Modelling Data Segmentation for Image Retrieval Systems

Leticia Flores-Pulido[1,2], Oleg Starostenko[1], Gustavo Rodríguez-Gómez[3] and Vicente Alarcón-Aquino[1]

[1] Universidad de las Américas Puebla, Puebla, C.P. 72820, México
[2] Universidad Autonoma de Tlaxcala, Apizaco, C.P. 90300, México
[3] Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, C.P. 72840, México
aicitel.flores@gmail.com, oleg.starostenko@udlap.mx, grodrig@inaoep.mx,
vicente.alarcon@udlap.mx

Abstract. The analysis of a large amount of data with complex structures is a challenging task in engineering and scientific applications. The data segmentation task usually involves either probabilistic or statistical approaches, however, global minima disadvantage is still present in each mentioned approaches. A combination of these two approaches is based on subspace arrangements avoiding global minima in classification methods. In this paper we propose a new approach for a generalized principal component analysis algorithm (GPCA) improving knowledge representation in images. The linear algebra concepts achieve an abstraction of data sets whose items as images, documents or stellar spectra could be handled providing a knowledge description for image classification process of involved data. We describe a solution to optimization GPCA function using Gutmann Algorithm for segmentation data sets.

1 Introduction

The best classification methods for Visual Image Retrieval systems have low performance when image set are noised or not homogeneous. Knowledge representation could imply definition of geometric dispersion, statistical analysis, abstract representation of data sets which improves segmentation data and classification results. The statistical methods improves classification ability, however they imply initialization parameters that lead to low performance and incorrect clustering. The probabilistic methods do not require initialization and provide significant advantages in discrimination process for segmentation of data sets. Probabilistic methods combined with statistical treatment allow an abstract representation that increases such reliability of classifications tasks as the classification of image data sets depending on knowledge representation parameters. In this work a GPCA algorithm (Vidal, 2004) requires an adaptive function improved with Gutmann algorithm (Regis et al. 2007) in classification tasks is described. The main problem goal is to decrease the error of data sets classification. Our proposal consist in hybrid linear model implementation for image data sets segmentation combining them best features of statistical and probabilistic approaches in order to overcome local minima problem presented in computational methods such as neural networks, EM, PCA, ICA, and others (Ma et al., 2008). We propose an alternative to stochastic methods for optimization of iteration process of GPCA algorithm. We purpose improving the GPCA algorithm for segmentation of data sets with high dimensionality (Section 3). A stochastic method can be used in cases phase, however a deterministic method also can be implemented GPCA optimizing (Section 5). The Gutmann algorithm allows optimization and adaptating process for objective functions with less complexity than a stochastic method.

2 Segmentation Data Problem

A typical learning uses statistical or probabilistical data analysis. Huge collections handled mixed data that are typically modelled as a a group of samples $\{s_1, s_2, s_n\} \subset \mathbb{R}^n$ are obtained from a learning approach with some probabilistic distribution. Each one of the samples has a domain of values and they are composed of a set of vectors. Every vector can be modelled as a singular value array that describes relevant knowledge about the features of a sample as an image, a stellar spectrum, or a document. These features represents knowledge content about image data set. The main problem here is the implementation of some approaches that not only avoid local minima disadvantages, but permit to use a hybrid approach as GPCA. GPCA can be improved in its optimization process trough not only evolution strategies but other kind of approaches as deterministic methods. One of the deterministic method employed for reduction of iterations is based on the Gutmann algorithm. Gutmann algorithm (Regis et al. 2007) allows a cheaper iterative function. We propose to design a mathematical abstraction for data segmentation, using GPCA algorithm by adding an optimization phase. GPCA makes calls to objective function, but if dimensionality of data increase, its optimization process has not so well performed. Iteration process in GPCA for segmentation data is computationally intractable when data dimensionality increase concluding propose an alternative approach to overcome former disadvantages with Gutmann implementation algorithm inside GPCA. Gutmann algorithm allows successfully global minima search and guarantee decreasing number of iterations comparing with stochastic methods like genetic algorithms, evolution strategies, or tabu search. Also, Gutmann algorithm encourage high data dimensionality analysis when iteration process are computationally intractable.

