Techniques Used for Geospatial Big Data Collection, Storage and Analysis

Liviu PORUMB, Andreea Florina JOCEA, Alexandru GRIVEI, Lucian NECULA and Dan RĂDUCANU

***Abstract*—Nowadays data are generated by different sources, at an incredible rate, and the traditional approaches for their collection, storage and analysis are not suitable. Big Data is analyzed and used by state institutions, business environment, transportation, health, communications, banking system, utilities, defense and other components of modern society in order to support their decisions as well as the human activities. Geospatial data is an important component of Big Data and aerial/satellite images offer a lot of details about the environment, events and their evolution in time. This paper presents the significant techniques used in three main stages of Geospatial Big Data lifecycle, namely data collection, storage and analysis. Geospatial data collections are mainly executed by using web crawlers in order to find meaningful data and, during this stage some preprocessing operations can be done (standardization, completion and integration). Cloud storage and distributed file systems are widely used for Geospatial Big Data storage and new types of non-relational and relational databases are developed. The very challenging aspects for Big Data analysis are related to feature identification and extraction from aerial or satellite images, using feature-based extraction and deep learning algorithms.**

***Index Terms*—distributed storage, feature extraction, Geospatial Big Data, web crawlers.**

I. INTRODUCTION

Technology evolution, internet and interconnected networks influences continuously our daily life. Widespread use of digital data and cyberspace has facilitated connecting a big part of the population, institutions and companies around the globe. Thus, humanity depends more and more on information and communications systems while data and information volume has increased exponentially.

Statistics indicate that a huge amount of data (2.5 quintillion bytes of data) called Big Data is generated every day [1].

Usually, this data comes from aerial and remote sensing images and from social media applications (WhatsApp, Facebook, Twitter, Instagram, etc.).

Although the term *Big Data* has been used since the 1990s, it is only in 2001 that the three main traits of this type of data were conceptualized by Doug Laney. In a comprehensive understanding, "*Big Data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information*" [2]. Beside the three main characteristics of Big Data, in a general meaning, another 4V attributes could be added to describe Geospatial Big Data: visualization, veracity, versatility and value [3].

Big Data term refers to data sets that are so large and complex that they are too difficult to work with using standard software, that can be structured or unstructured [1]. The main differences between these types of data are highlighted in Fig. 1.

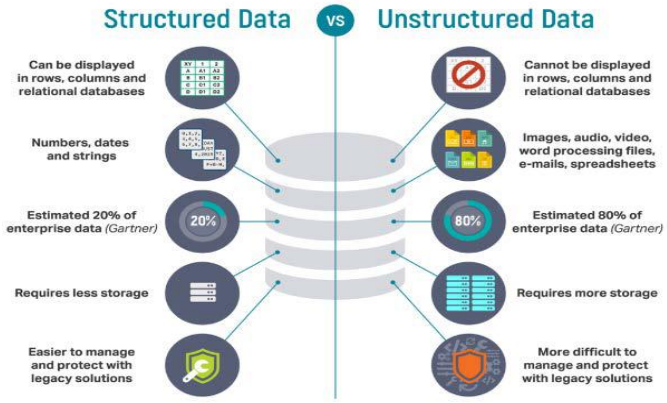

Figure 1. Structured versus unstructured data [4]

The processing of unstructured data, generated mostly by social media and the Internet, presents the biggest challenges, and Table I presents some examples of sources of large unstructured data sets.

TABLE I. UNSTRUCTURED BIG DATA SOURCES [5]

Data source	Production
Apple devices	47,000 applications/minute are downloaded
Facebook application	34,722 likes/minute are registers 100 Terabytes/day of data is uploaded
Google products	Over 2 million search queries/minute are executed
Instagram	40 million photos/day are shared by users
Twitter	More than 175 million tweets/day are generated
WordPress application	350 blogs/minute is published by bloggers

It should be noted that large volumes of data differ from classical data in terms of volume, data structure, generation rate and other factors. In order to extract characteristics or patterns there are used sets of algorithms named Big Data Analytics.

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI - UEFISCDI, project number PN-III-P2-2.1-SOL-2021-0084, within PNCDI III

L. PORUMB is lecturer at Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: liviu.porumb@mta.ro).

A.F. JOCEA is associate professor at Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: andreea.jocea@mta.ro).

