Centro Científico Conhecer

Enciclopédia Biosfera

PATTERN RECOGNITION IN DIGITAL IMAGES FOR IDENTIFICATION OF SPELLING LINES ADULTERATED BY USING K-MEANS ALGORITHM AND PCA

Hailton David Lemos*(PG)

*Mestrado em Engenharia de Produção e Sistemas (MEPROS), Pontifícia Universidade Católica de Goiás, Goiânia, GO, Brasil
*e-mail: (hailton.david@gmail.com)

Recebido em: 06/10/2012 – Aprovado em: 15/11/2012 – Publicado em: 30/11/2012

ABSTRACT

Pattern recognition in digital images for the identification of lines in adulterated spelling using k-means algorithm and Principal Component Analysis (PCA). Color is used to classify objects of interest from the analysis and comparison of the numerical values of the data, pattern recognition and classification information from the statistical information extracted from patterns through the PCA. The images of the traits used in the study were obtained on equipment that reflects an ultraviolet spectrum which focuses on the trait examined and captured by the camera. The choice of images in the ultraviolet region is made experimentally due to its higher absorbance capability compared with images obtained in other regions of spectrum. Two classes of traits: adulterated and unadulterated. The features are those that receive tainted overlapping somewhere alien. Traces unadulterated do not receive the addition of any strange spot.

KEYWORDS: handwriting; pca; k-mean; pattern recognition

RECONHECIMENTO DE PADRÕES EM IMAGENS DIGITAIS PARA A IDENTIFICAÇÃO DE LINHAS DE GRAFIA ADULTERADA USANDO ALGORITMO K-MEANS E PCA

RESUMO

Reconhecimento de padrões em imagens digitais para a identificação das linhas adulterada na ortografia usando o algoritmo k-means e análise de componente Principal (PCA). A cor é usada para classificar objetos de interesse da análise e comparação dos valores de dados numéricos, padrão de reconhecimento e classificação de informações estatísticas extraídas de padrões através da PCA. As imagens dos traços utilizados no estudo foram obtidas em equipamento que reflete um espectro ultravioleta que incide sobre o traço examinado e capturado pela câmera. A escolha de imagens na região ultravioleta é feita experimentalmente devido à sua capacidade de absorção superior em comparação com imagens obtidas em outras regiões do espectro. Foram utilizadas duas classes de traços:

adulterado e puro. Os traços adulterados são aqueles que recebem sobreposição em alguma parte ou região no traço original. Traços puros não recebem a adição de qualquer corpo estranho a ele.

PALAVRAS-CHAVE: escrita a mão; *pca*; *k-mean*; reconhecimento de padrões

INTRODUCTION

This work describes the solution to a problem presented by the National Sciences and Technologies Advanced Analytical Institute (INCTAA) in partnership with Department of Criminology of the Federal Police of Brazil, which is responsible for verifying the authenticity of documents presented in procedural parts.

The problem now presented is to identify frauds in documents, such as writing overlay, falsification or overlapping signatures, by identifying traits overwritten and underwritten in spelling and handwriting. The aim of the software, result of this work is to allow it´s application from a server in the WEB, which is going to make image processing and will return the results through a WEB page, helping the work of the criminal expert on the conclusion and the report of the document examined.

THEORETICAL BASIS

Here are presented the concepts of digital images, colors system, pixel, neighborhood, vector graphics, algorithm k-means and Principal Component Analysis (PCA), histogram.

DIGITAL IMAGING

Image (from the Latin imago) means the visual representation of an object. For computing, image is a two dimensional representation of an object as a finite set of full digital values, where each value is called Picture element, or pixel. Therefore, an image refers to a function of two-dimensional light intensity, denoted by f (x, y), where the value or range of f in special coordinates gives the intensity (brightness) of the image at the point (PEDROSA & GAMA, 2004; GONZALEZ & WOODS, 2010; SILVA, 2010).

