Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

Using Apache Spark Streaming and Kafka to Perform Face Recognition on Live Video Streams of Pedestrians

Vikas Tripathi[1], Durgaprasad Gangodkar[2], Devesh Pratap Singh[3], Dibyahash Bordoloi[4]

[1]Department of Computer Science & Engineering,Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002
[2]Department of Computer Science & Engineering,Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002
[3]Department of Computer Science & Engineering,Graphic Era Deemed to be University, Dehradun, Uttarakhand India, 248002
[4]Head of the Department, Department of Computer Science & Engineering,Graphic Era Hill University, Dehradun, Uttarakhand India, 248002

ABSTRACT

Through the use of Apache Spark Streaming, Kafka, and OpenCV on a distributed platform, we propose a method for recognising pedestrians in real time. This method intends to recognise motion and faces by matching new photos to a database of previously identified images. Apache Spark Streaming and Kafka have been utilised as real time analysis frameworks, which deliver event based decision making on Hadoop's distributed environment, because video processing and analysis from multiple resources is time consuming when using the Cloud or even any single highly configured machine. If real-time event analysis can be performed continuously, a choice may be made either immediately thereafter or simultaneously with the occurrence of the event in question. Hadoop is the foundation of all real-time analysis distributed solutions, so even processing massive amounts of films in parallel is no longer a bottleneck. Whenever continuous monitoring and decision making is involved with a large data set, this event based analysis can be put to use. This includes, but is not limited to, the monitoring of border areas of countries by cameras and drones, road traffic monitoring, the life science domain, airlines, logo recognition, and more

Keywords: .

INTRODUCTION

Importance of Distributed Data Processing: \sWith the rapid growth of data, it progressively becomes humungous and the phrase buzzing around the globe is BigData, it is a term for data sets that are so massive or complicated that standard data processing programmes are insufficient to cope with them. Data analysis, data collection, data search, data transmission, data visualisation, data querying, data updating, and data privacy are all areas of difficulty. Predictive analytics, user behaviour analysis, and other sophisticated data analysis approaches that derive value from data are often what people mean when they use the phrase "BigData," rather than a specific volume of data

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

[1]. Typically, the data sets that are part of BigData are too large for frequently used software tools to efficiently acquire, correct, manage, and analyse in a reasonable amount of time. The "size" of BigData is a continually evolving objective, spanning from many tens of terabytes to hundreds of petabytes or perhaps an Exabyte. Furthermore, it may supplement data warehouse systems by acting as a buffer for processing incoming data before it is added to the data warehouse or for removing old or seldom viewed data [2]. BigData's ability to provide insight into operational difficulties may help businesses enhance their operations as a whole. Machine data, which can come from anything from computers to sensors to metres to GPS devices, can be the basis for operational insights; meanwhile, BigData offers businesses unprecedented visibility into their customers' decision-making processes through the monitoring and analysis of shopping trends, recommendations, purchasing behaviours, and other drivers known to affect sales. One further area where BigData is being put to use is in cyber security and fraud detection. Businesses can improve security and intelligence analysis tools using real-time data. They can also process, store and analyse a greater range of data types to increase intelligence, security and law enforcement knowledge [3].

Hadoop's Significance:

Apache Hadoop is an open source software framework for storing and processing huge size of data-sets on clusters of cheap hardware. Hadoop is a top-level Apache project that is being developed and utilised by people all around the world.

Every component of Hadoop is built on the premise that hardware faults are inevitable and the system should be able to deal with them without any user intervention. Doug Cutting and Mike Cafarella developed Hadoop in 2005. It was initially intended to enable distribution for the Nutch search engine project. Hadoop was called after a yellow stuffed elephant that Doug's kid had. With both Hadoop and Big Data, the solution follows the problem. Hadoop and its related frameworks were created to address the issue of BigData in the information technology sector. When it comes to handling and analysing BigData, the Hadoop Eco-System has you covered from head to toe.

