Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080614

Dimension Reduction Parameters for Leukemia Diagnostic based in Subspace Arrangement Segmentation

Leticia Flores-Pulido[1,*], *Gustavo Rodríguez-Gómez*[2] *and Jesús A. González*[2]

[1] Faculty of Engineering and Technology, Computer Engineering Department, Autonomous University of Tlaxcala, Apizaco, Tlaxcala, Mexico
[2] Department of Computer Science, National Institute of Astrophysics, Optics and Electronics, Puebla, Mexico, Tonantzintla, Puebla, Mexico

Received: 15 Oct. 2013, Revised: 13 Jan. 2014, Accepted: 14 Jan. 2014
Published online: 1 Nov. 2014

Abstract: This paper presents a novel approach for the classification of acute leukemia subtypes using image processing and mathematical techniques. The preprocessing phase analyses 376 features from abdnormal leukocytes images. The features or parameters are Leukemia Parameters that helps to lymphoblastic subtypes detection which come from bone marrow images with heterogeneous staining. The second phase imply the robust generalized principal component analysis as segmentation method for data classification into a subspace arrangement with tree dimensions for each plane of lymphoblastic subtype and four dimension for the subspace arrangement. The novel of our proposal states that the two subtypes of acute leukemia can be classified into a subspace arrangement trough robust generalized principal component analysis method. The subspace arrangement is achieved with singular value decomposition, an hibrid linear model to noise samples detection and homogeneus polynomial. Test reveals that variation in dimension of subspace arrangement depends on features size, the outliers percentage and noise parameters are tunned, dimension of subspace and effective dimension are adjusted, time in execution algorithm and segmentation percentage are measured to lymphoblastic subtypes classification with only 4 parameters from 376 attributes set that are previously computed from cell images and their respective nucleous and cytoplasm.

Keywords: Leukemia feature extraction, generalized principal component analisis, lymphoblastic subtype, homogeneus polynomial, subspace arrangement.

1 Introduction

Leukemia is a type of cancer that starts in the bone marrow. The cause of its production is of immature leucocytes. This leucocytes replaces normal blood cells. The body is then exposed to many diseases let them without defenses. This cancer is one of the causes of many deaths in Mexico. The National Institute of Statistics, Geography, and Informatics [1] reported as the third cause of death in 65 of 100 people where 13.1 % were woman and 14.6 % were men, only for people of old age. Leukemia can be detected in early stage and can be treated with a complete blood count. The abnormalities in this count can be detected by morphological bone marrow smear analysis. This analysis is done to confirm the leukemic cells presence. The pathologist uses a microscopy to observe the cells looking for abnormalities in cytoplasm of the cells classify types and subtypes of leukemia. The classification of this data can be taken as support to diagnostic process in order to determine the kind of treatment given. The goal of this paper is devoted to subtypes detection of lymphocytic leukemia thorough feature information inside cytoplasm of cell images [2], [3] Specifically, there are two types of acute lymphocityc leukemia: L1, L2 and L3, but the samples of lymphocytic leukemia handled in this paper are only of L1 and L2 subtypes. Other approaches allows segmentations of leukocytes with markov random fields and teager energy in [4] and [5] and fuzzy approach as in [6], [7]

[*] Corresponding author e-mail: aicitel.flores@gmail.com

2 The Robust Generalized Algorithm

The RGPCA algorithm is implemented as a variation of the GPCA Algebraic algorithm in a semi-supervised fashion with noiseless data. The description of the RGPCA is as follows: The first phase depends about: (1) A number of subspaces that are defined for the number of sets or classes desired. (2) The total of the dimension is another data that depends of the number of features of the system plus 1 (maximum of dimension of the arrangement). (3) Other input data is the matrix of N points with the feature vector of one of each leukemia data. The second phase is about the polynomial embedding. This phase generates a polynomial set that allows the intersection between planes, the bases of each subspaces and the veronese map required for the final phase. The third phase imply the computation of polynomial fitting that allows the equation linear system. Inside this phase the computation of a singular value decomposition is performed. The fourth phase of the system obtains the Jacobian matrix to obtain the bases that allow the final segmentation or clustering.

The implementation of this stage requires to choose the most suitable version of GPCA algorithm that Yi Ma [8] offers, then the feature leukemia parameters are taken as input for the algorithm.

