THINK INDIA JOURNAL
ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Ad-hoc Removal Method of Small-Scale Images from Big Data

[1]Devendran V, [2]Nandini Sethi
[1,2]School of Computer Science Engineering,
[1,2]Lovely Professional University, Punjab (India)
devendran.22735@lpu.co.in , nandinisethi2104@gmail.com

Abstract: This paper focuses on identifying the unusual sizes of images in Big Data. As the term Big Data itself refers, it is already Big and it still will go bigger and bigger through many channels like audio, video, web sites, advertisements, social media and many. Companies are investing huge amount of money for storing the raw data assuming that all the data will be useful and it is of high importance. In the centralized data servers, multiple copies of the same document / file which are getting stored may also create unwanted junk of big data. Considering the image data, the scale invariants problem may be overcome by removing very small-scale images as it is taking much processing time and having less accuracy. This paper helps in removing the unusual small-scale images using the simple ad-hoc method. All the image features are extracted using two standard feature extraction techniques i.e., Gray Level Co-occurrence Matrix and Zernike Moments. Then, the standard deviation on those features is exactly identifying the unusual images, which can be the big-scale, or the small-scale images. The results are proving the effectiveness of the proposed method. The work is implemented on Matlab 7.0 and tested on the "Yahoo Flickr Two Dataset".
Keywords: Big Data, Image Categorization, Co-occurrence matrix, Zernike Moments, Standard Deviation.

1. INTRODUCTION

Big Data consist of data which is growing rapidly and which is dynamic in nature so it becomes difficult to manage such tremendous amount of data. Every day data- growing rate is quite high. Data comes from all the social media sites such as Facebook, Flickr, Google+, emails, video, audio etc contributes a lot to Big Data. Some of the data is Structured, Un- structured and semi-structured. Extracting information from rows and column i.e. structured data is simple but extraction from unstructured form of data is complex. Most of the data is usually unstructured, for example, tweets and blogs are having unstructured pieces of text, while images and video are in structure-format used for storage and display. Also, it transforms such content into a structured format for further research. Big Data is based on four basic dimensions [1][2][3][4] as in Fig 1, which are as follows. Volume: The amount of data, which is growing vastly every year. Social media sites are greatest contributor on increasing the volume of data. Now we have data in petabytes but in near future it will become zettabyte. Velocity: The speed of increasing data or we can say the rate of data flow from various resources. Variety: Data which is coming from various resources having different format such as structured, un-structured and semi- structured (log files, server logs, videos, emails etc). Veracity: Data that we have from various resources must be accurate. We cannot make decisions on the basis of inaccurate data [9].
Processing time is very large with respect to bigdata. To minimize the same, particularly with images, as the flow of images is tremendous through many sources, it is our responsibility to remove all unwanted images to ensure the integrity of the existing data. Large amount of data is available on internet available through various social websites and mobile phone. Due to this increasing datasets, it is not far to use the terms as ExaByte,

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

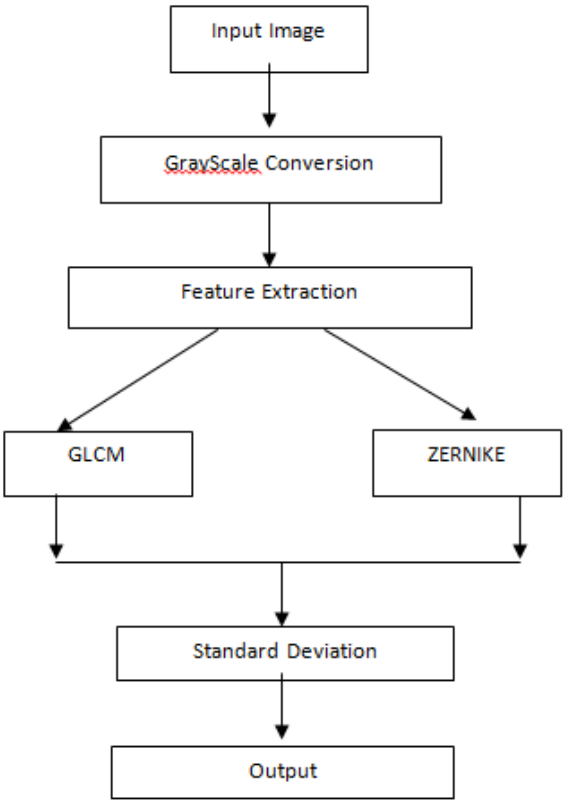
ZettaByte and PetaByte.