APACHE SPARK

Apache Spark's programming style extends automated fault tolerance to a considerably broader set of use cases than MapReduce did. Specifically, MapReduce is inefficient for multi-pass applications that need low-latency data exchange amongst several concurrent processes. The analytics field makes extensive use of iterative algorithms, which are used in many machine learning and graph algorithms (like PageRank) as well as other popular applications. The ability to load data into RAM across a cluster and frequently query that data for insights is a key feature of interactive data mining [3].

Streaming apps that save their global state over time.

Due to their reliance on acyclic data flow, traditional MapReduce and DAG engines aren't well-suited for such applications. Instead, such applications need running as a series of separate jobs, each of which takes data from stable storage (such a distributed file system) and sends it back. The data must be loaded and written back to mirrored storage after each step, which is an expensive process.

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

Face identification, shadow elimination, background subtraction, feature extraction, etc. are only a few examples of the many algorithms that may be used in the processing of a video. The stages may be completed in any order, with some requiring quick online processing. Face recognition (for instance, the algorithm included with OpenCV) may be done extremely rapidly, hence it is preferable to place such a lightweight algorithm on more than one node in order to achieve the fast response [4]. This has practical uses in a variety of contexts; for instance, a smart tourist guide may use this to keep track of the number of visitors in the area in real time. For example, if a system wants to analyse streaming video data, it will need to execute GMM (Gaussian mixture model) for background removal, which is a somewhat high weight method, in its batch processing layer offline.

For this reason, it is necessary to use the batch processing capabilities of certain platforms in addition to the real-time in-memory processing capabilities of others if we are to successfully and efficiently analyse video. This is a critical gap in the state of video processing studies. For this reason, this article presents efforts toward a high-performance video on a distributed platform that combines distributed with rapid processing [5].

For the reasons listed below, processing images needs Stream data. First, real-time processing of video is essential for security purposes. It's difficult and memory-intensive to process video data after some period of time has passed. Third, data streaming, or the flow of data in real time from a source to a stream data processing cluster, must be managed effectively for the processing of image data. Apache Kafka is an effective solution to the challenge of real-time data transport.

APACHE KAFKA IMPORTANCE

To publish and subscribe to streams of data, Apache Kafka serves as a distributed streaming platform. Both producers and consumers may benefit from this data's availability in a fault-tolerant database.

A free and open-source Scala-based message broker project created by the Apache Software Foundation [6]. The project's goal is to provide a standardised, high-throughput, low-latency infrastructure for processing streaming data. Transaction logs played a significant role in the development of this system's architecture.

No amount of data loss is acceptable if true value is to be extracted from large data. Apache Kafka's disc structures are O (1), allowing for constant-time performance even with petabytes (TB) of stored messages.

The Significance of Apache Storms

Fast and flexible, Apache Spark is a cluster-computing solution built for a wide variety of uses. There are modules for working with SQL data, as well as semi-structured and unstructured data, machine learning with MLlib, graph processing with GraphX, and streaming data with Spark. The Open Source Image Processing Library (OpenCV): It is a free and open-source library distributed under the BSD licence, optimised for speedy calculation and with a particular emphasis on real-time use cases. Although it is built in C++, this library also has a Java application programming interface (API). In OpenCV, you'll find hundreds of CV algorithms for handling and analysing your image

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

and video data. Free and open-source software for detecting pictures, which is then put to use to pull stills from moving video. This project has extensive documentation and sees extensive use. OpenCV has approximately 500 features that may be used for things like robotics, security, medical imaging, and inspection of factory-made goods.