A. GRIVEI is collaborating lecturer at Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: alexandru.grivei@mta.ro).

L. NECULA is lecturer at Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: lucian.necula@mta.ro).

D. RĂDUCANU is professor at Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: dan.raducanu@mta.ro).

Geospatial data mainly consists of three components: vector, grid data (i.e., elevation data) and raster/images. Even the dimension of vector data it is not comparable in size with raster/image data, there are large vector databases at global scale, proprietary or open source, like OpenStreetMap. Grid data covers specific zones, from individual countries to the whole world (for example, SRTM [6]), and comprises data evenly distributed at a specific spatial resolution. Raster data and images are by far the Big Data component of Geospatial data. Raster data comes from scanned maps and other cartographic documents while images originate from different airborne and satellite passive or active sensors. In the recent years other sources for large geospatial datasets have emerged, mobile devices, Internet of Things, social media etc.

Remote sensing technology used to collect information and images over a large area from Earth has been available for decades. For a better understanding the Earth's system, the technology is used for mapping and monitoring atmosphere, oceans, land surfaces, ecological forecasting etc. The increasing capabilities of remote sensing technology has led to improved spatial, temporal and radiometric resolution of satellite images. The processing and transformation of remote sensing data into information and applications has lead to problems due to the large amount of data.

Volume, complexity and heterogeneous nature of remote sensing data hide useful information that limits its applicability in real-world scenarios; therefore, image processing is usually done manually and according to the user's needs so that it can be done quickly and accurately. Fully autonomous processing is not yet possible, although remote sensing image analysis technologies and implicitly classification technology have evolved in recent years.

Although there are many approaches to the Big Data lifecycle, ranging from five to ten steps, the overall performed activities are the same. Fig. 2 shows a data life cycle that is more appropriate for Geospatial Big Data.

In the next sections we will refer to collection, storage and analysis of geospatial Big Data.

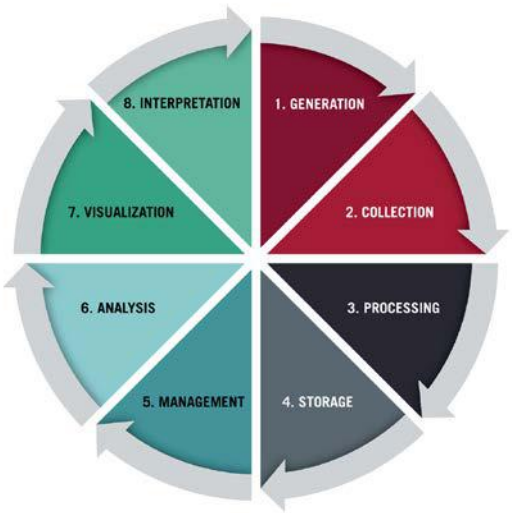

Figure 2. Data lifecycle stages [7]

II. DATA COLLECTION

Open-source data has the greatest potential for providing Big Data and is the most accessible source of information, due to the fact it is based on data available in the public space. Some of the limitations of open-source data refer to misinformation and propaganda and, in order to identify them, it is necessary to develop algorithms for initial screening. Isolating fake news requires human intervention and the use of appropriate grouping and classification algorithms.

Methodical, periodic Internet scanning is used to identify sites of interest using web crawlers to create a data index. Most search engines use web crawling as a means of providing up-to-date data and finding news on the Internet. Analytics companies and market researchers use web crawlers to determine customer and market trends in a particular geographic area. There are many web crawlers, and the most used of these are presented in [8].

A. Text data collection

The technology for collecting text data relevant to a specific activity is called text mining and is widely applied in government, research and business environments to manage records and search for documents relevant to their day-to-day activities. Many governments and military structures use text mining for national security and intelligence purposes.

Major companies and firms, including IBM and Microsoft, are interested in developing new methods and software for automated retrieval and analysis processes in order to improve results. In the public sector, for example, much effort has been focused on creating security applications, especially for monitoring and analyzing online text sources, such as Internet news, blogs etc., for national security purposes. The study of text encryption/decryption is also involved in this field. For Python programmers, there is a set of tools called NLTK (Natural Language Toolkit) for more general purposes. For more advanced programmers, there is also the Gensim library, which focuses on topic modelling, document indexing and similarity retrieval [9].

