Pattern Recognition 48 (2015) 1185–1196

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Pattern Recognition in Latin America in the "Big Data" Era ☆

Alicia Fernández [a], Álvaro Gómez [a], Federico Lecumberry [a,*], Álvaro Pardo [b], Ignacio Ramírez [a]

[a] Instituto de Ingeniería Eléctrica, Facultad de Ingeniería, Universidad de la República, J. Herrera y Reissig 565, Montevideo, Uruguay
[b] Department of Electrical Engineering, Universidad Católica del Uruguay, Av. 8 de octubre 2801 Montevideo, Uruguay

ARTICLE INFO

*Article history:*
Received 2 January 2014
Received in revised form
20 March 2014
Accepted 9 April 2014
Available online 18 April 2014

*Keywords:*
Pattern Recognition
Machine learning
Data mining
Image analysis
Computer vision
Large scale
High dimension
Big Data
Latin America

ABSTRACT

The "Big Data" era has arisen, driven by the increasing availability of data from multiple sources such as social media, online transactions, network sensors or mobile devices. This is currently a focus of interest among public and private organizations, governments, research institutes and companies operating in diverse fields as health, security, commercial recommendations, detection of anomalies and future trends among others. In this problem, the main objective is to recognize and extract meaningful information (patterns, structure, underlying relationships, etc.) from huge amounts of heterogeneous data. This task is complicated by new, significant storage and processing requirements due to unprecedented volumes of data. In this scenario, new algorithms in Pattern Recognition and related fields are being devised, while well known techniques are revisited and adapted to these new challenges. Latin American research in the "Big Data" problem is still incipient, but there is a significant body of recent works in the subjects of Pattern Recognition and related fields that indirectly addresses the problem. This paper reviews Latin American contributions in Pattern Recognition and related fields in the last lustrum. The focus is set on—but not restricted to—applications in the fields of Computer Vision and Image Analysis with large scale characteristics.



## 1. Introduction

The most accepted definition of "Big Data" goes back to [87], which defines the "three Vs" (Volume, Velocity and Variety) that characterize the new flows of data. These "three Vs" imply dealing with large, heterogeneous and rapidly and continuously flowing datasets comprising huge number of samples and features. Many other definitions of this concept have also appeared in recent years.[1]

Although the term has been around for over a decade, the results of web analytics on the term "Big Data" show an "explosion" in its use in 2012, evidencing the increasing interest of public and private organizations driven by the availability of data from multiple sources (social media, online transactions, sensor networks, mobile devices, etc.), and of new technologies for storing and manipulating them. This trend is confirmed at an academic level where, for example, a search using the Scopus[2] database also shows an impressive increase in works related to the "Big Data" keyword since 2012 (see Fig. 1).

However, "Big Data" is not only related to having more data. Most of this huge amount of data is related to "personal information" and some analysts believe that this will change our lives in many ways for good and for bad [143]. It is an era of great opportunities for research and development. However, these should not be conducted at the expense of loss in security and privacy of personal data.

From the point of view of the data analysis practitioner, the "Big Data" scenario implies a number of technical challenges. The most obvious of them is the computational issue of having to deal with very large data structures in memory (storage complexity), and to do so in a reasonable time (computational complexity). For example, the computational complexity of most algorithms depends super-linearly on the number of samples and/or the dimension of the samples, so that increasing any of those variables requires more than just optimizing the code; it calls for practical, approximate solutions, massive parallelization and also new algorithms.

But there are other, more fundamental issues with the aforementioned scenario. Large feature vectors summon the ever dreaded *curse of dimensionality*. Heterogeneous features require creative ways to perform inference in complex data spaces. In many cases (typically, data mining), little or no prior information is known about the data to be explored, so that only very general

☆This document is a collaborative effort.

* Corresponding author.
*E-mail addresses:* alicia@fing.edu.uy (A. Fernández), agomez@fing.edu.uy (Á. Gómez), fefo@fing.edu.uy (F. Lecumberry), apardo@ucu.edu.uy (Á. Pardo), nacho@fing.edu.uy (I. Ramírez).

[1] See for example http://www.opentracker.net/article/definitions-big-data.

[2] http://www.scopus.com

http://dx.doi.org/10.1016/j.patcog.2014.04.012
0031-3203/

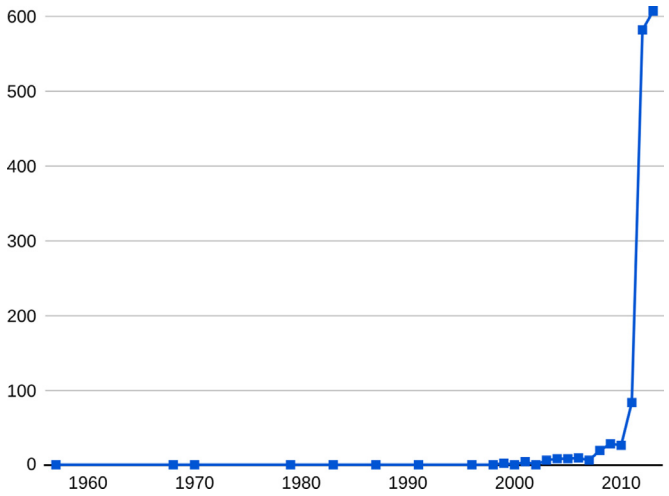


**Fig. 1.** Number of publications per year with the keyword "Big Data" in either the title, abstract or keywords sections, as obtained from a search in the Scopus database (September 2013). This graph illustrates the dramatic increase experienced since 2012.

