Principal Component Analysis in Digital Image Processing for Automated Glaucoma Diagnosis

Carla N. Neves¹, Yago de C. Souza¹, Denilson S. da Encarnação¹, Amanda O. R. da Silva¹, Francisco Bruno S. Oliveira¹ and Paulo E. Ambrósio¹

¹ Department of Exact and Technological Sciences, State University of Santa Cruz, Ilhéus, Brazil.

Abstract— In digital image processing (DIP), attribute vectors that tend to contain a large number of elements are generated, and some of these elements are irrelevant for image classification. When working with automated classification techniques, some commonly verified attributes may have low relevance for solving a specific problem or even worse the classification, unnecessarily increasing the dimensionality of the problem. Thus, using methods to reduce the dimension of the problem representation space, there may be better interpretation of the data. In this perspective, in this research the Principal Component Analysis (PCA), a widespread dimensionality reduction technique in the literature, was applied to the attribute vector generated by the DIP, aiming to increase the accuracy of the classifier. As a case study, retinal image classification for the diagnosis of glaucoma was used. The results showed a better classification of the images, validating the possibility of applying PCA to optimize the automated glaucoma diagnosis process.

Keywords— Principal component analysis, dimensionality reduction, glaucoma, digital image processing.

I. INTRODUCTION

Glaucoma is an irreversible neuropathy that is the second leading cause of blindness in the world [1] and represents a major challenge for many areas of science.

This neuropathy is a group of conditions characterized by progressive damage to the optic nerve, most often caused by increased pressure in the intraocular region, causing loss of the visual field. The main risk factors are age, high intraocular pressure, family history and inclusion in a susceptible ethnic group. Still, this disease usually has an asymptomatic evolution [2].

One of the main techniques for the diagnosis of glaucoma is retinography, which consists in the capture of fundus images [3]. It is possible to optimize this process by associating the images obtained with retinography to digital image processing (DIP) techniques, facilitating pattern recognition and characterization of visual occurrences related to the disease.

One of the DIP steps is the Co-occurrence Matrix, which is based on gray-tone spatial dependencies in an image. Its main purpose is the extraction of texture features. These elements compose a set of information defined from a grouping of 14 statistical features of an image, such as: homogeneity, contrast, structural organization, complexity, entropy and energy.

The use of this technique results in the generation of texture attribute vectors with length equivalent to the product of the amount of statistical measures that were used by the number of correlation angles between the pixels. These vectors tend to accommodate a large number of elements, increasing the dimensionality and complexity of the problem.

However, some of this data may be irrelevant or have low influence to the intended task or study, so it is possible to reduce the dimensionality of the problem by discarding certain attributes, turning it into a simpler dataset that will possibly describe the original data topology and will facilitate the achievement of the expected results.

There are many methods for dimensionality reduction for various tasks, such as data mining and pattern recognition, and that demonstrates their importance for problem solving.

Contact: Carla Neves, Campus Soane Nazaré de Andrade, Rod. Jorge Amado, Km 16 - Salobrinho, Ilhéus, Brazil, Phone +5573988433297, cnneves@uesc.br.
In [4] there is a comparative between some of these methods, such as Principal Component Analysis (PCA), Kernel PCA (KPCA), Isomap, Maximum Variance Unfolding (MVU), Laplacian Eigenmaps and others. According to the authors, PCA is one of the most used methods.

From this perspective, the present paper aims to evaluate the classification of fundus images that underwent Principal Component Analysis to validate the use of dimensionality reduction in glaucoma diagnosis.

II. DIMENSIONALITY REDUCTION

According to [5], some tasks such as mining, interpreting or modeling a dataset often do not demand a significant number of attributes, and it can even prejudice the results. As stated by [6], it is necessary to consider a smaller number of variables in a model whenever possible, so that it can be more easily interpreted.

The variables disposal is a technique called dimensionality reduction, and consists in transforming highdimensional data into structures with fewer attributes, which must maintain the most of the original data space topology to keep its relevant properties [7, 8].