This paper is organized as follows: Section II describes various challenges of Big Data; Section III explains batch processing and stream processing in abundant data; Section IV gives details of technology used for Big Data; Section V deals with the proposed work; Section VI deals with data set used in this work; Section VII describes feature extraction methods; Section VIII gives implementation of Proposed Work; Section IX concludes with a conclusion.

2. CHALLENGES OF BIG DATA

A. Heterogeneity and Incompleteness

The challenges of big data range from dealing with floods of information in real time to making sense of streams of disparate data coming in at variable speeds. Human can read any kind of information in any format comfortably. But, machines do not tolerate the heterogeneity. The natural languages can give very valuable information. The machine analysis algorithms do not perform well on heterogeneous data but expect homogeneous data. In this respect, the first step which is data processing the data should be carefully structured. Let us consider one example of a patient who has multiple medical procedures at a hospital. Multiple Records are created of a single patient as first record which maintain all the details of laboratory test, second record will maintain the stay information of the patient and the third record will be maintained for whole life hospital interaction of the patient. Instead of following the Above design, The three design choices listed will result in successively less structure and provide greater benefit [1][10][11]. However, the less structured design will be better and effective for many purposes. In this Author [12] is discussing about the better representation, nd analysis of semi-structured data in general. However, computer systems will work effectively if it has all identical data with respect to size and structure.



ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Fig. 1: Proposed Work

B. Scale

The first thing anyone who thinks about Big Data is its size only. It shows that the word "big" is having its real meaning. It has been a challenging issue to manage large and rapidly increasing volumes of data for many decades. In the past, computer processors getting faster and able to speed up to meet the requirement of handling bigdata, which provides us the ability to deal with increasing large amount of data. But, there is another issue is that the data is increasing very faster than the availability of computing resources and also to note that CPU speeds are static [14].

C. Timeliness

Another side of the size is computing speed. The larger the data set is to be processed, the longer duration it will take to analyze. The design of a system should be such a manner that it has to process effectively with size and also to result in a system [15]. Our proposed work will analyze the existing image data and trying to help the system by removing the unfavorables. This saves the processing time at least partially.

D.Privacy and Human Collaboration

Issues of privacy also increase with the increase in the data. There is a risk public feels in linking their personal data from different sources. Various researches has been performed to resolve this issue at least with respect to the modeling and analysis phase in the pipeline including assuming the similar value to human input at all stages of the analysis pipeline.

3. DATA ABUNDANCE – STREAM VS BATCH PROCESSING

Big Data analysis can be obtained through the stream processing of data and batch processing of data on the basis of processing time requirements. In Stream processing of data, tremendous amount of data is generated during the stream processing in the form of stream so rapidly processing of data is required. There is an assumption that we do mostly that fresh data is taken during stream processing [7][8]. A stream generates huge amount of data, which can't be stored in any memory place. That's why we need faster processing of data. For e.g. online application uses streaming processing where processing of applications takes time in milliseconds and seconds. In Batch processing of data, the data is stored in memory, which can be distributed and then analysed. Data is divided in to small size chunks and then processing of data is obtained through parallel computing i.e. where simultaneous execution of the data happens and intermediate results are obtained. After that all the intermediate results are combined together i.e. integrate to obtain the final results. For instance Map Reduce is used for doing batch processing. Batch processing is widely used for getting the result in real time.