"A panchromatic image is a function of light intensity 2D, f (x, y) where x and y are spatial coordinates and the value of f at the point (x, y) is proportional to the brightness of the scene at that point. If we have a multispectral image, f (x, y) is a vector whose component indicates the intensity of the scene at the point (x, y) and corresponds to the spectrum of band" (PETROU & PETROU, 2010).

"A digital image is an image f (x, y) which has been discretized both in spatial coordinates and brightness. The brightness value digitalized is called the gray level" (PETROU & PETROU, 2010).

In the process of representing the digital image we have the representation of the image by the function f (x, y), which refers to two-dimensional function of light intensity, where x and y denote the spatial coordinates and the value f (x, y) is proportional to the brightness (gray level) of the image at that point (GONZALEZ & WOODS, 2010).

COLORS SYSTEM

The system RGB (Red, Green, Blue) is a system of additive color representation and is based on the theory of the three stimuli. According to this theory, the human eye perceives the colors by stimulating three visual pigments found in cones of the retina, that have sensitivities to certain wavelengths, such as

630 nanometers (red - red), 530 nanometers (green - green) and 450 nanometers (blue - blue); (BIMBO, 1999).

The object color of m is determined by the medium of frequency of the electromagnetic wave packages that their constituent molecules reflects, which are perceived by the people in the specified range, the visible spectrum, approximately 380nm to the violet 740nm to the red (RIBEIRO & MENEZES, 2010).

Two concepts are particularly important for the understanding the concept of color perception. They are: the luminance and chrominance. The luminance contains the information about the amount of black and white colors in an image.

The human brain interprets this information as the quantity of gray present in the color (or brightness). The chrominance informs about the shade of a color. Is the frequency dominant beam. The combination of these two concepts, in different proportions, allows the brain to perceive the visible color spectrum in a particular image or scene. To represent the colors there are models or systems of colors, such as RGB, CMY ("cyan, magenta, yellow"), HSI ("hue, saturation intensity "), YIQ (luminance, emphasis, quadrature), LAB (lightness, green-red, blue and yellow), the standard YUV (intensity, tone, saturation) (GONZALEZ & WOODS, 2010).

PIXEL

A pixel ("picture element" or "pel") is the basic element in an image. The most common form for the pixel is rectangular or square (AGOSTON, 2005). However this study, we adopted an elliptical shape, which after several tests was more efficient on the verifications of the overlapping of colors. An adequate representation of a pixel is rectangular or square. Three are very important aspects in this context of representation: is the neighborhood, connectivity and distances (AZEVEDO & CONCI, 2003).

The pixel is also an element of finite dimensions in the representation of a digital image. Often, the organization of an image in the form of an array of pixels is carried out in a square symmetry in the form of a checkerboard.

With the set of thousands of pixels that form the whole image, it is possible to see each one of the points that make up the image and work this point individually, in its characteristics such as color, luminance, saturation, brightness, tone, array and positioning in the plan under which the image is superimposed (AGOSTON, 2005).

NEIGHBORHOOD

A pixel p at the coordinates (x, y) has four horizontal and vertical neighbors whose coordinates are:

(x +1, y) (x -1, y) (x, y +1) (x, y -1)

This set of pixels is called neighbors of 4 of p. Each pixel is a distance of p and some will be some outside of the image s and (x, y) the edge of the image. There are also 4 diagonal neighbors of p, whose coordinates are.

(x -1, y -1) (x -1, y +1) (x +1, y -1) (x +1, y +1)

Together with the 4 neighbors, this set of pixels are called neighbors of 8 pixel p (AZEVEDO & CONCI, 2003; AGOSTON, 2005). The connectivity between pixels is an important concept used in establishing of the edges of objects and components of the region. To establish whether two pixels are connected is necessary to determine if they are somehow adjacent (say if are neighbors of 4), if levels of color (or gray) satisfy some criterion of similarity. This criterion of connectivity plays a central role in the algorithms group, through them are defined criteria to evaluate whether two points are close and thus form part of a same group. These groups may define a region of interest in the image, for example, pixels that belongs to a network of vessels connected between each other by a criterion of connectivity (AGOSTON, 2005).