Exploiting the text involves the process of structuring the input text (analyzing it together with the addition of derived linguistic features and later inserting it into a database), obtaining models from the structured data, and finally evaluating and interpreting the output. Typical text extraction tasks include classifying text, grouping text, extracting concept/entities, producing granular taxonomies, analyzing feelings, summarizing documents, and modeling relationships between entities.

One of the most efficient solutions for collecting and processing text data is the Sintelix platform (Fig. 3). It contains the Harvester component, a customizable solution for extracting text data from multiple web sources, such as news, wikis, forums, blogs, and social networks.

Harvester identifies only relevant data, ignoring navigation, sidebars, footers, ads, and any other unwanted text, and transmits it to Sintelix to organize, associate, and build a network. Sintelix can identify individuals, organizations, geotagged locations, and twenty-four other types of data, providing meaningful content.

A single webpage can be processed at one time or more, grouped in a batch. The system extracts text as well as hyperlinks and processing can be recursive.

Sintelix Harvester component is able to recognize the entities, relationships, and properties of the extracted data and creates networks of entities that can be viewed through entity tables, link diagrams, and chronological occurrence [10].

 Digital Object Identifier 10.32754/JMT.2022.1.04

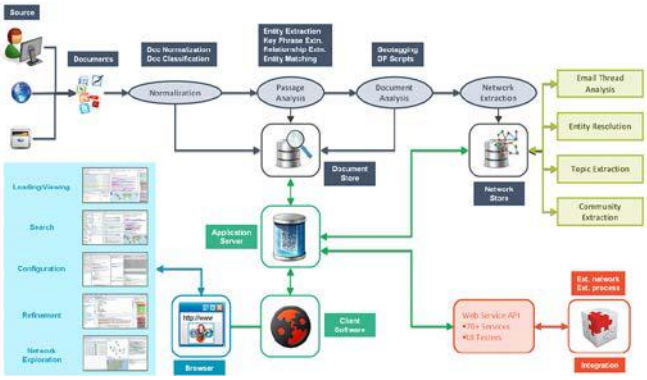

Figure 3. Sintelix platform - operational flow [10]

B. Remote sensing and airborne images collection

There are two major categories of systems that carry sensors used to collect geospatial data: satellite and airborne. Passive or active sensors, placed on these platforms, can provide panchromatic, RGB, infrared, RADAR (SAR), LiDAR, multispectral, hyperspectral and ultraspectral images, etc.

Very high-resolution optical satellite imagery can be provided based on commercial agreements from different distributors, such as Maxar (30 cm native and 15 m by resampling), Digital Globe (30 cm), Pleiades (30 cm and 50 cm) etc. The use of high-resolution RADAR images such as those provided by the TerraSAR-X/TanDEM-X and PAZ satellites (25 cm - 40 cm), may also be considered. The data provided by the Sentinel satellites, although not very high resolution (10 m), is available online for free through Copernicus platform [11]. Access to LiDAR data is not as easy, as it can be purchased from third parties if you do not have your own collection systems.

Due to their ability to stay in the air for long periods of time, Unmanned Aerial Vehicles are becoming an increasingly important source of geospatial data.

C. Geospatial information collection

Geospatial information mainly consists of vector and raster databases as well as numerical terrain models for an area of interest. Usually, vector databases are available in standard formats, to ensure interoperability and facilitate the exchange of geospatial information/data. Thus, DAFIF (Aeronautical Flight Information Files), DVOF (Digital Vertical Obstruction File), VMap1 (Vector map level 1), VMap2 (Vector map 2), MGCP (Multinational Geospatial Co-production Program), Local Topographic Datasets (LTDS) and other vector data can be obtained through agreements with commercial partners, from geospatial agencies, and from open sources (for example, Open Street Map [12]). Depending on the source of geospatial information/data, other vector data formats (ESRI shapefile, ESRI geodatabase, DXF etc.) may be used.

