Chapter 18
The Image as Big Data Toolkit: An Application Case Study in Image Analysis, Feature Recognition, and Data Visualization

Kerry E. Koitzsch
Kildane Software Technologies Inc., USA

ABSTRACT

This chapter is a brief introduction to the Image As Big Data Toolkit (IABDT), a Java-based open source framework for performing a variety of distributed image processing and analysis tasks. IABDT has been developed over the last two years in response to the rapid evolution of Big Data architectures and technologies, distributed and image processing systems. This chapter presents an architecture for image analytics that uses Big Data storage and compression methods. A sample implementation of our image analytic architecture called the Image as Big Data Toolkit (IABDT) addresses some of the most frequent challenges experienced by the image analytics developer. Baseline applications developed with IABDT, status of the toolkit and directions for future extension with emphasis on image display, presentation, and reporting case studies are discussed to motivate our design and technology stack choices. Sample applications built using IABDT, as well as future development plans for IABDT are discussed.

OVERVIEW

Rapid changes in the evolution of "Big Data" software techniques have made the ability to perform image analytics --- the automated analysis and interpretation of complex semi-structured and unstructured data sets derived from computer imagery --- with much greater ease, accuracy, flexibility and speed than has been possible before even with the most sophisticated and high-powered single computers or data centers. The "Big Data" processing paradigm" including Hadoop, Apache Spark and distributed computing systems have enabled a host of application domains to benefit from image analytics and the treatment of images as Big Data including medical, aerospace, geospatial analysis and document processing applications. However, several challenges exist when developing image analytic applications.

DOI: 10.4018/978-1-5225-3142-5.ch018

Modular, efficient and flexible toolkits are still in formative or experimental development. Integration of image processing components, data flow control and other aspects of image analytics remain poorly defined and tentative. The rapid changes in Big Data technologies have made even the selection of a 'technology stack' to build image analytic applications problematic. The need to solve these challenges in image analytics application development have led us to develop an architecture and baseline framework implementation specifically to support Big Data image analytics.

In the past, low level image analysis and machine learning modules have been combined within a computational framework to accomplish domain tasks. With the advent of distributed processing frameworks such as Hadoop and Apache Spark, it has been possible to build integrated image frameworks that integrate seamlessly with other distributed frameworks and libraries and in which the 'image as Big Data' concept has become a fundamental principle of the framework architecture.

IABDT provides a flexible modular and plug-in oriented architecture which makes it possible to combine many different software libraries, toolkits, systems and data sources within one integrated, distributed computational framework. It is a Java and Scala-centric framework as it uses both Hadoop and its ecosystem as well as the Apache Spark framework and its ecosystem to perform the image processing and the functionality of image analytics. IABDT may be used with NoSQL databases such as Neo4j or Cassandra as well as with traditional relational database systems such as MySQL to store computational results. Apache Camel and the Spring Framework may also be used as "glue" to integrate components with one another.

One of the motivations for creating IABDT is to provide a modular, extensible infrastructure for performing preprocessing, analysis and visualization and reporting of analysis results specifically for images and signals. Leveraging the power of distributed processing as with the Apache Hadoop and Apache Spark frameworks and inspired by such toolkits as BoofCV, HIPI, LIRE, Caliph, Emir, ImageTerrier, Apache Mahout and many others, IABDT provides frameworks, modular libraries and extensible examples to perform Big Data analysis on images using efficient, configurable and distributed data pipelining techniques.

Image as Big Data toolkits and components are becoming resources in an arsenal of other distributed software packages based on Apache Hadoop and Apache Spark as shown in Figure 1.

Some potential modules being investigated as distributed module technologies in IABDT include:

- Genetic Systems; Genetic Algorithm (GA based) algorithms are an effective way to process image feature data as well as to solve many optimization problems within the individual data flows in the system.
- Bayesian Techniques. Naïve Bayes Classification and other Bayesian classifiers.
- Statistical Library Support. Distributed R and Distributed Weka are components in the statistical library support module.
- Hadoop Ecosystem extensions (including Apache Spark); Many new and interesting machine learning toolkits are associated with Apache Spark as well as with Apache Flink. The modular construction of the IABDT allows a mix-and-match approach while evaluating the appropriate libraries to use.
- Search Engines including Lucene (The Apache Software Foundation, 2011-2016), Nutch (The Apache Software Foundation, 2004-2014), Solr (The Apache Software Foundation, 2016) as well as customized search code for images.