SCALABLE VECTOR GRAPHICS

SVG (Scalable Vector Graphics) deals with a language XML to describe in a vector form two-dimensional drawings and graphics, whether a static form whether dynamic or animated. One of the main characteristics of the vector graphics is that don´t lose quality when they are resized. The big difference between the SVG and other vector formats, is the fact that it is an open format, not being property of any company. It was created by World Wide Web Consortium, responsible for definition of other standards like HTML and XHTML (EISENBERG, 2002). In this work the images processed are presented in this format.

K-MEANS ALGORITHM

The K-means algorithm (or k-means) is one of most clustering algorithms known and used. A clustering is no more than one type of learning not supervised, which has the objective to group a set of objects into subsets or clusters (WITTEN, 2000; HAYKIN, 2008). In the context of this study the subsets are the standard colors RGB, each pixel of the image analyzed, grouped according to criteria of luminance, hue, saturation, intensity, color similarity and proximity.

PCA

Principal Component Analysis (PCA) is a method that reduces the dimensionality of the data by conducting an analysis of covariance between the factors analyzed in the image luminance, hue, saturation, intensity, similarity and proximity of color. The PCA applies to tables of data where the lines are considered as an individual and quantitative variables as columns (CHAMBERS, 2008; HUSSON *et al;* 2011). That method is suitable for the various sets of data dimensions, which is the case of the data acquired in this work.

HISTOGRAM

Each pixel in the gray or color image is calculated with a luminance value between 0 and 255. The histogram represents graphically the pixel count value of each possible luminance, or brightness. The total gamma tone value of a pixel is 8 bits 0 .. 255, where 0 is black, and 255 is white.

A histogram is a representation of the frequency distribution of a set of measurements. Typically represented by a bar chart form. The histograms are useful to represent various types of information, such as pattern recognition, for example (GONZALEZ & WOODS, 2010).

METHODOLOGY

In this work the image acquisition was done electronically from original documents using a digital camera. The color model RGB is used by the digital camera used to obtain hadwriting images along with the model of JPEG image compression. Once acquired the image, the calculations are made of statistical data and colorimetric image using k-means algorithm and also the PCA, which allows a grouping by luminance, saturation, intensity and verisimilitude of colors, and generates a file with a vector graphic image highlighting points of interest. During the image processing are also generated two files, one file in HTML standard containing

VBScript code which allows to generate dynamically an electronic spreadsheet with various statistic data and colorimetric of the image, and when it is executed by the browser triggers a plugin by free will of the software Microsoft Excel that shows the graphics and spreadsheet on this software. The second is a file comprising the arrays of the image data processed and also functions needed to be analyzed by statistical software R, such information together which are used to demonstrate which trace was made in the first plan and if so, which trace was done in the second plan in the image. All this information is used by the expert at the time of the report.

In the formation of the processed image to the SVG format, each pixel of the original image read and converted to a new geometric shape, an ellipse, and the interior of the ellipse is filled by the color of the pixel of the original image, and will have a bigger axis with a size 0.8 u.m. (units of measurement), while its smaller axis has length of 0.5 u.m. and these values can be changed later in execution mode, which helps the visual analysis of the processed image. The axis values were selected from the information acquired by k-means algorithm and by the PCA, which was considered the relationship of the variation in brightness, saturation and appearance of the original image colors.

RESULTS

The following shows an original image and the results of processing the image. Because it is a controlled experiment was known beforehand, which were the superscript and subscript features of the spelling, and there was also overlap by a pencil writing with a blue ballpoint pen. Thus it was possible to validate the proposed model of visual form with the help of this program developed in this work. Figure 1 shows the raw image, in the context of this study, the raw image is the one acquired and which will go through the process to generate the statistical data and colorimetric. Figure 2 shows the result of processing indicating the traces above and below of the spelling.