Raster geospatial data is usually obtained by scanning analog maps and other cartographic materials, but also by converting existing satellite and aerial images. The usual standard formats for raster data are GeoTIFF, TIFF World, IMG, JPEG2000, ESRI Grid, ADRG (Arc Digitized Raster Graphics), CADRG (Compressed ARC Derived Raster Graphics), CIB (Controlled Image Database) etc. [13]. Other elevation models for an area of interest can be stored in different formats (DEM, DTED1, DTED2, etc.) and can be obtained from producers or can be downloaded online, namely from the US Geological Survey site [14].

Other sources of geospatial data can be point coordinates databases or files (GML, JSON, CSV, etc.).

It is important for subsequent analysis that geospatial information to be as up-to-date as possible.

III. DATA STORAGE

Data storage is a constant challenge for Big Data sets due to their volume and velocity, which make it very difficult to properly estimate the required storage capacity. Therefore, unlike conventional data, storage systems for large data sets must first and foremost be easily and quickly scalable and provide protection against faults.

Although relational database management systems are used for storing structured and semi-structured large data sets, the variety of Big Data sets has inherently led to the emergence of new storage and management methods, such as cloud storage, NewSQL databases, NoSQL databases and distributed management file systems like Hadoop File Distributed System (HDFS).

In the case of *cloud storage*, physical storage extends to multiple servers in different locations and is usually owned and operated by a hosting company, which is responsible for keeping data available and accessible in a physically secure environment. Some known examples of cloud storage services are Amazon Web Services S3, Oracle Cloud Storage, Microsoft Azure Storage, EMC Atmos and EMC ECS.

NewSQL databases provide scalable performance similar to that of NoSQL systems combining the ACID properties of a traditional database management system. VoltDB, NuoDB, Clustrix, MemSQL, and TokuDB are some of the examples of NewSQL database [1].

NoSQL (Not only SQL) databases, useful for working with large sets of distributed data, are non-relational databases and were designed to store and manage unstructured and semi-structured data, organized in key: value pairs.

In addition to SQL, NoSQL databases supports use other query languages such as HQL to query structured data, XQuery to query XML files, SPARQL to query RDF data etc. The most known NoSQL database implementations are Amazon DynamoDB, Microsoft Azure Table Storage, Apache Cassandra, CouchDB, SimpleDB, and Google Bigtable [1].

The *Hadoop Distributed File System* (HDFS) is a highly fault-tolerant distributed file system designed to run on low-cost hardware. Having a master/slave architecture (Fig. 4), it has many similarities as well as significant differences with existing distributed file systems. HDFS provides high throughput access to application data and is suitable for applications that have large data sets [15].

Nowadays, HDFS is a subproject included in Apache Hadoop, an open-source framework developed by the Apache Software Foundation used to store and process large data in a distributed environment for processing simultaneously, based on simple programming model. Highly suitable for unstructured and semi-structured data, it is designed to scale up from single servers to thousands of machines and provides distributed storage and computing between computer clusters.

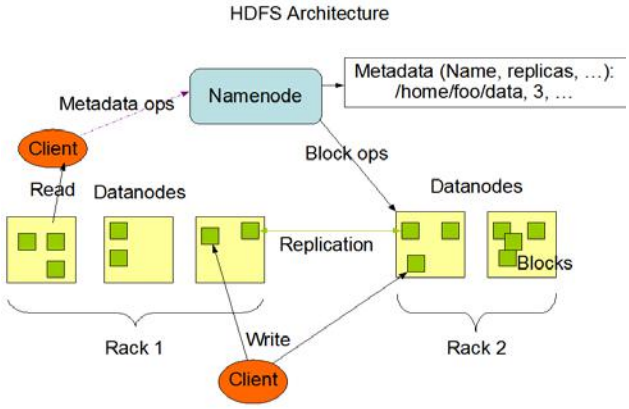

Figure 4. HDFS architecture [15]

Apache Hadoop provides high-availability service over a group of computers, due to the fact that it detects and manage application-level failures instead of counting on hardware reliability. There are four major components to Hadoop architecture: HDFS, Hadoop MapReduce, Hadoop common and Hadoop YARN [15]. Fig. 5 presents an example of big data integration.