3 Modelling Data Segmentation

The general outline of our proposal titled *Modelling Data Segmentation* (MDS) is shown in Figure 1. The proposal is divided in to three levels: The abstraction level that describes an improving of generalized principal components analysis algorithm (GPCA) for data segmentation. The abstraction level is related to procedures such as *rings of polynomials*, *veronese maps*, and *vanishing ideals*

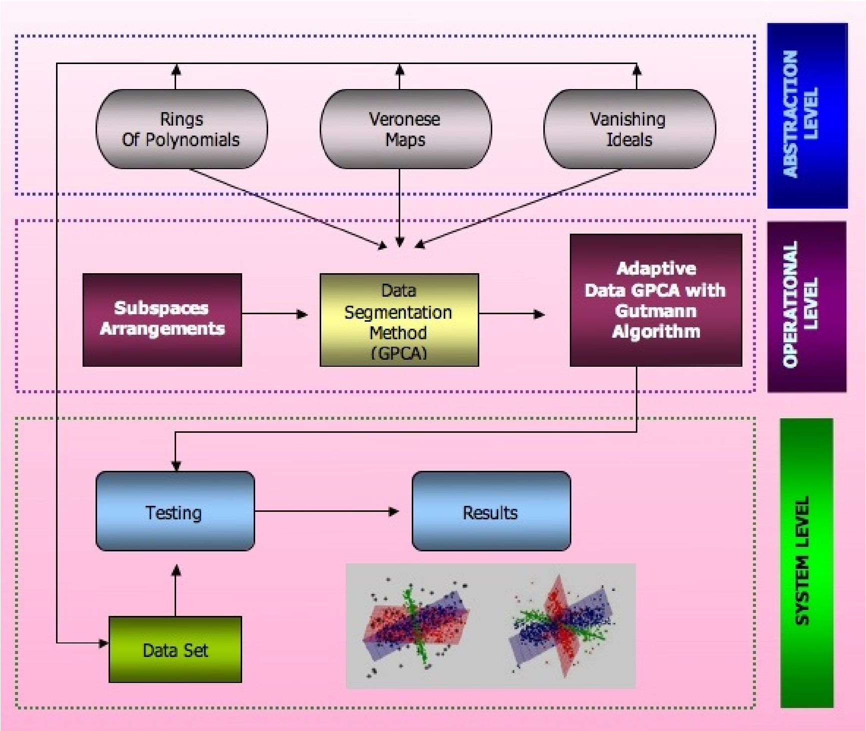


Fig. 1. General outline of modelling data segmentation for image retrieval systems. In the top of the figure the abstraction level, in the middle of the figure the operational level, and in the bottom of the figure the level system.

for design an alternative approach based on subspace arrangements. The operational level is composed by *subspace arangements, data segmentation method, and GPCA improving procedures* at this level the new contributions will be implemented. Finally, the system level consists of experimentation with input data sets, testing and evaluation of the model, including analysis of classification results of subspace arrangements with the method proposed.

4 State of Art

Several works about segmentation methods for subspace arrangements of data sets have been developed. Subspace arrangements can be computed using vanishing ideal in polynomial rings. Some of the works that describe this process are (Bjorner, 2005), (Derksen, 2005), and (Vidal, 2004). The concept of outlier in subspace arrangements refers to elements that lie out of some well classified data group. Some research papers written by (Sugaya, 2003) and (Yang, 2006) help to detect and to avoid outliers in subspace arrangements. Other models for detection of subspaces are described in the following papers: (Kanatani, 2001), (Kanatani, 2004) and (Roweis 2000) as alternatives for well known approaches for subspace arrangements. The relevant related works that describe GPCA improvement are (Huang, 2004), (Ma, 2008), (Ma, 2005), (Rao, 2005) and (Vidal, 2005).

5 Methodology

The adopted methodology computes subspace arrangements from knowledge representation of visual data sets, representing images or multimedia documents frequently used in engineering and science applications. Our proposal imply mathematical treatment of data using extension of GPCA algorithm (Ma et al., 2008). The novelty of our proposal is detailed in Figure 2.

GPCA algorithm describes subspace arrangements with special features. Hilbert function is a special kind of subspace arrangement that represents a data set with particular invariances (Lang, 2002) and (Bjorner, 2005). Radial Basis Function is the core of Gutmann algorithm (Regis et al. 2007) through which GPCA improvement could be achieved. In this section the mentioned previously approaches are presented by formal definitions as it follows.