4. TECHNOLOGY

In Big Data processing of data by the existing methods is too complex and also time consuming. So there is a need of the new technologies are required. To process large amount of data, one the best technique is to perform parallel computing in which multiple tasks are performed simultaneously at single point of time. There are some of the technologies [5][6][21], which are being used in the big data. They are as discussed below.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

Hadoop: Hadoop is an open source project of apache which works on clusters of computers to process large distributed data sets . It consist of various features like fault tolerance, cost efficiency, felexibilty and scalability.

Hadoop Distributed File System (HDFS): This system gives the storage to large distributed files across distributed servers.

MapReduce: Originally owner of MapReduce is Google but now it is incorporated by the apache. The MapReduce performs two functions. Those are Map and Reduce. In map function it will firstly filter the data and then sort it. And in reduce function it will summarize the data.

Apache Hive: It is based on data warehouse infrastructure, which is built in top of Hadoop. It is used for summarizing data, query and making analysis [13].

NoSQL: It provides less constrained database system then SQL. It provides mechanism for storing and retrieving data, which is less constrained.

5. PROPOSED WORK

This paper analyses images in respect of scales. Of course, the images in dataset in dealing with foreground object of size (480x640) colored. The conversion into grayscale is mandatory in this work. We have taken 13 images of test data from yahoo Flikr-shoe images and extracted gray scale Co-occurrence features and calculated the standard deviation. The results are given in Table 1, sharing that the small-scale images are used and extracted the features using Zernike moments. The results of calculated standard deviation are presented in Table 1. Surprisingly, the same small-scale images are identified correctly with the help of simple standard deviation.

6. YAHOO FLIKR BIG DATA SET

The data used in this work has been taken from “Yahoo Flikr Two” Dataset having 109 categories of Yahoo Shopping shoe photos. This standard database consists of 5250 colored images of size (480x640). In this work, the first category of Shopping shoe photos ie., ‘aerosols sandals’ is taken for our work which is having only thirteen images. The sample images of the dataset are given in Table 1. The entire dataset is concentrating on foreground objects excluding the background objects. Hence, pre-processing is not needed in our work and features are extracted without any segmentation processes.

FEATURE EXTRACTION METHODS

A. Texture Features

In this paper the texture feature extraction is based on the spatial gray-level co-occurrence-matrix (SGLCM). SGLCM is used to provide a graphical representation in the form of histogram which quantized two images pixels which are having spatial relationship on the basis of grey-level value. On the basis of above representation various texture features are calculated such maxprob; energy; entropy; inertia; homogeneity; contrast; inverse; correlation by verging over different angles 00, 450, 900 and 1350 degrees. For each image 8 features were extracted on the basis of single angle, so in total 32 features are extracted for each image on the basis of all above mentioned angles.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

B. Zernike Moments

From past many years, Moments acts as the crucial part in image processing applications. Main issue in these methods is that these were unable to completely describe the object in such a way, so that it can be reconstructed on the basis of moments sets.Basically these descriptors are not orthogonal. We can use the kernel of Zernike moments which is a set of orthogonal Zernike polynomials and it is defined inside a unit circle over the polar coordinate space The two-dimensional Zernike moments of order p with repetition qof an image intensity function f(r, θ) are defined as;

$$Z_{pq} = \frac{p+1}{\pi}\int_0^{2\pi}\int_0^1 V_{pq}(r,\theta) f(r,\theta) r dr d\theta; |r| \le 1 \quad (1)$$

whereVpq(r, θ) which are Zernike polynomials are defined as:

$$V_{pq}(r,\theta) = R_{pq}(r) e^{-\hat{j}q\theta}; \hat{j} = \sqrt{-1} \quad (2)$$

andRpq(r), is defined as follows:

$$R_{pq}(r) = \sum_{k=0}^{\frac{p-|q|}{2}} (-1)^k \frac{(p-k)!}{k!\left(\frac{p+|q|}{2}-k\right)!\left(\frac{p-|q|}{2}-k\right)!} r^{p-2k} \quad (3)$$

where $0 \le |q| \le p$ and $p - |q|$ is even.