It also has an MLL that may be used for a variety of purposes and is geared toward statistical pattern detection. The OpenIMAJ (OPEN Intelligent Multimedia Analysis in Java) library and tool, on the other hand, is a Java-based solution for large-scale multimedia content analysis and picture indexing. Broad cutting-edge computer vision methods are included. The jar files used in the distribution are modular, and it is licenced in the spirit of the BSD licence. Maximum code maintainability is achieved by keeping all components modular throughout design and implementation [7]. Both packages use equivalent methods for picture recognition. Based on these shared characteristics, we decided to utilise Open IMAJ as the foundation for implementing our code to extract and characterise interesting regions of pictures through the Scale-Invariant Feature Transform (SIFT). Random Sample Consensus (RANSAC) is used to fit a geometric model known as an Affine Projection to locate analogue pixels. Revert back to the original group of matches. When we run a pattern detection algorithm, we get a list of similarities or matches. The concept of proximity to a threshold is useful. If the sum is more than the threshold, the logo will appear inside a box. You should take note that the quality of the library has a significant impact on the quality of the systems.

PRESENT METHODS For Detecting Pedestrians and Faces
In addition to facial recognition, the paper's authors also successfully deployed pedestrian detection methods. Although it potentially save a life, most studies in this area have concentrated on locating persons in a standing position survival in a catastrophic environment, among other applications. Astounding uses have been found for pedestrian detecting technology [8].

Among published detectors, those included in OpenCV, an open vision library, are typical.

Dalal's HOG detector from 2005, which was taught by the INRIA People Database histogram of oriented gradients. Template objects smaller than 64(w) x 128(h) are difficult to identify (h). The HOG(Daimler) algorithm uses the Daimler Pedestrian Dataset to learn how to recognise pedestrians. Objects fitting a tiny template size were inside its detection range.

Hogcascades is a detector that uses a cascade algorithm to analyse the HOG feature. The Haarcascades Viola2001 Detector can quickly identify items, making it a competitive option.

Since the focus of this study is on detecting in a dispersed environment rather than on the precision of such detection, we have chosen to employ Haarcascades as our detector for locating pedestrians [9].

Face recognition is a highly sought after biometrics issue with several real-world applications. Researchers in the biometrics, pattern-recognition sector, and computer-vision disciplines are all drawn to the challenges of facial-recognition technology. Besides their usage in biometrics, certain

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

face recognition algorithms have found widespread use in other fields, including video compression, indexing, etc. They may also be used for the purpose of categorising multimedia content, making it possible for the user to quickly and easily search for information of interest. With the recent rise in terrorist acts, the challenge of facial recognition has taken on more significance. Using a person's face as a form of authentication not only makes it less likely that they will need to remember a password, but it also has the potential to significantly increase security when used in tandem with other authentication and authorization methods. Considering that a licence for a reliable commercial face recognition system may cost as much as $150,000, you can see how important this issue is.

PROPOSED METHODOLOGY

To publish and subscribe to streams of data, Apache Kafka serves as a distributed streaming platform. Both producers and consumers may benefit from this data's availability in a fault-tolerant database.

A free and open-source Scala-based message broker project created by the Apache Software Foundation. The project's goal is to provide a standardised, high-throughput, low-latency infrastructure for processing streaming data. Transaction logs played a significant role in the development of this system's architecture.

No amount of data loss is acceptable if true value is to be extracted from large data. Apache Kafka's disc structures are O (1), allowing for constant-time performance even with petabytes (TB) of stored messages.

The Significance of Apache Storms

Fast and flexible, Apache Spark is a cluster-computing solution built for a wide variety of uses. There are modules for working with SQL data, as well as semi-structured and unstructured data, machine learning with MLlib, graph processing with GraphX, and streaming data with Spark. The Open Source Image Processing Library (OpenCV): It is a free and open-source library distributed under the BSD licence, optimised for speedy calculation and with a particular emphasis on real-time use cases. Although it is built in C++, this library also has a Java application programming interface (API). In OpenCV, you'll find hundreds of CV algorithms for handling and analysing your image and video data. Free and open-source software for detecting pictures, which is then put to use to pull stills from moving video. This project has extensive documentation and sees extensive use. OpenCV has approximately 500 features that may be used for things like robotics, security, medical imaging, and inspection of factory-made goods.