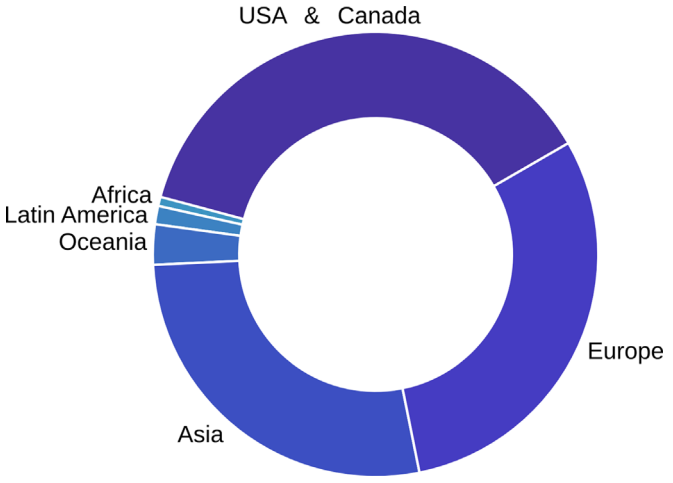


**Fig. 2.** Distribution of publications per continent with the keyword "Big Data" in title, abstract or keywords section in the Scopus database (September 2013). The contribution of Latin America is about 1.5%.

assumptions can be made about them. In such cases, sophisticated methods are often needed to determine whether and up to what point a given method can discover any structure within the data.

On the good side, when reasonably accurate models can be assumed about the data generating process, the large number of samples often implies that analytical formulas for the asymptotic behavior of certain learning methods can be accurate enough, thus simplifying and speeding-up, for example, model selection procedures.

In summary, the new "Big Data" paradigm has opened up an opportunity to expand and enrich Pattern Recognition and related fields. One of the objectives of this work is to understand whether Latin America is up to this challenge, and how much of this opportunity can be seized by its research community.

Considering this, the paper explores the contributions on Pattern Recognition and related fields in Latin America connected, in a wide sense, with the so-called "Big Data" problem. The focus is set on applications in the fields of Computer Vision and Image Analysis with large scale characteristics.

Following this introduction, Section 2 introduces the methodology applied to conduct the bibliographic search. Section 3 presents a relevant sample of the documents found in the aforementioned search, organized by application area. In Section 4, some statistics are shown with the objective of identifying areas of research and collaboration relationships among countries. Concluding remarks are presented in Section 5.

## 2. Methodology

Fig. 2 shows the distribution of the publications retrieved with the keyword "Big Data" in the title, abstract or keywords section by author affiliation country in the Scopus database. The figure shows that, for this specific keyword in the search, the contribution of Latin America is about 1.5%. However, even if the keyword in question is not explicitly mentioned, there are several recent publications in Latin America that, indirectly, address the "Big Data" problem through the various subjects treated therein, for example large scale datasets, and high dimensionality.

In order to identify relevant articles in the focused area, a wide search was performed on the Scopus database. Scopus is a large abstract and citation database that comprises most of the peer-reviewed literature of the areas of interest for this paper. An initial search was performed with the objective of getting most of the works related to applications involving "Big Data" in a wide sense in Latin America since 2008. In this search, only documents where one or more of the keywords "big data", "huge data", "large data", "high dimension", "dimensionality reduction" and "high scale" were present in either their title, abstract or keywords section were selected. From the result of the initial search, new keywords were identified and added to the search criteria: "large data", "higher dimensions", "dimensionality reduction", "large datasets", "high dimensions", "data mining", "large data sets", "data reduction", "high dimensional data", "information retrieval", "image retrieval", "very large datum", "high-dimensional", "big datum". Also, only publications with at least one author with affiliation country in Latin America were considered. The final search yielded 3724 documents (September 2013). Statistics were extracted from this search in order to identify collaboration relationships between countries in the region and with the rest of the world (see Section 4).

These 3724 documents were manually examined, and relevant examples (covering the areas which of interest of this work) were selected from this search; these are presented in Section 3. More specifically, those works on Pattern Recognition and related fields that were more related to "Big Data" and large scale problems were selected. Also, whenever a series of closely related works, along the same line of research, within the same research group was found, only the most relevant one or two were referenced.

At this point, we would like to note that, despite our effort to perform a search as wide as possible, the aforementioned search procedure may have excluded relevant publications. This being said, we believe that the documents retrieved can be considered as a fair sample of the areas of interest of this work.

## 3. Applications

This section presents relevant examples of the publications found in the search described in Section 2. The publications are grouped into four major application areas (Remote Sensing, Multimedia, Biometrics and Biology/Biomedical) and a fifth category that deals with generally applicable Pattern Recognition and Machine Learning methods and theoretical developments.

### 3.1. Remote sensing

The term remote sensing refers to the task of obtaining information from distant sources via some sort of signal acquisition. Taken literally,