Definition 1. (Subspace arrangement). *A subspace arrangement in* $\mathbb{F}^D$ *is the union*

$$\mathcal{A} \doteq V_1 \cup V_2 \cup ... \cup V_n. \tag{1}$$

of n subspaces $V_1, V_2, ..., V_n$ of $\mathbb{F}^D$.

Definition 2. (Radial Basis Function Interpolation). *Given* n *distinct points* $x_1, ..., x_n \in \mathbb{R}^D$ *where the function values* $f(x_1, ..., x_n)$ *are known, we use an interpolant of the form*

$$s_n(x) = \sum_{i=1}^{n} \lambda_i \phi(\| x - x_i \|) + p(x), x \in \mathbb{R}^d \tag{2}$$

where $\| \, . \, \|$ is the Euclidean norm, $\lambda_i \in \mathbb{R}^D$ for $i = 1, ..., n, p \in \Pi_n^d$ (the linear space of polynomials in d variables of degree less than or equal to m), and ϕ is a real valued function that can take many forms. Gutmann algorithm (Regis et al. 2007) works better with $\phi(r) = r^2 logr, r > 0$ and $\phi(0) = 0$.

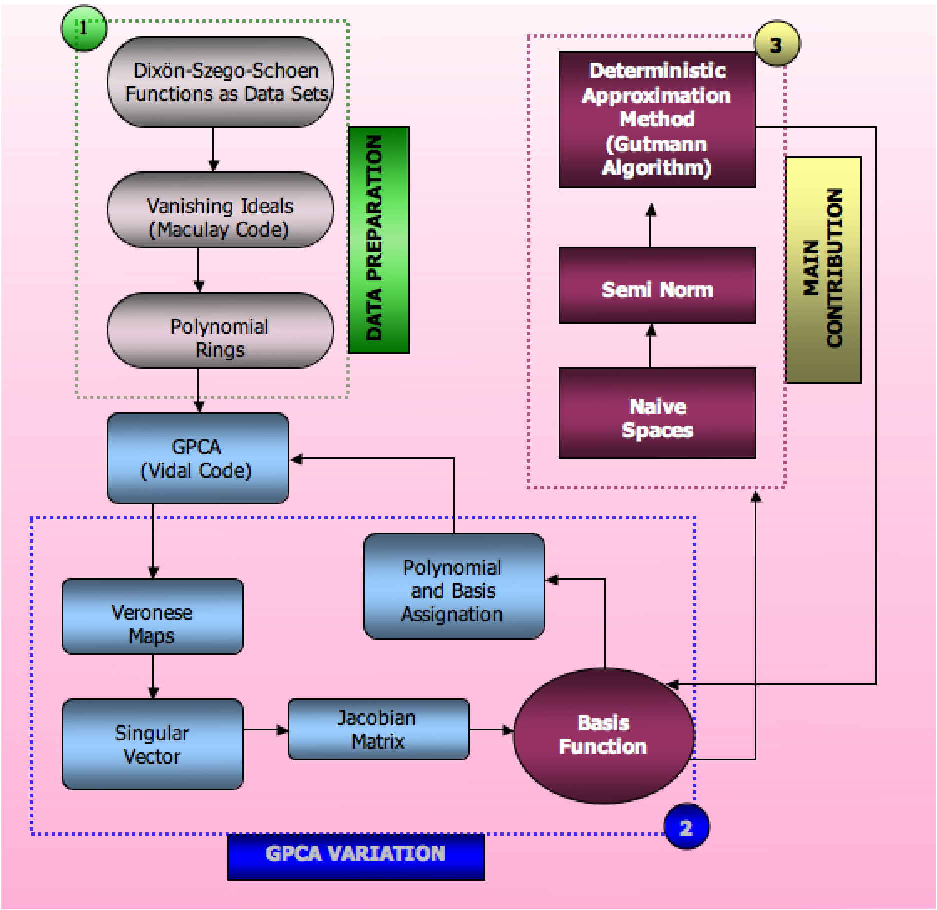


Fig. 2. Variation of Modelling Data with GPCA adding Gutmann Algorithm for VIR systems. It is observed that the first phase is about data preparation, second phase contains GPCA variation and third phase is about Gutmann algorithm implementation as the main contribution of our work.