It also has an MLL that may be used for a variety of purposes and is geared toward statistical pattern detection. The OpenIMAJ (OPEN Intelligent Multimedia Analysis in Java) library and tool, on the other hand, is a Java-based solution for large-scale multimedia content analysis and picture indexing. Broad cutting-edge computer vision methods are included. The jar files used in the distribution are modular, and it is licenced in the spirit of the BSD licence. Maximum code maintainability is achieved by keeping all components modular throughout design and implementation. Both packages use equivalent methods for picture recognition. Based on these shared characteristics, we decided to utilise Open IMAJ as the foundation for implementing our

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

code to extract and characterise interesting regions of pictures through the Scale-Invariant Feature Transform (SIFT). Random Sample Consensus (RANSAC) is used to fit a geometric model known as an Affine Projection to locate analogue pixels. Revert back to the original group of matches. When we run a pattern detection algorithm, we get a list of similarities or matches. The concept of proximity to a threshold is useful. If the sum is more than the threshold, the logo will appear inside a box. You should take note that the quality of the library has a significant impact on the quality of the systems.

PRESENT METHODS For Detecting Pedestrians and Faces

In addition to facial recognition, the paper's authors also successfully deployed pedestrian detection methods. Although it potentially save a life, most studies in this area have concentrated on locating persons in a standing position. Survival in a catastrophic environment, among other applications. Astounding uses have been found for pedestrian detecting technology. Among published detectors, those included in OpenCV, an open vision library, are typical. Dalal's HOG detector from 2005, which was taught by the INRIA People Database histogram of oriented gradients. Template objects smaller than 64(w) x 128(h) are difficult to identify (h). The HOG(Daimler) algorithm uses the Daimler Pedestrian Dataset to learn how to recognise pedestrians. Objects fitting a tiny template size were inside its detection range. Hogcascades is a detector that uses a cascade algorithm to analyse the HOG feature [10]. The Haarcascades Detector can quickly identify items, making it a competitive option. Since the focus of this study is on detecting in a dispersed environment rather than on the precision of such detection, we have chosen to employ Haarcascades as our detector for locating pedestrians. Face recognition is a highly sought after biometrics issue with several real-world applications. Researchers in the biometrics, pattern-recognition sector, and computer-vision disciplines are all drawn to the challenges of facial-recognition technology. Besides their usage in biometrics, certain face recognition algorithms have found widespread use in other fields, including video compression, indexing, etc. They may also be used for the purpose of categorising multimedia content, making it possible for the user to quickly and easily search for information of interest. Forensic science, identification for law enforcement, surveillance, authentication in banking and security systems, and access management for secure locations are just few of the many potential applications of a reliable face recognition system. With the recent rise in terrorist acts, the challenge of facial recognition has taken on more significance. Using a person's face as a form of authentication not only makes it less likely that they will need to remember a password, but it also has the potential to significantly increase security when used in tandem with other authentication and authorization methods. Considering that a licence for a reliable commercial face recognition system may cost as much as $150,000, you can see how important this issue is.

ALGORITHM

The Haarcascades – Detector of Viola2001 is superior to other pedestrian detectors in terms of speed. When it comes to face identification using eigenfaces, the high dimensionality of the picture representation we're given is a major issue. A pixelated picture, as contrast to a grayscale image, which occupies a, is already in a -dimensional image space. Is it true that all dimensions provide us with the same amount of benefit? Since we need some kind of discrepancy in the data in order to make a call, we need to isolate the factors that make up the bulk of that data. Visualize a scenario in which the light is the external variable responsible for the observed variation in your data. It is hard

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

to classify the projected samples since the components found by a principal components analysis (PCA) may not include any discriminative information at all the described how he had applied the technique to the challenge of categorising flowers. Instead of trying to maximise the overall dispersion, Linear Discriminant Analysis focuses on finding the combination of characteristics that distinguishes best across classes. The concept is straightforward: in the lower-dimensional representation, classes that share certain characteristics should be close together, whereas classes that have various characteristics should be as far apart from each other as feasible.