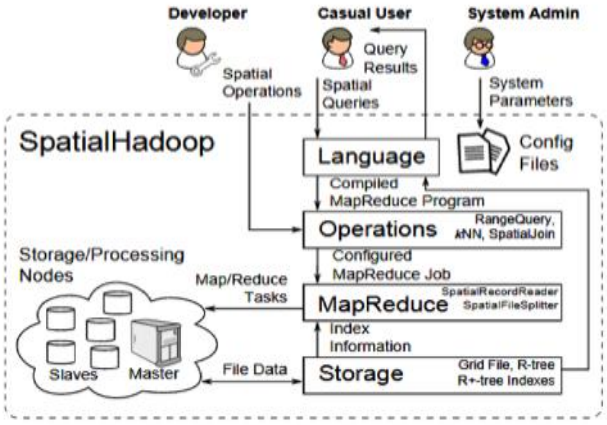

Figure 5. Big data integration with Apache Hadoop

Large datasets from different data warehouses, like Hadoop Distributed File System (HDFS), NoSQL databases and relational data stores (i.e., Apache Hive) can be processed by Apache Spark, a unified analytics engine for large-scale data processing.

It provides an optimized engine that supports general computation graphs for data analysis, high-level APIs in Java, Python, and R and offers a rich set of higher-level tools like Spark SQL for SQL, pandas API on Spark, MLlib for machine learning, GraphX for graph processing, and Structured Streaming [16].

In order to combine and integrate large datasets, usually there are used different storage and computing platforms.

IV. Data Analysis

Existing commercial and open-source geospatial software, among them being ArcGIS, Intergraph GeoMedia, QGIS, ERDAS Imagine, ENVI, Global Mapper, etc. offers different tools to import/export, store, analyze, visualize and distribute geospatial data and information. While the analysis of vector and grid data does not require sophisticated and very complex algorithms to obtain usable information, the features' identification and their extraction from images require the development of existing algorithms as well as the creation of new ones.

In order to identify image content, the oldest technologies use manual semantic labeling. The searches focus on keywords, which are often information that contains the type of sensors, acquisition mode, acquisition date, number of spectral bands, etc.

Due to the high amount of remote sensing data, researchers have started to analyze the potential of integrating deep learning algorithms such as neural networks into image searching process.

Search methods based on deep learning, used for remote sensing images, are able to extract information with great accuracy if the systems are trained with a very large number of examples.

Image search methods can generally be divided into: text-based methods - text based image retrieval (TBIR) and content - based methods image retrieval (CBIR). Of these, TBIR is relatively straightforward to operate and is widely used, however, with disadvantages such as the need to use a large amount of image data tags, ambiguity in tags, and difficulty in modifying tags. Given the limitations of TBIR, CBIR has received increasing attention in recent years and has been frequently adopted in image search applications.

CBIR can be further divided into three categories, depending on the information with which the search is performed: original data, characteristics and semantics. Processing based on raw data often measures the similarity of the pixels, and the returned images will be the most similar to the examples used. The accuracy of search methods based on raw data is high, but with low accuracy. Feature-based image search is based on the values of the visual characteristics of images. These are often hidden features of images, including information about color, texture, and shape.

The search methods focused mainly on the proper selection of image features and on appropriate methods for measuring similarity, such as QBIC developed by IBM, Photobook developed by MIT Multimedia Lab, SaFe/Visual-SEEK and webSEEK system developed by Columbia University etc. These types of search engines are suitable for searching in portions of an image. A single feature often fails to be sufficient to express the full content of the images. Therefore, some scientists use more features to get better results in the search process. Most have also applied image segmentation techniques before extracting features to obtain images with diverse content. For example, Netra, developed by the University of California, can search for similar images based on the shape and texture of the image [17].

Similar solutions are Blobworld, developed by the University of Berkeley, and SIMPLicity developed jointly by Stanford University and the University of Pennsylvania. Although these methods simulate the process by which people search for images, the accuracy largely depends on the performance of the image segmentation process and the needs of the users. With user feedback-based technologies, the image search process can be adjusted over time depending on the application. Methods for extracting and measuring similarity can be fully adjusted to reduce the gap between extracted features and associated semantics. Popular methods of image search that adopt feedback-based technology include multimedia analysis and the search system developed by the University of Illinois, MARS, and PicHunter, respectively.