The discrete approximation will be calculated for equation (1) where N will be considered as number of pixels among each axis of the image:

$$Z_{pq} = \lambda(p,N)\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} R_{pq}(r_{ij}) e^{-\hat{j}q\theta_{ij}} f(i,j); 0 \le r_{ij} \le 1 \quad (4)$$

In the above equation λ(p, N) is the normalizing constant. The image coordinate transformation for interior of the unit circle is calculated on the basis of following equation:

$$r_{ij} = \sqrt{x_i^2 + x_j^2}; \theta = \tan^{-1}\left(\frac{y_i}{x_i}\right); x_i = c_1 i + c_2;$$

$$y_i = c_1 j + c_2; \quad (5)$$

Zpq consist of two parts i.e. real and imaginary part, Z_{pq}^c, Z_{pq}^s which is calculated as follows:

$$Z_{pq}^c = \frac{2(p+1)}{\pi}\int_0^{2\pi}\int_0^1 R_{pq}(r)\cos(q\theta) f(r,\theta) r dr d\theta \quad (6)$$

$$Z_{pq}^s = \frac{2(p+1)}{\pi}\int_0^{2\pi}\int_0^1 R_{pq}(r)\sin(q\theta) f(r,\theta) r dr d\theta \quad (7)$$

where $p \ge 0$, $q > 0$.

$$\lambda(p,N) = \frac{4(p+1)}{(N-1)^2\Pi}; c1 = \frac{\sqrt{2}}{N-1}; c2 = \frac{-1}{\sqrt{2}} \quad (8)$$

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

For the purpose of implementation, the image (NxN) is firstly transformed and then normalized over a unit circle; i.e., x2 + y2 ≤ 1, in which the transformed unit circle image is bounding the square image. In the square-to-circular transformation,Therefore,

$$x_i = \frac{\sqrt{2}}{N-1} i + \frac{-1}{\sqrt{2}} \text{ and } y_i = \frac{\sqrt{2}}{N-1} j + \frac{-1}{\sqrt{2}} \quad (9)$$

8. RESULTS AND DISCUSSION

This paper aims to find the better approach for identifying the unusual pattern of images to remove from storing in the data servers and avoiding it becoming the garbage. Here, in this work, we have used standard deviation approach in the Yahoo Data Set. This data set, thirteen colored images are available of size (480x640). All the images are converted into gray scale images of size (512x512). Data are collected from each images using gray-scale co-occurrence feature extraction method and also using Zernike moments. Standard deviations are derived from each data extracted from two different methods separately as in Table 1. Features are of 32 nos. from each images and hence, the 32x13 matrix is formed in the case of co-occurrence matrix. For Zernike, 36 nos. of data is extracted from each image, and hence 36x13 matrixes are formed. These matrices are used for the calculation of standard deviations without any flag values.

Table 1. Results of yahoo-Flickr Dataset

Images (1-7)	Zernike based Std. Dev.	Co-occurrence based Std. Dev.	Images (8-13)	Zernike based Std. Dev.	Co-occurrence based Std. Dev.
	1.57	5.4984		1.877	6.0156
	1.3075	2.2261		1.9459	6.2156
	1.672	5.5749		1.828	6.0635

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

	**3.1736**	**7.7668**		**2.8752**	**7.8448**
	1.9518	6.1015		1.521	5.2321
	1.9844	5.9665	-	-	-

It is clear from the Fig 2 and Fig 3 showing the values of standard deviations obtained from Zernike moments and co-occurrence matrix respectively. In Fig 2, the values depicts that the serial number (respective image number) 5 and 12 are having extraordinary difference among the values of other images.