The original GPCA algorithm presented in [9] and improved in [8] is included in this Section. This algorithm will be applied at leukemia diagnostic. The GPCA algorithm is given below in Table 1.

One of the first versions is based in robust GPCA with influence (RGPCA-I), second version is based in the robust GPCA with influence speedup (RGPCA-IS) and the third version is about robust GPCA with multivariate timming (RGPCA-MVT). One of each version was briefly described as follows:

Robust GPCA with influence (RGPCA-I):] This approach classifies outliers form a set of small probability samples with respect to the distribution in question. The given data set is therefore an atypical set if such samples constitute a significant portion of the data.

Robust GPCA with influence speedup (RGPCA-IS): The second approach classifies outliers form a set of samples that have relatively large influence on the estimated model parameters. A measure of influence is normally the difference between the model estimated with and without the sample in question.

Robust GPCA with multivariate timming (RGPCA-MVT): In this case, outliers form a set of samples that are not consistent with (the model inferred from) the remainder of the data. A measure of inconsistency is normally the error residue of the sample in question with respect to the model.

The Multivariate timming process (MVT) is described as follows: First, an initially robust mean of samples are obtained, then a trimming parameter need to be specified equivalent to the outliers percentage. A Mehalanobis distance is computed and a Mehalanobis distance also using samples of polynomials. Then a difference between both distance is iterated as stop criteria that ends the algorithm [8]. We made an analysis about visual classification of the three of the versions where the MVT results the highest in classification, spare data and clearest definition more than one model of subspaces.

Table 1: Algorithm 2: GPCA (Taken from [8])

```
         Given a set of samples $(z_1, z_2, ..., z_n)$
         from a (transversal arrangement) of $n$
         linear subspaces with dimensions
         $(d_1, d_2, ..., d_n)$ in $\mathbb{R}^D$
STEP 1.  Construct the matrix
         $L_n = (v_n(z_1), v_n(z_2), ..., v_n(z_N))$.
STEP 2.  Compute the singular value
         decomposition (SVD)
         of $L_n$ and let $C$ be the matrix
         whose columns are the
         singular vectors associated with all
         zero singular values.
STEP 3.  Construct the polynomials
         $Q(X) = C^T v_n(X)$.
STEP 4.  for all $1 \leq i \leq n$ do
STEP 5.      Pick one point $z_i$
             per subspace $V_i$
             and compute the Jacobian
             $J(Q)(z_i)$.
STEP 6.      Compute a basis
             $B_i = (b_1, b_2, \ldots, b_{i_d})$ of $V_i$
             from the right null space
             of $J(Q)(z_i)$ via the singular
             value decomposition
             of $J(Q)(z_i)$ .
STEP 7.      Assign samples
             $z_j$ that satisfy
             $B_i^T z_j = 0$
             to the subspace $V_i$.
STEP 8.  end for
```

3 Existence and analysis of Subspace Arrangements

This section provides a technical explanation about theorical approaches of algebraic concepts that hold the fundamentals of this research. Some of the most important concepts to define distance between polynomials between planes are the Sampson Distance that is explained in this section. Other of the basic concepts are the singular value decomposition whose intention is to expose the importance of discrimination fratures or attributes to define between the two classes of data treatment in leukemia diagnostic. The subspace arrangement concept is provided with the intention to understand the hyperplanes array that represents the final classification or segmentation of data. Finally, the GPCA

Algorithm is provided in detail to understand how the previous concepts are handled in this segmentation method.

Sampson Distance We assume that he polynomials in $\mathscr{Q}(\mathscr{X})$ are linearly independent [8]. Given a point z close to the zero set of $\mathscr{Q}(\mathscr{X})$, i.e., the subspace arrangement $\mathscr{A}$, we let $\hat{z}$ denote the point closest to z on $\mathscr{A}$. Using the Taylor series of $\mathscr{Q}(\mathscr{X})$ expanded at z, the value of $Q(X)$ at $\hat{z}$ is given by

$$Q(\hat{z}) = Q(z) + J(Q)(z)(\hat{z} - z) + O(\|\hat{z} - z\|^2). \tag{1}$$

After ignoring the higher order terms and nothing that $Q(\hat{z}) = 0$, we have

$$z - \hat{z} \approx (J(Q)(z)^T J(Q)(z))^\dagger J(Q)(z)^T Q(z) \in \mathbb{R}^D \tag{2}$$

where $(J(Q)(z)^T J(Q)(z))^\dagger$ is the pseudo-inverse of the matrix $(J(Q)(z)^T J(Q)(z)$. Thus, the approximate square distance from z to $\mathscr{A}$ is given by

$$\|z - \hat{z}\| \approx Q(z)^T (J(Q)(z) J(Q)(z)^T)^\dagger Q(z) \in \mathbb{R} \tag{3}$$

