*
Value F	477.445	8660.670	85.948
Table F	3.05	3.9	3.9

Figure 2 is an overall analysis regarding the k factor of confidence intervals applied in dimensions with images 120x90, 32x24 and 12x9. In Figure 2, it is noticeable that in the dimension 12x9, k = 1 has the highest accuracy, followed k = 4 and k = 3 for 5% level of significance.

An overlap between k = 5 and k = 6 and between k = 9 and k = 10, so there is no way to tell who is greater or lower between these values. As well, for the dimensions 32x24 and 120x90, there is overlap between k = 1 and k = 4. Therefore, for the dimensions 32x24 and 120x90 it is difficult to know which is the best value of k that produces larger values for the metric accuracy. To certify it, it was done a statistical comparison test (F-test) described in Table 4.

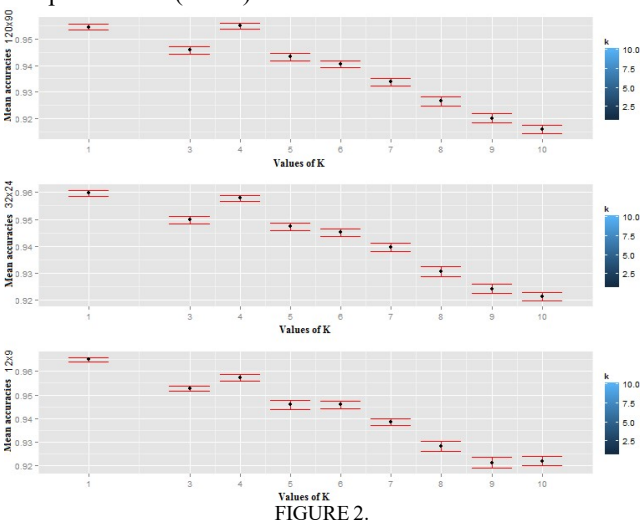

FIGURE 2.
INTERVAL OF CONFIDENCE OF FACTOR K.

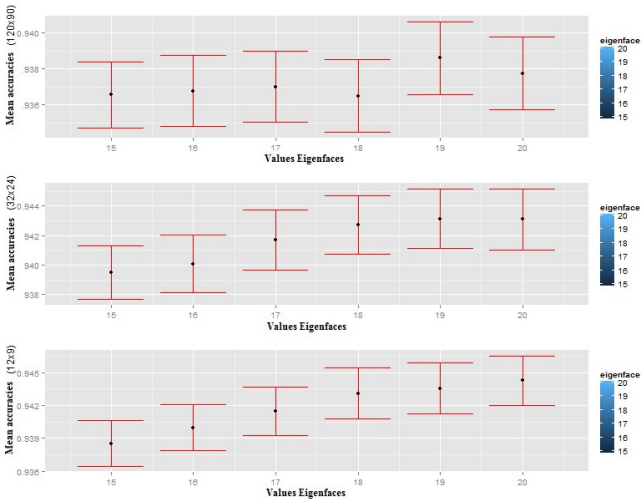

FIGURE 3.
INTERVAL OF CONFIDENCE OF FACTOR EIGENFACES.

Figure 3 presents analysis referring to the confidence intervals comparing the Eigenfaces factor applied to images with the dimensions 120x90, 32x24 and 12x9.

Examining Figure 3, it is seen an overlaps between the Eigenfaces values in all dimensions, then there is no way to say which Eigenface value produces the highest or lowest value of accuracy, and whether or not Eigenfaces factor affects the accuracy of the estimated facial recognition system. To certify it, it was done a statistical comparison test (F-test) which the result is described in Table 4.

Discussion of verifications Hypotheses

According to the visual analysis of the confidence intervals (Figures 1, 2 and 3) and the F-test of all factors mentioned above (Table 4), there was obtained the following findings on the hypotheses addressed.

- As $\mathbf{F_{distance} > 3.05}$ implies the **rejection of the hypothesis H1-0**. Thus, the distance factor affects the variable accuracy. Consequently, it answers the question **QP1** research, because it can be said by analyzing Figure 1 and Table 4, the *Euclidean* distance is normalized to produce more effect on the accuracy with the highest value, followed by the *Manhattan* and finally the *Euclidean*.
- As $\mathbf{F_{Eigenface} > 3.9}$, implies the **rejection of the hypothesis H2-0**. Thus, the *Eigenface* factor generates different accuracies. Consequently, it answers the question of **QP2** research, therefore it can be said, analyzing Figure 3 and Table 4, which, in the dimension *120x90*, the best *Eigenfaces* in ascending order are: *19; 20; 17; 16; 18; 15*. In the dimension *32x24* the best Eigenfaces in ascending order are: *19, 20, 18, 17, 16, 15*. And at the size *12x9 Eigenfaces* are the best in: *20; 19; 18; 17; 16; 15*;
- As $\mathbf{F_k > 3.9}$, implies the **rejection of the hypothesis H3-0**. Thus, the *k* factor affects the variable accuracy.