 Digital Object Identifier 10.32754/JMT.2022.1.04

The development of feature-based search methods has received increasing attention. By describing the semantic features of the image (for example, scenes, behaviors, objects, and object relationships, etc.), efficient and complete image search patterns can be established. For example, machine learning methods (Bayesian, SVM, and various deep learning models) can obtain semantic information for images by combining fundamental image features and keywords to automatically label images.

A. Feature-based search methods

Feature-based information retrieval (CBIR) is a mechanism that depends on the content (visual features) of the input images [18]. Since the first occurrence of CBIR, research has focused primarily on image exploration, feature selection, and feature combinations. Combined with similarity measurement algorithms, CBIR can improve the efficiency of the remote sensing image search process. The features used in the process include, but are not limited to: spectral characteristics, texture characteristics, and image shape characteristics.

The spectral characteristic is the information contained in the reflection of ground objects. It is very sensitive to changes in lighting conditions, which means that noises often occur during the purchase process. For example, [19] proposed a search method based on points and areas used as examples of the area of interest.

Texture features can be understood as patterns revealed by a combination of pixels, which do not depend on the brightness or color of the image. In most cases, several texture features can be combined. Commonly used texture characteristics are obtained using Gabor algorithms, wavelet, gray level co-occurrence matrix (GLCM), local binary patterns (LBP) and Tamura texture. Another approach is to use the Gabor wavelet to construct several filters with the same direction but with different scales to extract information about the difference between the bands. The researchers also explored improving texture features using random fields Gaussian Markov (PS-GMRF) [20] and Gabor Opponent Colors (CGOT) [21]. Combining traditional texture features with enhanced texture features can further improve the efficiency of the image search process.

Shape characteristics can be obtained after image segmentation and edge extraction. These are also popular steps for detecting and recognizing the target in images. Image search based on shape characteristics is based on similarity measures that describe shape characteristics, which can be further divided into border-based and region - based extraction. Fourier analysis and moment invariance proved to generate the best results [22].

However, the complexity of remote sensing images and the ubiquitous overlapping of objects often leads to unsatisfactory search results when retrieval methods based on traditional features (spectrum, texture and shape) are used. In addition, local point of interest features such as Scale Invariant Feature Transform (SIFT), HOG feature, SURF - Speeded Up Robust Features) are also used to automate the image search process.

Fundamental features such as spectrum, texture, and shape can rarely fully describe the content of the image. The semantic information of the image content is usually represented by the mid-level coded features. Compared to low-level features, mid-level features incorporate original descriptors into a space representative of a visual vocabulary. After encoding the features in the vocabulary space, the semantic information of the remote sensing images is obtained by counting the information about their spatial distribution. Compared to low-level features, mid-level features are invariant to the differences caused by changes in scale, rotation, and lighting.

The semantic feature of remote sensing images is generally to encode the description of local features. Commonly used word encoding methods include Bag of Words (BoW) [23], Fisher Vector (FV) [24] and Vector of Locally Aggregated Descriptors (VLAD) [25]. Of all these encoding methods, BoW is the most widely used in applications. BoW uses the K-Means group to build word vocabulary and incorporates key features.

B. Methods based on deep learning

With multiple hidden layers, deep learning algorithms apply multiple nonlinear transformations to perform multiple levels of input data abstraction to progressively extract higher-level features. Today, convolutional neural networks, a popular approach to deep learning, provide solutions for a variety of tasks that include image recognition, speech recognition, and natural language processing [26]. In terms of feature extraction, it is necessary to design the feature space based on professional knowledge in a particular field. Deep learning plays an important role in extracting characteristics. For urban planning and construction, [27] proposed Building Residual Refinement Network (BRRNet), a new deep learning network whose goal is to extract buildings from images. BRRNet has maintained the integrity of buildings and improved the accuracy of extracting buildings with complex shapes. Deep learning has also been applied in the field of urban road planning. For example, [28] proposed Multitask Road-related Extraction Network (MRENet), used to extract the road surface and the road axis with a very good classification accuracy. For higher resolution images, such as UAV images, [29] proposed the Global Density Fusion Convolutional Network (GDF-Net) for target detection in UAV images.