The expression on the right-hand side is known as the Sampson distance [3]. Thus, the average Sampson distance:

$$\frac{1}{N}\sum_{i=1}^{N} Q(z_i)^T (J(Q)(z_i) J(Q)(z_i)^T)^\dagger Q(z_i) \tag{4}$$

is an approximation of the mean square distance. Minimizing the Sampson distance typically leads to a good approximation to the maximum-likelihood estimate that minimizes the mean square distance. There is however, a certain redundancy in the expression of Sampson distance. If $\mathscr{A}$ is the zero set of $Q(X)$, it is also the zero set of the polynomials $Q(X) = MQ(X)$ for any nonsingular matrix $M \in \mathbb{R}^{mxm}$. It is easy to check that the Sampson distance is invariant under the nonsingular linear transformation M. Thus the estimate of polynomials in Q that minimize the average Sampson distance (or the mean square error) is not unique, at least not in terms of the terms of the coefficients of the polynomials in $Q(X)$.

One way to reduce the redundancy is to impose some constraints on the coefficients of the polynomials in $Q(X)$. Notice that

$$(J(\hat{Q})(z_i) J(\hat{Q})(z_i)^T) = MJ(Q)(z_i) J(Q)(z_i)^T M^T \tag{5}$$

and, if there is no polynomial of lower degree (than those in $Q(X)$) that vanishes on $\mathscr{A}$, the matrix

$$\frac{1}{N}\sum_{i=1}^{N} (J(Q)(z_i) J(Q)(z_i)^T) \varepsilon \mathbb{R}^{mxm} \tag{6}$$

is a positive definite symmetric matrix. Therefore, we can choose the matrix M such that the following is the identity:

$$\frac{1}{N}\sum_{i=1}^{N} (J(Q)(z_i) J(Q)(z_i)^T) = I_{mxm} \tag{7}$$

Thus, the problem of minimizing the average Sampson distance now becomes a constrained nonlinear problem:

$$Q^* = argmin_P \frac{1}{N}\sum_{i=1}^{N} (Q)(z_i)^T (J(Q)(z_i) J(Q)(z_i)^T)^\dagger Q(z_i) \tag{8}$$

subject to

$$\frac{1}{N}\sum_{i=1}^{N} J(Q)(z_i)(J(Q)(z_i)^T = I_{mxm} \tag{9}$$

Many nonlinear optimization algorithm can be employed here to minimize the above objetivo function via iterative gradient-descent techniques. However, in order for the iterative process to coverge quicly to the global minimum, a good initization is needed. Below we discuss one such method.

Singular Value Decomposition The principal components of a set of data in $\mathbb{R}^p$ provide a sequence of the best linear approximations to that data, of all ranks $q \leq p$ [8]. Denote the observations by $x_1, x_2, \ldots, x_N$ and consider the rank-q linear model for representing them

$$f(\lambda) = \mu + V_q \lambda, \tag{10}$$

where μ is a location vector in $\mathbb{R}_p$, V_q is a pxq matrix with q orthogonal unit vectors as columns, and λ is a q vector of parameters. This is the parametric representation of an affine hyperplane of rank q. Fitting q value of such a model to the data by least squares amounts to minimizing the *reconstruction error*

$$min_{\mu,\{\lambda_i\},V_q} \sum_{i=1}^{N} \| x_i - \mu - V_q \lambda_i \|^2 \tag{11}$$

We can partially optimize μ and the λ_i to obtain

$$\hat{\mu} = \bar{x} \tag{12}$$

$$\hat{\lambda}_i = V_q^T (x_i - \bar{x}) \tag{13}$$

This leaves us to find the orthogonal matrix V_q:

$$min_{V_q} \sum_{i=1}^{N} \|(x_i - \bar{x}) - V_q V_q^T (x_i - \bar{x})\|^2 \tag{14}$$