A. Principal Component Analysis

According to [9], several methods have been developed for this purpose, but principal component analysis (PCA) is one of the oldest and most widely used methods. The authors also state that the expanded use of large datasets in the area of image analysis has brought methodological advances in data analysis which often finds their roots in PCA.

This method can be seen as the projection of a point swarm in a large dimensionality space down on a lower-dimensional subspace [10]. Its objective is to reduce the dimensionality of a dataset. According to [4], PCA constructs a lowdimensional representation of the data that describes as much of the variance in the data as possible and performs dimensionality reduction by embedding the data into a linear subspace of lower dimensionality.

In this method, the dataset variance is maximally described to represent the different aspects of its topology with less data. It identifies new variables with maximum data variance, called principal components, which are linear combinations of the original variables, and thus the dataset can be represented by a smaller number of attributes [12].

The following table is an adaption from [12]. It describes PCA for a dataset with p variables for n individuals.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>PRINCIPAL COMPONENT ANALYSIS DATA FORMAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>X₁</td>
</tr>
<tr>
<td>1</td>
<td>a₁₁</td>
</tr>
<tr>
<td>2</td>
<td>a₂₁</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>aᵌ₁</td>
</tr>
</tbody>
</table>

Data Format for Principal Component Analysis from n observations of variables X₁ to Xₚ.

According to Table I, each principal component is a linear combination of Xₚ:

\[ Zᵢ = aᵢ₁X₁ + aᵢ₂X₂ + \cdots + aᵢₚXₚ \]  

There are up to p principal components Zₚ for p variables, and they are obtained from an analysis that consists in finding the eigenvalues of a sample covariance matrix which is symmetric and has the form [11]:

\[
C = \begin{bmatrix}
c_{11} & c_{12} & \cdots & c_{1p} \\
c_{21} & c_{22} & \cdots & c_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
c_{n1} & c_{n2} & \cdots & c_{np}
\end{bmatrix}
\]  

Where the elements of the main diagonal (cᵢᵢ) are the variances of Xᵢ and those of the secondary diagonal (cᵢⱼ) represents the covariance between the variables Xᵢ and Xⱼ. There are p eigenvalues λ representing the variances of the principal components in the sample covariance matrix, whose value is either positive or zero. With the eigenvalues ordered as λ₁ > λ₂ > ... > λₚ > 0. The eigenvalues λᵢ are the principal components described in Eq. 1 [11]. The constants aᵢ in Eq. 1 are the eigenvectors, staggered in a manner that

\[ a₁² + a₂² + \cdots + aₚ² = 1 \]  

In [10] it is shown that this procedure is based on the fact that the eigenvectors corresponding to the largest eigenvalues contain the most useful information related to the specific problem and that the others mainly comprise noise. Therefore, these eigenvectors are usually written in descending order according to associated eigenvalues.

From this information, it is understood that the eigenvectors with the largest associated eigenvalues represent the most significant relationship between their dimensions.

For (cᵢᵢ) is the variance of Xᵢ and λᵢ is the variance of Zᵢ, the sum of the variances of the principal components is equal to the sum of the variances of the original variables, and, in this perspective, the principal components contain all variation of the original data [12].

To summarize the process, the main components are linear combinations with coefficients equal to the eigenvectors of the covariance matrix of the data, ordered decreasingly by the associated eigenvalues.

A lower-dimensional subspace is formed with the main components and the matrix with the original data is projected on it, and thus the coordinates of the data in this new plane can be obtained.

B. PCA of a digital image.

The PCA seeks to maximize variance in the attributes that describe the objects of an image, so they must be uncorrelated [13]. Remember that correlation is a measure that indicates how much a linear function can model a relationship between attributes.

In the PCA algorithm, the attributes that describe the image to be analyzed are an input parameter. The steps of this algorithm are described below in Fig. 1 (adapted from [13]):

![Fig. 1: PCA execution](image-url)
III. MATERIALS AND METHODS

A. Imagery database.

This research used an open retinal image database for optic nerve evaluation, RIM-ONE\(^1\). This base is divided into three versions. The second version is being used in this article because it contains a similar number between retinograms of normal and glaucomatous eyes, totaling a set of 455 images (255 healthy and 200 pathological).