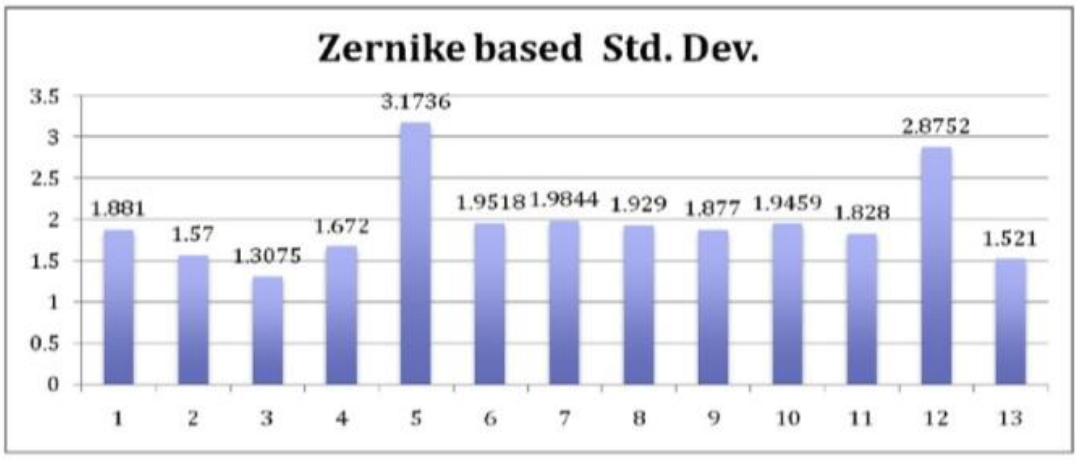


Fig 2. The Results of Standard Deviations using Zernike Moments

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

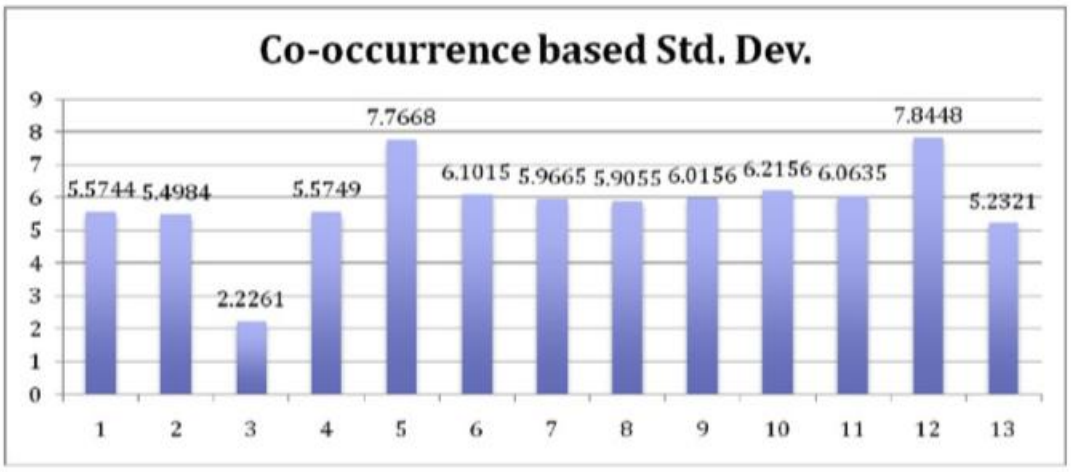


Fig 3. The Results of Standard Deviations using Co-occurrence Matrix

Similarly, the effects of standard deviations are observed when Zernike moments used for testing. Hence, the proposed ad-hoc method of standard deviation is giving better results in a short span of time identifying the images that can be neglected from wasting the memory and become garbage in the Big Data Servers. This kind of images may mislead to wrong decisions that never should happen.

9.CONCLUSION

This paper discussed the issue of maintenance of Big Data, specifically on image data. The importance of filtering the unusual data is focused and given importance on the real-time environment in data collecting issues. In turn, this reflects the companies in reduction of investment in hardware and also benefit in valued data. Maintenance in the sense, keeping an eye on valued data that is to be checked and processed immediately before storing in the Big Data Servers. In this paper, simple standard deviation is used and tested in the image dataset for identifying unusual pattern using Zernike and Co-occurrence feature extraction methods. The results are giving prominent identification unexpectedly, this method identifies the unusual images fairly well and the results are given. The entire work tested on the “Yahoo Flickr Two” dataset.

