Procedia Computer Science
Volume 53, 2015, Pages 1–8
2015 INNS Conference on Big Data

A Nested Hierarchy of Dynamically Evolving Clouds for Big Data Structuring and Searching

Plamen Angelov[1] and Pouria Sadeghi-Tehran[2]

[1] Lancaster University, Lancaster, U.K
p.angelov@lancs.ac.uk
[2] Lancaster University, Lancaster, U.K
pouria.sadeghi-tehran@rothamsted.ac.uk

Abstract
The need to analyse big data streams and prescribe actions pro-actively is pervasive in nearly every industry. As growth of unstructured data increases, using analytical systems to assimilate and interpret images and videos as well as interpret structured data is essential. In this paper, we proposed a novel approach to transform image dataset into higher-level constructs that can be analysed more computationally efficiently, reliably and extremely fast. The proposed approach provides a high visual quality result between the query image and data clouds with hierarchical dynamically nested evolving structure. The results illustrate that the introduced approach can be an effective yet computationally efficient way to analyse and manipulate stored-images which has become the centre of attention of many professional fields and institutional sectors over the last few years.

Keywords: dynamically evolving hierarchy of data clouds, content-based image retrieval, recursive density estimation, evolving clustering

1 Introduction

The number of images and videos captured by humans and uploading to social media is staggering. Surveys show that images and videos make up about 80 percent of all corporate and public unstructured big data [19]. It has been estimated that about 100 hours of video are uploaded to Youtube every minute and average of 350 million photos are uploaded to Facebook daily which as a result it is estimated that Facebook currently has more than 250 billion photographs in its collection [13]. Nevertheless, dealing with such gigantic image collections are not restricted to social networks. Users in other domains such as healthcare, defence and astronomy are exploiting the opportunities offered by the ability to access and manipulate gigantic image and video datasets efficiently. Image management at this scale requires highly computationally efficient approaches which provide accurate and visually meaningful results to be able to search and browse of the files. On the other hand, manual tagging for such large datasets is not feasible

Selection and peer-review under responsibility of the Scientific Programme Committee of INNS-BigData2015

doi:10.1016/j.procs.2015.07.273

and is prone to errors due to user's subjective opinions; therefore, having an efficient, fast and accurate retrieval system is essential more than ever.

CBIR is a field of research which attracts professionals and industries and it has been used both in the form of commercial [14, 1] and research [16, 12] products. Nevertheless, there are still many open research issues needs to be addressed in terms of effectiveness and handling real-time scenarios. One of the important attribute in any CBIR system is extracting visual features such as colour, texture and shape from an image or keypoint based feature extraction techniques such as SIFT and SURF; however, once the size of the image dataset increases a processing time to compute the matching between images deteriorates extensively.

The other important component of any CBIR system is clustering and organising multi-dimensional data extracted from dataset images. Variety of clustering techniques has been used to cluster/group relevant digital items (e.g. images) such as hierarchical, K-means clustering, etc. [8, 21]. The main drawback of such approaches is that the clusters are fixed and not evolving; therefore, adding even a single new image to the database requires the whole procedure, including the clustering to be repeated 'from scratch'. Furthermore, in high dimension, data becomes very sparse and distance measures become increasingly meaningless caused the performance of clustering techniques to be degraded.

The other recent works in CBIR systems include Relevance feedback which enables users to have more interaction with the system and provide feedback on the relevance of the retrieved images; then, the feedback is used for learning and improving the performance of the system. Navigation-Pattern-based Relevance Feedback (NPRF) is an example of such a system used in [18] to reduce the iterations of feedback by implementing the navigation patterns discovered from the query log. However, the proposed method cannot integrate user's profile into their system to improve the retrieval quality nor capable of applying it on a multimedia retrieval system. Learning from negative example in relevance feedback [11] is another approach which is a combination of relevance feedback with discriminators to combine a negative examples with positive ones to identify and implement important features in retrieval process.

In this paper, we proposed a dynamically evolving nested structure for Big Data structuring and searching with an application to CBIR systems. The approach builds automatically a hierarchically nested structure from unorganised Big Data streams facilitating an efficient and fast searching of most relevant data (e.g. images). Due to the combination of the hierarchical nested evolving structure and using local recursive density estimation (RDE), the proposed approach is extremely fast and computationally efficient. The proposed approach was tested on a dataset of 30,000 images on a desktop PC to evaluate the performance of the method in terms of accuracy and computational efficiency.

2 Proposed Method

The aim of the proposed method, subject of a pending patent application [3], is to provide an efficient, robust and fast approach for Big Data streams such as images for searching and retrieving information in large dynamically evolving data streams based on the query one requested by a user. The three main components of the novel approach consist of: First, extracting multiple features from image dataset to cover different aspects of image contents. Second, construct a hierarchically nested dynamically evolving data clouds for fast and computationally efficient searching and organising of image dataset using evolving local mean (ELM) clustering. Finally, evaluating the similarity between the query image and data clouds using RDE formula.