B. Texture features.

The analysis of the texture pattern of an image provides a set of statistical information that will aid its classification, such as: spatial distribution, brightness variation, roughness, regularity, etc [14].

For obtaining such information, it was used the cooccurrence matrix, turning the image into a two-dimensional structure of gray level intensity based on the spatial relationship between the pixels.

The co-occurrence matrix is mainly used to apply statistical calculations for feature extraction. In [15] fourteen features were postulated to describe the spatial behavior between the pixels of an image.

However, in this research only nine texture features were applied: angular second moment, contrast, sum average, variance, correlation, sum variance, inverse difference moment, entropy and measure of correlation.

C. Dimensionality reduction stage.

With the mentioned texture features, it was possible to create attribute vectors where the number of elements is equivalent to the product of the number of angles that the pixels of an image were submitted to the co-occurrence matrix by the number of implemented texture features. In this case, 9 texture features were submitted to 4 angles of the co-occurrence matrix, resulting in 36 attributes for each image.

This process generates a vector that is the input parameter for PCA algorithm, which identifies the principal components and creates smaller space to project the original data.

In this research, dimensionality reduction was applied aiming to improve the predictive ability of the classifier for fundus images. This was done by obtaining a more compact representation of the data space to be studied, focusing only on the attributes that are really important for classification.

D. Classification.

Data patterns processing techniques are applied to identify relevant information in large datasets. There are several supports that use these techniques to obtain their results, and this research used WEKA, a free software that contains a collection of machine learning algorithms and rigorous classifiers in the process of automatic glaucoma classification and detection [16].

All mining techniques go through a training process and are used later in the supervised learning phase that uses data to the specified class. In this work, the algorithm implemented from WEKA was the Random Forest (RF) algorithm, which results in a set of results from several decision trees.

Basically, in the classification of random images, this algorithm will use a decision tree results to generate a mode, that is, to return the most frequent class. The methodology employed was cross-validation 10 fold.

This classifier was used on the original attribute space and on space obtained after the dimensionality reduction, to verify the precision obtained with each one of them and compare the results.

IV. RESULTS

The application of the descriptors cited in the previous section generated a vector with 36 attributes for each image, a matrix of 455 rows and 36 columns. PCA was used to reduce the original data space until only 2 attributes remained.

As previously discussed, the attributes that make up the space with new dimensional pattern are linear combinations of the original attributes. Thus, each space with \(p\) attributes shown in Table II is formed by the first \(p\) principal components (PCs) obtained with PCA.

To compare the data spaces performances in the classification, it was observed the percentages of precision, sensitivity and specificity of each classification, taking as a parameter the values of the same rating criteria for the original attributes.