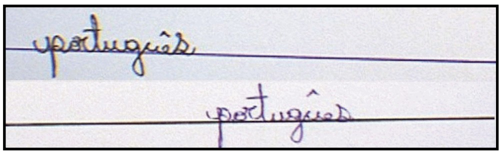

FIGURE 1 - Original Image

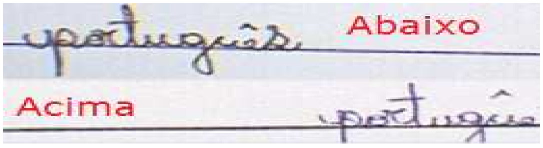

FIGURE 2 - processed image

The following are images generated and processed following the pattern established. Each pixel of the image is an ellipse of bigger and smaller axis with 0.8 u.m. dimensions and a 0.5 u.m. respectively, in this case are different luminance, Saturation and verisimilitude of color.

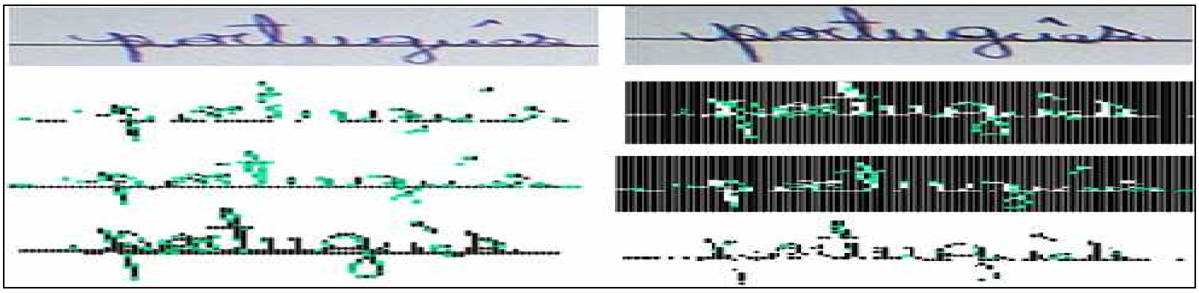

FIGURE 3 - Images processed with different luminosities, saturation and verisimilitude.

Following the same reasoning applied to the previous image, were made changes in the measurement units of the ellipse´s axis and also the luminance and saturation of the image. As the ellipse´s axis increases the colors of the pixels that composes the image becomes sharper, and there exists superposition of colors, in this case graphite of a pencil and the blue color of a ball pen.

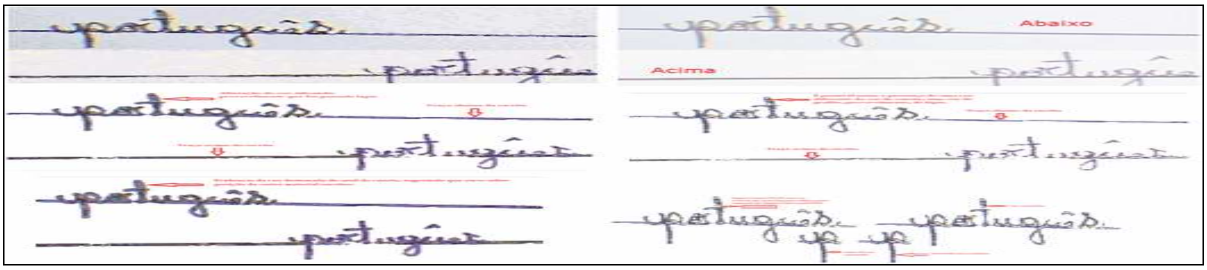

FIGURE 4 - Images processed with different brightnesses and saturations and axis of the ellipse.