For convenience we assume that $\bar{x} = 0$ (otherwise we simply replace the observations with their centered versions $\bar{x}_i - \bar{x}$). The $p \times p$ matrix $H_q = V_q V_q^T$ is a projection matrix, and it maps each point x_i on to its rank-q reconstruction $H_q x_i$, the orthogonal projection of x_i

onto the subspace spanned by the columns of V_q. The solution can be expressed as follows. Stack the (centered) observations into the rows of an $N \times p$ of matrix X. We construct the ***singular value decomposition*** of X:

$$\mathbf{X} = \mathbf{UDV}^T \tag{15}$$

Here $\mathbb{U}$ is an $N \times p$ orthogonal matrix $(\mathbf{U^T U = I_p})$ whose columns u_j are called the *left singular vectors*, and D is a $p \times p$ orthogonal matrix $(\mathsf{V^T V = I_p})$ with columns v_j called the *right singular vectors*, and D is a $p \times p$ diagonal matrix, with diagonal elements $d_1 \geq d_2 \ldots \geq 0$ known as the *singular values*. For each rank q, the solution V_q to (14) consist of the first q column of V. the columns of UD are called the principal components of X. The N optimal λ_i in equation (13) are given by the first q principal component (The N rows of the $N \times q$ matrix $U_q D_q$).

The SVD is specially used in discrimination of leukemia cell features as it is explained in [10], [11], [12], and [13] with the intention to reduce the search space and increasing segmentation percentage.

The Veronese map of degree h is the map

$$v_h : F^D \to F^{M_h^{[D]}} \tag{16}$$

given by

$$v_h \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} = \begin{pmatrix} x_1^h \\ x_1^{h-1} x_2 \\ \vdots \\ x_D^h \end{pmatrix} \tag{17}$$

An arbitrary homogeneous polynomial $q(X)$ of degree h in $X = \{X_1, X_2, ..., X_D\}$ can be written as $q(X) = c^T v_h(X)$ for some vector $c \in \mathbb{F}^{M_h^{[D]}}$ that collects all the coefficients associated with the monomials [8].

Vanishing Ideal Let $I_1, ..., I_r$ be the linear ideals in an infinite field $k[x_1, ..., x_n]$ that are the defining ideals of the subspaces in $\mathscr{A}$. Denote by $V_{\mathscr{A}}$ the union of the subspaces in $\mathscr{A}$ [9]. The vanishing ideal of $V_{\mathscr{A}}$ is the reduced ideal $(rad(I_r))$

$$I_{\mathscr{A}} = I_1 \cap \ldots \cap rad(I_r) \tag{18}$$

When $\mathscr{A}$ is an arrangement of hyperplanes its vanishing ideal $I_{\mathscr{A}}$ is a very simple object - a principal ideal generated by the product of linear forms that define the hyperplanes. In general, the ideal $I_{\mathscr{A}}$ is generated by products of linear forms up to a radical, since:

$$rad(I_1 \ldots I_r) = rad(I_1) \cap \ldots \cap rad(I_r) = I_1 \cap \ldots \cap Ir = I_{\mathscr{A}} \tag{19}$$

but is difficult to construct a nice system of generators of $I_{\mathscr{A}}$ itself. Geometrically, is required to find generators of $I_{\mathscr{A}}$.

Subspace Arrangement A subspace arrangement in F^D is a union

$$\mathscr{A} \doteq V_1 \cup V_2 \cup \ldots \cup V_n. \tag{20}$$

of n subspaces $V_1, V_2, ..., V_n$ of F^D.

For a non empty subset S of the index set $\{1, 2, ..., n\}$, we define the intersection

$$V_S \doteq \cap_{s \in S} V_s \tag{21}$$

with dimension $d_S \doteq dimV_S$ and co dimension $c_S \doteq D - d_S$ [9].