<table>
<thead>
<tr>
<th>Number of attributes</th>
<th>Precission (%)</th>
<th>Sensibility (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>36 (Original space)</td>
<td>79.7</td>
<td>80.3</td>
<td>79.0</td>
</tr>
<tr>
<td>35 - PCA</td>
<td>79.2</td>
<td>78.4</td>
<td>80.2</td>
</tr>
<tr>
<td>34 - PCA</td>
<td>79.5</td>
<td>79.8</td>
<td>79.1</td>
</tr>
<tr>
<td>33 - PCA</td>
<td>79.4</td>
<td>78.9</td>
<td>80.0</td>
</tr>
<tr>
<td>32 - PCA</td>
<td>79.8</td>
<td>79.2</td>
<td>80.6</td>
</tr>
<tr>
<td>31 - PCA</td>
<td>81.8</td>
<td>80.9</td>
<td>83.0</td>
</tr>
<tr>
<td>30 - PCA</td>
<td>79.6</td>
<td>78.7</td>
<td>80.6</td>
</tr>
<tr>
<td>29 - PCA</td>
<td>79.1</td>
<td>79.0</td>
<td>79.2</td>
</tr>
<tr>
<td>28 - PCA</td>
<td>81.1</td>
<td>79.4</td>
<td>83.3</td>
</tr>
<tr>
<td>27 - PCA</td>
<td>80.7</td>
<td>79.9</td>
<td>81.7</td>
</tr>
<tr>
<td>26 - PCA</td>
<td>80.2</td>
<td>78.7</td>
<td>82.1</td>
</tr>
<tr>
<td>25 - PCA</td>
<td>81.5</td>
<td>79.9</td>
<td>83.5</td>
</tr>
<tr>
<td>24 - PCA</td>
<td>80.3</td>
<td>79.6</td>
<td>81.1</td>
</tr>
<tr>
<td>23 - PCA</td>
<td>80.8</td>
<td>80.9</td>
<td>80.8</td>
</tr>
<tr>
<td>22 - PCA</td>
<td>80.3</td>
<td>79.6</td>
<td>81.1</td>
</tr>
<tr>
<td>21 - PCA</td>
<td>80.9</td>
<td>80.2</td>
<td>81.8</td>
</tr>
<tr>
<td>20 - PCA</td>
<td>79.7</td>
<td>79.6</td>
<td>79.9</td>
</tr>
<tr>
<td>19 - PCA</td>
<td>79.4</td>
<td>78.8</td>
<td>80.0</td>
</tr>
<tr>
<td>18 - PCA</td>
<td>78.6</td>
<td>78.6</td>
<td>78.7</td>
</tr>
<tr>
<td>17 - PCA</td>
<td>78.7</td>
<td>78.4</td>
<td>79.0</td>
</tr>
<tr>
<td>16 - PCA</td>
<td>79.3</td>
<td>79.5</td>
<td>79.0</td>
</tr>
<tr>
<td>15 - PCA</td>
<td>78.9</td>
<td>78.5</td>
<td>78.9</td>
</tr>
<tr>
<td>14 - PCA</td>
<td>79.5</td>
<td>79.6</td>
<td>79.4</td>
</tr>
<tr>
<td>13 - PCA</td>
<td>79.9</td>
<td>80.8</td>
<td>78.7</td>
</tr>
<tr>
<td>12 - PCA</td>
<td>78.6</td>
<td>78.6</td>
<td>78.7</td>
</tr>
<tr>
<td>11 - PCA</td>
<td>79.2</td>
<td>79.9</td>
<td>78.4</td>
</tr>
<tr>
<td>10 - PCA</td>
<td>78.6</td>
<td>78.6</td>
<td>78.7</td>
</tr>
<tr>
<td>9 - PCA</td>
<td>78.8</td>
<td>79.8</td>
<td>77.5</td>
</tr>
<tr>
<td>8 - PCA</td>
<td>78.6</td>
<td>79.3</td>
<td>77.7</td>
</tr>
<tr>
<td>7 - PCA</td>
<td>80.1</td>
<td>81.6</td>
<td>78.2</td>
</tr>
<tr>
<td>6 - PCA</td>
<td>77.5</td>
<td>75.9</td>
<td>78.2</td>
</tr>
<tr>
<td>5 - PCA</td>
<td>78.6</td>
<td>78.6</td>
<td>78.7</td>
</tr>
<tr>
<td>4 - PCA</td>
<td>69.3</td>
<td>71.5</td>
<td>66.3</td>
</tr>
<tr>
<td>3 - PCA</td>
<td>59.4</td>
<td>63.4</td>
<td>54.3</td>
</tr>
<tr>
<td>2 - PCA</td>
<td>57.8</td>
<td>62.2</td>
<td>52.1</td>
</tr>
</tbody>
</table>

Percentages of precision, sensitivity and specificity for the data spaces according to WEKA classification. Highlighted: Spaces with 7, 25, 31 and 36 attributes.

The results in Table II shows that while the classification with the original attributes obtained 79.7% precision, with 31 attributes a precision of 81.8% was obtained.
The sensitivity result for the original attributes was 80.3%, while by reducing the attributes to 7, a sensitivity of 81.6% was found. The best specificity was obtained with PCA, by reducing the original space to 25 attributes, corresponding to a percentage of 83.5%. The original space obtained 79.0% for this measure.

Since sensitivity represents the probability of a positive diagnosis in sick people (true positive), the specificity, the probability of a negative diagnosis in non-sick people, and the precision represents the probability of a correct diagnosis, it is necessary to consider the levels of precision and sensibility more significant in the analysis than those of specificity.