Eigenfaces and Fisherfaces, which use local binary patterns histograms, are two examples of more holistic approaches to facial identification. You think of your information as a vector in a very large picture space. Since we all agree that having too many dimensions is harmful, we look for lower-dimensional subspaces where (hopefully) meaningful information is still available. Reality, however, is far from ideal. Whether you take one picture or ten pictures of the same individual, you can never be sure that the lighting will be quite right. If just one picture exists for each individual, then what? It's possible that the subspace covariance estimations, and therefore the recognition, are completely off. we can see how well Eigenfaces and Fisherfaces do in terms of Rank-1 recognition rates. Thus, the Fisherfaces approach is not very useful for improving identification rates, and you will require at least 8(+- 1) photos for each individual. Therefore, various studies aimed to learn how to extract regional traits from photographs. Instead of describing an object's appearance in its whole, as a high-dimensional vector, this method focuses on its individual parts. The dimensions of the characteristics you extract in this approach will be naturally reduced. Excellent plan! Still,

You'll quickly see that the picture we're provided has more problems than just varying lighting. The local description has to be somewhat resistant to changes in the picture, such as those caused by scaling, translation, or rotation. Local Binary Patterns is another technique that has its origins in 2D texture analysis, alongside SIFT. By comparing each pixel to the ones around it, Local Binary Patterns may provide a summary of the local structure in a picture. Apply a threshold to the pixels around a given pixel. Assign a value of 1 if the central pixel's intensity is higher than or equal to that of its neighbour, and a value of 0 otherwise. For each pixel, you'll get a binary number like 11001111. Therefore, there are 28 possible combinations with 8 neighbouring pixels; these are known as Local Binary Patterns (or LBP codes). It was a fixed 3x3 neighbourhood that was used by the first LBP operator described in the literature.

The suggested framework is meticulously crafted to allow high parallelism and remove performance constraints as compared to baseline solutions. This framework's performance claims are supported by experimental evaluations using real data.

REFERENCES

1. Dhamija, J., Choudhury, T., Kumar, P., & Rathore, Y. S. (2017, October). An Advancement towards Efficient Face Recognition Using Live Video Feed:" For the Future". In *2017 3rd international conference on computational intelligence and networks (CINE)* (pp. 53-56). IEEE.
2. Gafni, O., Wolf, L., & Taigman, Y. (2019). Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9378-

Webology, Volume 18, Number 5, 2021
ISSN: 1735-188X
DOI: 10.29121/WEB/V18I5/85

9387).

3. Cho, M., & Jeong, Y. (2017). Face recognition performance comparison between fake faces and live faces. *Soft Computing*, *21*(12), 3429-3437.
4. Manjunatha, R., & Nagaraja, R. (2017). Home security system and door access control based on face recognition. *International Research Journal of Engineering and Technology (IRJET)*, *4*(03), 2395-0056.
5. Wang, J., Amos, B., Das, A., Pillai, P., Sadeh, N., & Satyanarayanan, M. (2018). Enabling live video analytics with a scalable and privacy-aware framework. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *14*(3s), 1-24.
6. Deeba, F., Memon, H., Dharejo, F. A., Ahmed, A., & Ghaffar, A. (2019). LBPH-based enhanced real-time face recognition. *International Journal of Advanced Computer Science and Applications*, *10*(5).
7. Parchami, M., Bashbaghi, S., & Granger, E. (2017, August). Cnns with cross-correlation matching for face recognition in video surveillance using a single training sample per person. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 1-6). IEEE.
8. Bakshi, N., & Prabhu, V. (2017, December). Face recognition system for access control using principal component analysis. In *2017 International Conference on Intelligent Communication and Computational Techniques (ICCT)* (pp. 145-150). IEEE.
9. Kumar, S., Singh, S., & Kumar, J. (2018). Live detection of face using machine learning with multi-feature method. *Wireless Personal Communications*, *103*(3), 2353-2375.
10. Jose, E., Greeshma, M., Haridas, M. T., & Supriya, M. H. (2019, March). Face recognition based surveillance system using facenet and mtcnn on jetson tx2. In *2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)* (pp. 608-613). IEEE.