4 Perspective

This section explains the tests achieved for the leukemia diagnosis. The segmentation process handles parameters obtained from features extracted from samples of cells. The cells, the nucleus and the cythoplasm reveals important features about abnormalities in bone marrow for the cancer detection. The approach tested in this work apply RGPCA in segmentation data for classification of abnormalities in two types of leukemia: L1 and L2. There were three kind of evaluations that compares results obtained in segmentation data of leukemia features:

Evaluation One: Parameters variation in dimension of subspaces. This test obtains 10 important results. Where the noise level takes values from $\langle 0.01, 0.015, 0.02 \rangle$. The outliers percentage is changed only once from $\langle 0.06 - 0.02 \rangle$. The variation of segmentation error is of 15.32 (in the best case) and 51.26 (in the worst case). The execution time was meassured and the best case was obtained with 11 seconds. The sets dimension size (of L1 and L2 sets) was variated from $\langle 2,2 \rangle, \langle 3,3 \rangle, \langle 3,3 \rangle, \langle 4,4 \rangle, \langle 5,5 \rangle and \langle 6,6 \rangle$. The max dimension must be the max dimension of the sets plus one, so, this parameter is increased in a range of $\langle 3 - 7 \rangle$. So, the best case result in the Test Number 5 where the lowest segmentation error was of 15.32% with 0.02 of noise level, 0.2 of outliers percentage, the execution time is of 1 minute with 32 seconds, with size dimension of the sets of $\langle 3,3 \rangle$ and max dimension of 4. This results can be observed in Table 2.

Evaluation Two: Decreasing Noise Level and Outlier Percentage. This test obtains 6 important results. Where the noise level takes values from $\langle 0.005, 0.01, 0.02 \rangle$. The outliers percentage is changed only once from $\langle 0.01 - 0.06 \rangle$. The variation of segmentation error is of 15.52 (in the best case) and 21.23 (in the worst case). The execution time was meassured and the best case was obtained with 21 seconds. The sets dimension size (of L1 and L2 sets) was stated with $\langle 3,3 \rangle$. The max dimension must be the max dimension of the sets plus one, so, this parameter is increased in 4. So, the best case result in the Test Number 4 where the lowest segmentation error was of 15.52% with 0.01 of noise level, 0.2 of outliers percentage, the execution time is of 1 minute with 16

Table 2: Evaluation One: Variation of the dimension.

no.	Noise Level	Out. %	Seg. Error	Exec. Time	Sets Dim.	Dim.
1	0.01	0.06	31.91%	19s	$[2,2]$	3
2	0.015	0.06	32.50%	30s	$[2,2]$	3
3	0.01	0.06	32.30%	25s	$[2,2]$	3
4	0.02	0.06	48.21%	11s	$[2,2]$	3
5	0.02	0.02	15.32%	92 s	$[3,3]$	4
6	0.02	0.06	42.48%	48 s	$[4,4]$	5
7	0.02	0.06	48.67%	41 s	$[5,5]$	6
8	0.02	0.06	51.26%	88 s	$[3,2]$	4
9	0.02	0.06	43.77%	73 s	$[6,6]$	7
10	0.02	0.06	31.70%	20 s	$[2,2]$	3

seconds, with size dimension of the sets of $\langle 3,3\rangle$ and max dimension of 4. This results can be observed in Table 3.

Evaluation Three: Variation of Angle between Planes. This test obtains 3 important results. Where the noise level take a value of 0.01. The outliers percentage is of 0.02. The variation of segmentation error is of 15.52% (in the best case) and 20.35% (in the worst case). The execution time was meassured and the best case was obtained with 22 seconds. The sets dimension size (of L1 and L2 sets) was stated with $\langle 3,3\rangle$. The max dimension must be the max dimension of the sets plus one, so, this parameter is stated in 4. The angle between planes is stated in following values: $\pi/4, \pi/8$ and $\pi/16$. So, the best case result in the Test Number 1 where the lowest segmentation error was of 15.52% with 0.01 of noise level, 0.2 of outliers percentage, the execution time is of 1 minute with 16 seconds, with size dimension of the sets of $\langle 3,3\rangle$ and an angle between planes of $\pi/4$. This results can be observed in Table 4. The best case for three test can be visually observed in Figure 1.

The subspace arrangement has been succesfully used in image retrieval segmentation, and ordinary differential equations. The leukemia pathologies imply a deep analysis and a carfully selection or discrimination process of features extracted frome the bone marrow.

It is important to observe that singular value decomposition method and the simpson distance are relevant concepts to compute the segmentation ideal to model leukemia classification of lymphoblastic classification of L1, L2 and L3.