In terms of image search, [30] explored the application of deep learning methods for characteristics representation and similarity measurement. There are several methods in the literature for image search based on deep learning: (1) extracting features from the convolution layer (conversion) or the complete connection layer through CNN; (2) fine-tuning the existing convolution layer; neural network using domain-related data sets; (3) the development of a convolutional neural network-specific architecture and the use of large-scale remote sensing data sets for training.

When using fine-tuned models, there is always a certain level of uncertainty in transferability and adaptability of learning model, which, to some extent, restricts image retrieval performance. To solve the problem of high dimensionality in the image search process, two strategies can be adopted: 1) improvement of search method or 2) reducing the size of the descriptors. The first strategy can be achieved by implementing data partitioning algorithms. These algorithms recursively divide the characteristic space into subspaces and record these partitions through a tree

structure. The tree-based partitioning method can improve search speed [31]. Therefore, it is imperative to reduce the size of feature vectors for the process of searching for remote sensing images in very large volumes of data.

Hashing algorithms are widely used for image storage and retrieval due to low storage volume and high query speed. Moreover, with the involvement of users, supervised hash-based search methods often outperform unsupervised methods. The purpose of the hash code-based learning method is to associate High-Dimension Feature Vectors (HDFV) with Low Dimension Binary Feature Vectors (LDBFV). Therefore, compared to HDFV, the complexity of exhaustive searches using LDBFV is significantly reduced.

Some researchers have used high-level semantic features, extracted by deep learning methods, to replace key features extracted by hand, by combining the hash-based algorithm with deep learning. To take full advantage of the combination of deep learning and hashing algorithms, Deep Hashing Neural Networks (DHNN) [32] has been proposed in the field of computer image processing, achieving satisfactory image search performance in a wide range of applications. In general, remote sensing images are different from natural images in terms of spectrum and spatial resolution. Due to these differences, DHNN trained on natural image data sets could not be applied directly in remote sensing image search applications. Thus, DHNN's construction and training of models dedicated to the classification of remote sensing images requires experimentation and exploration. For example, [33] proposed a method of searching for DHNN-based remote sensing images, but the training was done with a small number of examples.

V. CONCLUSIONS

A very important component of Big Data is represented by Geospatial Big Data. In this paper, we focused on more important aspects regarding the types of Geospatial Big Data and the methods used for their collection, storage and analysis.

The location associated with geospatial data could be exploited in order to obtain very useful information about the environment and events, which can be used further to support decisional structures, either governmental or private.

Existing geospatial software provides spatial analysis and statistical functionality for vector, grid and raster data, but challenges arise in identifying and extracting elements from aerial and satellite imagery.

Feature based search methods have been extensively studied and reliable algorithms have been implemented in order to identify features in images, while deep-learning methods require further experimentation and exploration.

ACKNOWLEDGMENT

The authors gratefully thank to the reviewers for their support and valuable comments, which helped to improve this article.

REFERENCES

[1] B. Balamurugan, A. R. Nandhini, S. Kadry and Gandomi A. H., *Big Data, Concepts, Technology, and Architecture*. Pondicherry: Wiley, 2021.

[2] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Wo´zniak, and F. Herrera: *A survey on data preprocessing for data stream mining: Current status and future directions*. Neurocomputing **239**, 39–57, 2017.

[3] F. Sassite, M. Addou and F. Barramou, "*A smart data approach for Spatial Big Data analytics*," 2020 IEEE International Conference of Moroccan Geomatics (Morgeo), 2020, pp. 1-6, doi: 10.1109/Morgeo49228.2020.9121920.

[4] Promptcloud. http://promptcloud.com/blog/big-data-evolution-technology-modern/.

[5] Aboul E. Hassanien and A. Darwish, *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*. Cham: Springer, 2021.

[6] SRTM. https://www2.jpl.nasa.gov/srtm/.

[7] Data life cycle. https://online.hbs.edu/blog/post/data-life-cycle.

[8] Top 50 open source web crawlers for data mining. https://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining/.

[9] Gensim. https://github.com/RaRe-Technologies/gensim, accessed: Jun. 2022.