6 Experimentation

This research implies the first step of generation of data sets and data functions. The data sets can be generated randomly, and they need a definition of vector size and a number of dimensions. This can be done with `haltonseq.m` code written by Daniel Dougherty (MCFE link). An alternative for random data sets generation are Schoen functions (Schoen, 1992) useful in global optimization problems with special features such as: easy to built for any dimension, global minimum and maximum known *a priori*, controllable smoothness as well as number and location of stationary points. After generation of random points it is important to extract vanishing ideals (Ma et al. 2008) from data sets. The vanishing ideals can be obtained with by applying MACULAY routine (Grayson, 1997). One example of its use is shown in Figure 3(a). Vanishing ideals provides information about number and dimension of subspace arrangements of segmented data sets. Once vanishing ideals has been detected, we will accomplish experimentations with original code of Kanatani used for subspace detection *http://perception.csl.uiuc.edu/gpca* (Kanatani, 2002). This code has several tests for global optimization problem using generalized principal component analysis applied to data segmentation. We will use this code for applications with image data sets. Figure 3(b) shows an example of GPCA algorithm applied to data sets.

7 Conclusions

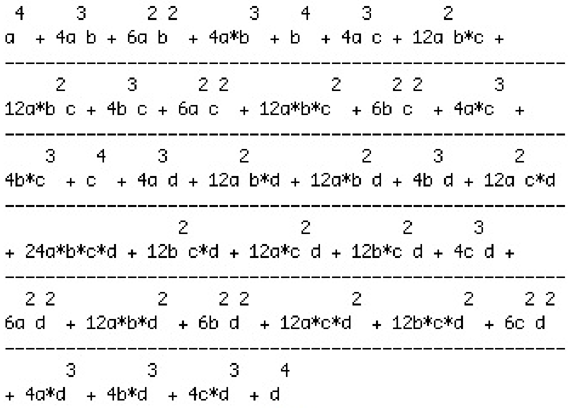
We have exposed a novel approach for modelling data segmentation. We purpose GPCA for extracting knowledge representation from image data sets. Our method allows GPCA improvement to high dimensional data classification. The methodology involves vanishing ideals computation, and subspace detection. Improve GPCA is stated in Gutmann Algorithm useful for high dimensional data sets. Gutmann Algorithm works with radial basis function interpolation. We have been explained the methodology that improves in knowledge representation methods for image data sets.

References

[Bjorner, 2005] A. Bjorner, I. Peeva, and J. Sidman, Subspace arrangements defined by products of linear forms, J. London Math. Soc. (2) 71 (2005) 273–288.

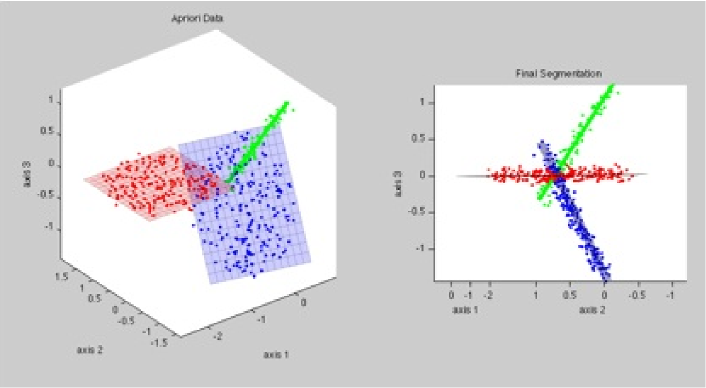
[Derksen, 2005] H. Derksen: Hilbert Series of Subspace Arrangements, preprint, arXiv.org, 2005; available online from *http://arxiv.org/abs/math/0510584*.