In this perspective, the best result set was found in the reduction to 31 attributes, equivalent to a precision of 81.8%, sensitivity of 80.9% and specificity of 83.0% compared to the original values of 79, 7%, 80.3% and 79.0% for the same measures, respectively.

However, it is important to emphasize that some smaller data spaces also had better results in the classification than the original attribute space. For example, the space that contained only 7 attributes had a precision of 80.1% and sensitivity of 81.6% according to the classification, almost the same results that the space containing 31 attributes, that was also obtained through PCA.

V. CONCLUSIONS

One of the main techniques for the diagnosis of glaucoma is retinography, a process that obtains fundus images. One way to optimize retinography is associating the images obtained with digital image processing techniques, such as the Co-occurrence Matrix.

The mentioned procedure results in the generation of texture features vectors that tend to have a large number of elements, increasing the dimensionality and complexity of the problem.

In this perspective, it was possible to describe this data topology with a simpler dataset through a dimensionality reduction method, to improve results in the classification of the retinography images.

This paper presented an approach for optimizing glaucoma diagnosis using a dimensionality reduction technique called Principal Component Analysis (PCA) in fundus images.

By applying a dimensionality reduction technique, the predictive capability of the classifier was improved. That happened because these techniques seek to remove redundant or irrelevant attributes from the database, allowing the generation of a less error-prone classifier.

The objective of the PCA application is to obtain approximate or better results than the original dataset by reducing the robustness of the data in the problem, considering the influence of each one on the classification to optimize the process.

PCA is a consolidated method of dimensionality reduction, used as a basis for several other methods and is still widely used today. This is because its procedure is simple to understand and has a solid theoretical basis.

The results showed an improvement of the classification after the application of PCA, and that allows the use of this method for better results in the automated glaucoma diagnosis.

In addition to showing the operation and application of the PCA and validating the use of this method in glaucoma diagnosis, this work confirms the importance that dimensionality reduction has in the field of image classification.

AKNOWLEDGMENTS

The authors acknowledge financial support from Coordination of Superior Level Staff Improvement (CAPES).

REFERENCES

**Autor 1** has a degree in Mathematics, Computing and its technologies from the Federal University of Southern Bahia (2018). She is currently a master's student in Computational Modeling in Science and Technology at State University of Santa Cruz. Areas of interest: Computational Modeling and applications, dimensionality reduction and general computational intelligence. E-mail address: cnneves@uesc.br

**Autor 2** has a degree in Mechatronic Engineering from UNIFACS (2017). He is a master's student in Computational Modeling in Science and Technology at State University of Santa Cruz. His areas of interest are control theory, optimization and artificial intelligence. E-mail address: ycsouza@uesc.br

**Autor 3** is an Electrical Engineering student at State University of Santa Cruz, currently attending the 6th semester. Has interest in the area of digital processing of medical images and biological signals and related topics. E-mail address: encarna.denilson@gmail.com

**Autor 4** is an Electrical Engineering student at State University of Santa Cruz, currently attending the 6th semester. She is also a student of Hospital Management at University Center of Maringá. She is interested in medical engineering, rehabilitation engineering and related fields. E-mail address: amanda.eng.uesc@gmail.com

**Autor 5** is a Mathematician, doctor in Computational Modeling and Full Professor at Santa Cruz State University. Area of interest: Computational Modeling and applications. Adress: State University of Santa Cruz, Department of Exact and Technological Sciences, Rodovia Jorge Amado, km 16, 45662-900, Ilhéus, BA. fbsoliveira@uesc.br

**Autor 6** is Associate Professor at the State University of Santa Cruz, Brazil, and vice director of the Research Center on Radiation Sciences and Technologies (CPqCTR) in this same University. He is deputy coordinator of the Special Committee on Computing Applied to Health of the Brazilian Computer Society. Has research interest in pattern recognition and computational biology. Adress: State University of Santa Cruz, Department of Exact and Technological Sciences, Rodovia Jorge Amado, km 16, 45662-900, Ilhéus, BA. peambrosio@uesc.br