[10] Sintelix. https://sintelix.com/technology/.

[11] Copernicus. https://www.copernicus.eu/en/accessing-data-where-and-how/conventional-data-access-hubs.

[12] Open Street Map data download, https://download.geofabrik.de/.

[13] EverySpec, http://everyspec.com/MIL-SPECS/MIL-SPECS-MIL-A/.

[14] USGS, https://earthexplorer.usgs.gov.

[15] Apache Hadoop. https://hadoop.apache.org/.

[16] Apache Spark. https://github.com/apache/spark, accessed May 2022.

[17] C. Wang, "*Research and Implementation of Remote Sensing Image Retrieval Technology*," Taiyuan University of Technology, 2011.

[18] T. Deselaers, D. Keysers and H. Ney, "*Features for image retrieval: an experimental comparison*," Kluwer Academic Publishers, 2008.

[19] P. J. Du, Y. H. Chen, F. Tao and H. Tang, *Spectral feature-based hyperspectral RS image retrieval*, Spectroscopy and spectral analysis, 2005, 25(08), pp. 1171-1175.

[20] Y. Zhao, L. Zhang, P. Li and B. Huang, "*Classification of high spatial resolution imagery using improved Gaussian-Markov random-field-based texture features*," in IEEE Transactions on Geoscience and Remote Sensing, vol. 45, no. 5, pp. 1458-1468, May 2007, doi: 10.1109/TGRS.2007.892602.

[21] Z. Shao, W. Zhou, L. Zhang and J. Hou, "Improved color texture descriptors for remote sensing image retrieval," *Journal of Applied Remote Sensing*, vol. 8, no. 1, 2014.

[22] X. Y. Wang, W. Y. Li, H. Y. Yang, P. Wang and Y. W. Li, "*Quaternion polar complex exponential transform for invariant color image description*," Applied Mathematics and Computation, vol. 256, no. C, pp. 951-967, 2015.

[23] S. A. Zisserman, "*Video google: a text retrieval approach to object matching in videos*," in Proc. of 9th IEEE International Conference on Computer Vision, 2003.

[24] F. Perronnin and C. Dance, "*Fisher kernels on visual vocabularies for image categorization*," in IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[25] H. Jegou, M. Douze, C. Schmid and P. Perez, "*Aggregating local descriptors into a compact image representation*," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.

[26] G. S. Xia, X. Y. Tong, F. Hu, Y. Zhong, M. Datcu and L. Zhang, "*Exploiting deep features for remote sensing image retrieval: A systematic investigation*," 2017.

[27] Z. Shao, P. Tang, Z. Wang, N. Saleem and S. Yam and C. Sommai, "BRRNet: *A fully convolutional neural network for automatic building extraction from high-resolution remote sensing images*, Remote Sensing, vol. 12, no. 6, pp. 1050, 2020.

[28] Z. Shao, Z. Zhou, X. Huang and Y. Zhang, "MREnet: simultaneous extraction of road surface and road centerline in complex urban scenes from very high-resolution images," *Remote Sensing*, vol. 13, no. 2, 2021.

[29] R. Zhang, Z. Shao, X. Huang, J. Wang and D. Li, "*Object detection in UAV images via global density fused convolutional network*," Remote Sensing, vol. 12, no. 3, 2020.

[30] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang and J. Li, "*Deep learning for content-based image retrieval : a comprehensive study*," in *Acm International Conference on Multimedia*, 2014.

[31] G. J. Scott, M. N. Klaric, C. H. Davis and C. R. Shyu, "*Entropy-balanced bitmap tree for shape-based object retrieval from large-scale satellite imagery databases*," IEEE Transactions on Geoscience & Remote Sensing, vol. 49, no. 5, pp. 1603-1616, 2011.

[32] H. Liu, R. Wang, S. Shan and X. Chen, "*Deep supervised hashing for fast image retrieval*," in IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[33] Y. Li, Y. Zhang, X. Huang, H. Zhu and J. Ma, "*Large-scale remote sensing image retrieval by deep hashing neural networks*," IEEE Transactions on Geoscience & Remote Sensing, pp. 1-16, 2017.