The images below were generated by the standard YUV, which highlights the overlap of the traces written with pencil and superimposed by a blue ink pen

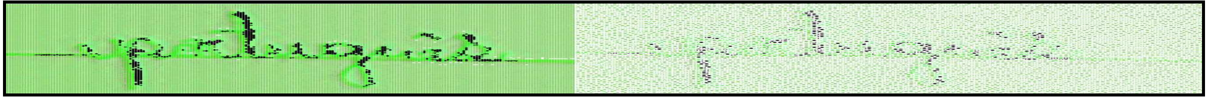

FIGURE 5 - Images processed according to standard YUV.

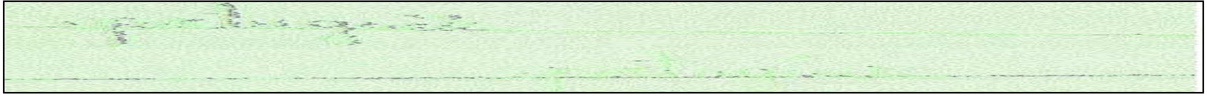

FIGURE 6 - Images processed according to standard YUV increasing the brightness.

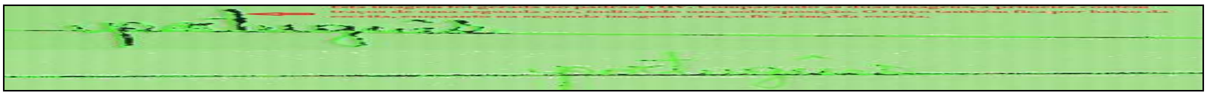

FIGURE 7 - Images processed according to standard YUV highlighting the overlay pencil / pen of the first image, the second image there was no overlap.

CONCLUSION

This study presents the results of the application of image processing techniques using the k-means algorithm and the PCA for the detection of overlapping of the spelling and traces, as well as overlapping the printing of fixing elements such as: ballpoint pen ink and graphite materials among others, that may change the composition of the original document, these elements being analyzed and posted during processing of the image analysis, indicating the criminal expert if there is the possibility of fraud or not the document analyzed. This prototype is being tested with other controlled experiments and their results analyzed and certified, and this software will be available to the Department of National Criminalist Federal Police in Brasilia, as an auxiliary tool in forensic for forensic for the analysis documents adulteration.

As future work, which is already under development, this Being able to detect other types of fraud that is characterized by the aging of paper, to give the impression, for example, it is an old document. That analysis is done taking into consideration morphometric characteristics, folds, edges and Also colorimetric and the images were collected using a digital camera using light ultraviolet, and we will be releasing soon.

REFERENCES

AGOSTON, M.K. **Computer Graphics And Geometric Modeling**. Springer, 2005.

AZEVEDO, E.; CONCI, A. **Computer Graphics: Image Generation**. Campus, 2003.

BIMBO, A.D. **Visual Information Retrieval**. Academic Press, 1999.

CHAMBERS, J. M. **Software For Data Analysis: Programming With R**, Crc Press, 2008.

EISENBERG, J. D. SVG Essentials. O'reilly, 2002.

GONZALEZ, R.C., WOODS, R.E. **Digital Image Processing**. São Paulo: Addison Wesley Brasil, 2010.

HAYKIN, S. **Neural Networks: Principles and Practice**. Bookman - 2008.

HUSSON, F. ; LEE, S. PAGES, JEROME. **Exploratory Multivariate Analysis By Example Using R**. Crc Press, 2011.

PEDROSA, A.C.;GAMA, S.M.A. **Introduction To Computational Probability and Statistics**. Porto, Portugal: Porto Editora, 2004.

PETROU, M.;PETROU, C.**Image Processing: The Fundamentals 2nd Ed** John Wiley & Sons Ltd, 2010.

RIBEIRO, M.M.; MENEZES, M.A.F.**A Brief Introduction to Computer Graphics**. Publisher Modern Science, 2010.

SILVA, I.N.**Artificial Neural Networks: For Science And EngineeringApplied**. São Paulo: Artliber, 2010.

WITTEN, I.H. **Data Mining: Practical Machine Learning Tools With Java Implementations**. Morgan Kaufmann, 2000.