Consequently, it answers the question of **PQ3** research, therefore it can be said, analyzing Figure 2 and Table 4, *k = 1* is what produced the most accurate, followed by *k = 4, k = 3, k = 5, k = 6, k = 7, k = 8 k = 9 k = 10*.

Lastly, it was concluded that the factor k is that produces the greatest effect on the results of the metric accuracy, followed by the distance factor, dimension and *Eigenface* respectively.

THREATS TO VALIDITY OF THE EXPERIMENT

To evaluate the validity were considered two types of evaluations, namely: external and construction. External validation is related to the approximate accuracy of the conclusions and with the generalization of the population, since the validity of construction concerns the measures used.

Regarding external validation, a threat that can come to arise is related to the quality of the images of the database. Various factors such as low brightness, contrast or brightness may cause the accuracy rate of recognition behaves as not the same observed during the experiment. In addition, the experimental units of the survey are selected from a single source (own database), which may have characteristics that do not apply to all other facial image databases. So, there is a threat to external validity in the interaction of selection and treatment, making it difficult to generalize the results beyond the scope studied.

Finally, a threat to the validity of the construction can take place with respect to the processing time of the algorithm, since a Machine with inferior characteristics may increase the time considerably, it should be considered that the processing will occur on a Machine with same characteristics as the one used in the experiment. Furthermore, it is possible that the levels of selected factors (number of dimensions of the images, number of features) are not enough to observe significant differences in efficacy of face recognition. Thus, has a threat to construct validity caused by confusion between constructs and their levels.

CONCLUSIONS AND FUTURE WORK

The conclusions we can to detail after the development of this project can be divided into two equally relevant approaches. The first approach is the process itself. The scientific method and empirical research are indispensable for conducting quality research. Talking about the collected data, it was performed an exploratory statistical analysis, always directed trying to accept or refute the hypotheses raised about the phenomenon under study. It is presented a logical sequence as was performed the treatment of the data emphasizing the importance of each factor in the model, and, after checking the most important factor, if it has significance.

It was verified that there is significance in the effects of all factors (distances, k, Eigenfaces, size) in the accuracy response-variable. Also, it was observed that the k factor is the most significant in explaining the variation of the values obtained for the response-variable after system execution. As well as, it was verified that the 12x9 dimension produces better accuracies rates in the facial recognition. Finally, it was observed that the best combination of factors that produced the best accuracy in face recognition are: 12x9 size, normalized Euclidean distance, k = 1, Eigenface = 20.

For future work can be detached the study of methods and techniques that can be used to expand the template by the incorporation of other combining techniques classifiers for example: SVM, Neural Networks and Bayesian networks. In addition, the search for new bases of images to perform more experiments and the search for new factors that can significantly influence the recognition of faces with great variations in pose and distinct dimensions of images.

ACKNOWLEDGEMENT

The authors thank CAPES and CNPq for partial support to this research.

REFERENCES

[1] Agarwal, M.; Agrawal, H. and Jain, N. Kumar, M. Face Recognition Using Principle Component Analysis, Eigenface and Neural Network. International Conference on Signal Acquisition and Processing – ICSAP, pp.310-314, 2010.

[2] Zhao W., Chellappa R. Face Recognition: A Literature Survey. ACM Computing Surveys, v.35, n.4, pp.399-458, 2003.

[3] Diniz, F. A.; Mendes Neto, F. M.; Lima Junior, F. C.; Fontes, L. M. O. RedFace: um sistema de reconhecimento facial baseado em técnicas de análise de componentes principais e autofaces: comparação com diferentes classificadores. Revista Brasileira de Computação Aplicada, v. 5, p. 42-54, 2013.

[4] Kshirsagar, V. P.; Baviskar, M. R. and Gaikwad, M. E. Face recognition using Eigenfaces. 3rd International Conference on Computer Research and Development – ICCRD, v.2, pp. 302-306, 11-13, 2011.

[5] Jiangsheng, Y. Method of k-Nearest Neighbors. Institute of Computational Linguistics, Peking University, China, 2002.

[6] Stan Z. Li and Anil K. J. Handbook of Face Recognition, 2nd Edition, Springer. 2011.

[7] Wohlin, C., Runeson, P., H¨ost, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. Experimentation in software engineering: an introduction. Kluwer Academic Publishers, Norwell, MA, USA. 2000.

[8] Prado, M. P.; Souza, S. R. S.; Maldonado, J. C. Resultados de um estudo de caracterização e avaliação de critérios de teste estruturais entre os paradigmas procedimental e OO. 7th Experimental Software Engineering Latin American Workshop – ESELAW, pp.91-100, 2010.

[9] Liao, S., Chung, A. C. S. A novel Markov random field based deformable model for face recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2675,2682, 13-18. 2010.