REFERENCES

[1] R. Hong, Y. Yang, M. Wang and X. S. Hua, "Learning Visual Semantic Relationships for Efficient Visual Retrieval," in IEEE Transactions on Big Data, vol. 1, no. 4, pp. 152-161, Dec. 2015.
[2] Qingchen Zhang et al., “A survey on deep learning for big data“, Information Fusion, Vol. 42, pp. 146-157, 2018.
[3] IbrarYaqoob et al., “Big data : From beginning to future“, International Journal of Information Management, Vol. 36 (6), pp. 1231-1247, 2016.
[4] Hameed, Z., Hing, YS., Cho, YM., Ahn, SH., Song, CK., ”Condition monitoring and fault detection of wind turbines and related algorithms: A review”, Renewable and Sustainable Energy Reviews, Vol. 13(1), pp. 1-39, 2009.
[5] W. H. Inmom, Building the Data Warehouse, 4th Edition, John Wiley & Sons, Inc., 2007.

ISSN: 0971-1260
Vol-22-Issue-16-August-2019

[6] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy and B. Becker, The Data Warehouse Lifecycle Toolkit, 2nd Edition, John Wiley & Sons, Inc., 2008 .
[7] Ahmed Ismail, AbdulazizShehab et al., Healthcare Analysis in Smart Big Data Analytics: Reviews, Challenges and Recommendations, Security in Smart Cities: Models, Applications and Challenges, pp.27-45, 2018.
[8] SjaakWolfert, LanGe et. al., "Big Data in Smart Farming – A Review", Agricultural Systems, Vol. 153, pp. 69-80, 2017.
[9] Cichy, R.M., Pantazis, D., &Oliva, A."Resolving human object recognition in space and time", Nature Neuroscience, Vol. 17(3), pp. 455-464, 2014.
[10] MuktiAkter and LivijaJakaite, "Extraction of Texture Features from X-Ray Images: Case of Osteoarthritis Detection", 3rd International Congress on Information and Communication Technology, pp. 143-150, 2019.
[11] Wei Ji, Shubin Yin, Lihui Wang, "A big data analytics based machining optimization approach", Journal of intelligent manufacturing, pp.1483-1495, 2019.
[12] Konkle, T., Brady, T.F., Alvarez, G.A., &Oliva, A., "Scene memory is more detailed than you think: the role of categories in visual long-term memory". Psychological Science, Vol. 21(11), pp. 1551-1556, 2010.
[13] T. White. Hadoop: The Definitive Guide. Yahoo Press, 2010.Apache Hive. http://hive.apache.org/
[14] Andreas Kamilaris et al., "A review on the practice of big data analytics in agriculture", Computers and Electronics in Agriculture, Vol. 143, pp. 23-37, 2017.
[15] Eric Redmond, Jim R. Wilson, Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement, O'Reilly, 2012.
[16] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Morgan & Claypool, 2011
[17] J. Lehmann et al., "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia", Semantic Web Journal, Vol.5(2), 2014.
[18] Z. Nie, J., R. Wen, W.Y. Ma, "Statistical Entity Extraction from the Web", Proceedings.of the IEEE, Vol. 100(9), 2012.
[19] Devendran V, HemalathaThiagarajan, Amitabh Wahi, "Zernike based scene categorization using artificial neural networks", IEEE Proceedings of International Conference on Computing Communication and Networking (ICCCN'08), Chettinad College of Engineering and Technology, Karur, pp-18-20, December 2008.
[20] Kedong Wang, Han Wang, Jinling Wang, "Terrain Matching by Fusing HOG with Zernike Moments", IEEE Transactions on Aerospace and Electronics Systems, 2019.
[21] Amir Gandomi, MurtazaHaider, "Beyond the hype: Big data concepts, methods, and analytics", International Journal of Information Management, Volume 35, Issue 2, Pages 137–144, April 2015.