Table 3: Evaluation Two: Decreasing Noise Level and Outlier Percentage.

no.	Noise Level	Out. %	Seg. Error	Exec. Time	Sets Dim.	Dim.
1	0.02	0.06	21.33%	91s	$[3,3]$	4
2	0.01	0.05	19.21%	67s	$[3,3]$	4
3	0.005	0.02	20.35%	21s	$[3,3]$	4
4	0.01	0.02	15.52%	76s	$[3,3]$	4
5	0.005	0.05	19.69%	92s	$[3,3]$	4
6	0.01	0.01	19.57%	89s	$[3,3]$	4

Table 4: Evaluation Three: Variation of Angle between Planes.

no.	Noise Level	Out. %	Seg. Error	Exec. Time	Sets Dim.	Ang. Var.
1	0.01	0.02	15.52%	76s	$[3,3]$	$\pi/4$
2	0.01	0.02	20.35%	22s	$[3,3]$	$\pi/8$
3	0.01	0.02	19.58%	35s	$[3,3]$	$\pi/16$

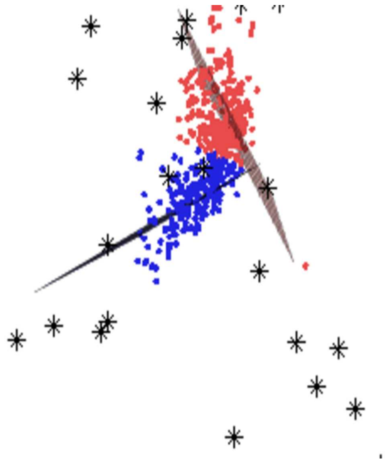


Fig. 1: GPCA final segmentation for L1 and L2 categories.

References

[1] Niegi, Statistics about the Cancer in the World, INEGI, México, (2012).

[2] J. González, Intelligent Data Analysis, Vision and Data Mining Strategy, (2011).

[3] P. Sampson, Computer Vision, Graphics and Image Processing, (1982).

[4] C. Reta, L. Altamirano, J.A. González Advances in Experimental Medicine and Biology (AEMB), (2011).

[5] B. Kumar, 14th International Conference on Digital Signal Processing. IEEE, (2002).

[6] S. Colantonio, LNCS, (2007).

[7] N. Theera-Umpon, LNCS, (2007).

[8] P. Sampson, SIAM Review, (2008).

[9] R. Vidal, Transactions on Pattern Analysis and Machine Intelligence, (2005).

[10] N. Halko, SIAM Review, (2011).

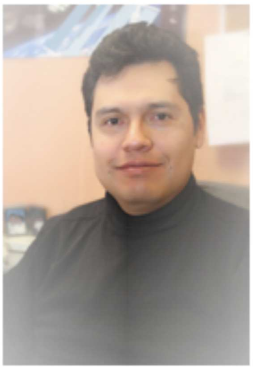
[11] L. Parsons, Sigkdd Exploration, (2004).

[12] C. Aggarwal, In Proceedings of the 2001 ACM SIGMOD international conference on Management of data ACM, (2001).

[13] R. Aggarwal, In Proceedings of the 2001 ACM SIGMOD international conference on Management of data ACM, (1998).

Leticia Flores Pulido received the Ph. D. degree in Computer Science in Universidad de las Américas Puebla and her Master Degree in National Institute of Astrophysics, Optics and Electronic (INAOE) in Computer Science. Her research interests are in the areas of mathematical modelling including visual information retrieval systems, image processing, wavelet transform, radial basis functions and pattern recognition. She has published research articles in international conferences of computer sciences.

G. Rodríguez Gómez Bachelor's degree in mathematics from the Faculty of Sciences UNAM, 1969-1973 generation. Full time Profesor in UNAM Science Faculty form 1974 to 1980. He worked inElectric Research Institute in Simulation Area Simulacin inside the Project titled D̈esign, Construction and Execution of Thermoelectric Simulation for Operators Training, as the first simulator of this type in Latinoametica. He has worked in Honeywell, Simex, CFE-México. In 1998 he gets his Master Degree in Mathematics in UNAM, in July of 2002 his Ph.D. in computer Science inside the National Institute in Astrophysics, Optics and Electronics (INAOE). He is a full time research in Computer Science Department in INAOE.



Jesus A. Gonzalez obtained his bachelor degree in Computer Science and Engineering in 1992 from las Americas University, Puebla. In 1999 and 2001 respectively, he got his Masters degree and Ph. D. in Computer Science and Engineering from the University of Texas at Arlington, Texas. He is currently enrolled in the Computer Science deparment at the Institute of Astrophysics, Optics, and Electronics (INAOE), Puebla.