[Vidal, 2004] R. Vidal, Y. Ma, and J. Piazzi, A new GPCA algorithm for clustering subspaces by fitting, differentiating and dividing polynomials, in Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (2004) 510–517.



```
 4     3      2 2       3    4     3       2
a   + 4a b + 6a b   + 4a*b   + b   + 4a c + 12a b*c +
-------------------------------------------------------
     2      3      2 2          2     2 2       3
12a*b c + 4b c + 6a c  + 12a*b*c  + 6b c  + 4a*c  +
-------------------------------------------------------
    3    4     3       2           2      3       2
4b*c  + c  + 4a d + 12a b*d + 12a*b d + 4b d + 12a c*d
-------------------------------------------------------
                 2           2         2      3
+ 24a*b*c*d + 12b c*d + 12a*c d + 12b*c d + 4c d +
-------------------------------------------------------
  2 2          2     2 2          2          2     2 2
6a d  + 12a*b*d  + 6b d  + 12a*c*d  + 12b*c*d  + 6c d
-------------------------------------------------------
      3       3       3    4
+ 4a*d  + 4b*d  + 4c*d  + d
```

(a)



(b)

Fig. 3. (a) Vanishing Ideals for random points, an example taked from MACULAY code with vanishing ideals extracted from a dense data set like halton points. (b) Subspaces detected by vanishing ideals, a simulation taken from (Vidal, 2004) with 3 subspace arrangements.

[Sugaya, 2003] Y. Sugaya and K. Kanatani, Outlier removal for motion tracking by subspace separation, IEICE Trans. Inform. Systems E86-D (2003) 1095–1102.

[Yang, 2006] A. Yang, S. Rao, and Y. Ma, Robust statistical estimation and segmentation of multiple subspaces, in Workshop on 25 years of RANSAC IEEE International Conference on Computer Vision and Pattern Recognition (2006) 99.

[Kanatani, 2001] K. Kanatani, Motion Segmentation by Subspace Separation and Model Selection, The 8th Internacional Conference in Computer Vision July (2001) Vancouver Canada (2) 586-591.

[Kanatani, 2004] K. Kanatani, For geometric inference from images, what kind of statistical model is necessary? Systems and Computers in Japan (35) 6 (2004) 1-9.

[Roweis, 2000] S. Roweis and L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.

[Huang, 2004] K. Huang, A. Wagner, Y. Ma: Identification of hybrid linear time-invariant systems via subspace embedding and segmentation, in Proceedings of the IEEE Conference on Decision and Control 3 (2004) 3227–3234.

[Ma, 2005] Y. Ma and R. Vidal, A closed form solution to the identification of hybrid ARX models via the identification of algebraic varieties, in Proceedings of the International Conference on Hybrid Systems Computation and Control (2005) 449–465.

[Rao, 2005] S. Rao, A. Yang, A. Wagner, and Y. Ma, Segmentation of hybrid motions via hybrid quadratic surface analysis, in Proceedings of IEEE International Conference on Computer Vision (2005) 2–9.

[Vidal, 2005] R. Vidal, Y. Ma, and S. Sastry: Generalized principal component analysis (GPCA), IEEE Trans. Pattern Anal. Machine Intelligence 27 (2005) 1–15.

[Ma et al. 2008] Y. Ma, A. Y. Yang, D. Harm, R. Fossum: Estimation of Subspace Arrangements with Applications in Modelling and Segmenting Mixed Data, SIAM Review Society for industrial and applied mathematics, Vol 50 Num 3 413–458 (2008)

[Regis et al. 2007] R. G. Regis , Ch. A. Shoemaker: Improved strategies for radial basis function methods for global optimization, J Glob Optim (2007) 37:113–135, DOI *10.1007/s10898-006-9040-1*

[Lang, 2002] S. Lang: Algebra, Springer-Verlag, New York, 2002.

[Kanatani, 2002] K. Kanatani, Motion segmentation by subspace separation: Model selection and reliability evaluation, Int. J. Image Graphics, 2 (2002), 179–197.

[Gutmann, 2001] H.M. Gutmann, A Radial Basis Function Method for Global Optimization, Journal of Global Optimization 19: 201–227, (2001).

[MCFE link] Matlab Central File Exchange, url: *http:/www.mathworks.com/matlabcen- tral/fileexchange/*.

[Schoen, 1992] F. Schoen, A wide class of test functions for global optimization, Journal of Global Optimization, 3: 133–137, 1993.

[Grayson, 1993] D. Grayson, M. Stillman, Macaulay 2– a system for computation in algebraic geometry and commutative algebra, *http:/www.math.uiuc.edu/Macaulay2*, 1997.