Survey on Multiple Objects Tracking in Video Analytics

Anjali Parihar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Nigdi, Maharashtra

Priyanka Nagarkar
Department Of Computer Engineering
Pimpri Chinchwad College of Engineering
Nigdi, Maharashtra

Vishakha Bhosale
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Nigdi, Maharashtra

Ketan Desale
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Nigdi, Maharashtra

ABSTRACT
Multiple object tracking is being used for many applications nowdays such as automated surveillance, Robotics,self driving cars,medical and many more. There have been continuous improvements in existing state of art MOT(multiple object tracking) methods through many methods and global optimization techniques.This paper focuses on various MOT techniques and how to achieve speedup and efficiency using MOT methods.

General Terms
Hadoop, MapReduce, Greedy Algorithm, K-means algorithm, Deformable Part-Based Models

Keywords
Multiple Object Tracking (MOT); Parallel Systems; Hadoop; MapReduce

1. INTRODUCTION
Video analytic application in this era has became an extremely important part in many areas such as video surveillance,medical,robotics,self driving vehicles etc. Of various video analytic components,Multiple Object Tracking(MOT) plays an important role in engaging area of research . MOT consists of two stages: first is object detection stage and the other one is object association stage. In object detection stage spacial location of the objects is being identified that are present in a video sequence which is basically the time independent process while the[1] object association stage receives the out from from the object detection stage and associate the detections by assigning some unique identity for all the detections from object detection stage. Object association is time dependent process and handle the inter frame dependencies. Thus association stage of MOT produces a set of continuous object trajectories across the frames where each trajectories represent a single object detections, therefore dependencies between the frames will get introduced. There are number of State of art object detection algorithms present ,some of them has been illustrated in [2,3,4] and data association for the detected object is being illustrated in [5,6,7] which provides good qualitative performance for MOT(multiple object tracking) these algorithms mainly focuses on accuracy of MOT that MOTA(multiple object tracking accuracy). Since the generation of videos with high speed sometimes makes the analytic algorithms computationally slow therefore for analytic applications distributed and parallel platforms are fast growing specially from last five years various methods have also been published for parallel and distributed platform based computing papers[8.9,10] illustrated these methods for distributed platforms.

2. BACKGROUND STUDY
Video Analytics is the growing trend of this era where many videos are generated through many sources such as cctv footage or from any sources to elicit our desired footage or say desired video from millions and billions of videos it becomes the cumbersome job for everyone for the purpose of handling the huge number of videos video analytics have came in picture. Various researches are still going on.

The state-of-the-art object detection methods which build the objects to be detected by iteratively linking the smaller parts of objects. The object detection methods make use of neural networks structure. These object detection methods focus on improving the qualitative performance that concern the accuracy of tracking. The methods in [2, 3, 4] can take 2 to 18 seconds for a single frame depending on the frame resolution using general purpose computers. Similarly, the state-of-the-art data association methods such as network flow-based based [5] and minimum-cost subgraph multi-cut problem-based [6] are computationally slow, exhibiting a processing speed of 0.3 to 2 frames per second (fps) [13]. The data association techniques based on greedy algorithms were proposed in [14, 15]. These methods are able to achieve a higher computational speed, however, the complete MOT methods still have a high computational cost. We have proposed ffmpeg for video analytics applications. A Hadoop-based video processing framework is proposed in [8]. It provides a model for storing video stream to Hadoop Distributed File System (HDFS) and uses the Opencv library to perform video analytics-related jobs. However, this method stores multiple small size videos and processes each of the stored videos independently. Another Hadoop-based video analytics solution is given in [10] that speedup the demonstrated video using multiple small video files. Although the method in [9] also splits a single video file for parallel processing, it does not effectively utilize the Map and Reduce phases of MapReduce for the video processing operations. Instead of utilizing the Map phase for parallelizing the operations a single map task for splitting a video file is used and Reduce phase for video processing operations. The Hadoop-based methods focus on using the Hadoop architecture for storing and processing general video analytics applications. These methods do not describe parallel techniques required for handling of MOT, and do not handle time-dependencies. To address the shortcomings of the

existing literature, we used three MapReduce-based techniques for implementing MOT. The techniques handle time-dependencies connected with distributing a single video file on a multi-node Hadoop cluster. A first measurement and measurement-based performance comparison of MOT on a single node and MOT on a MapReduce cluster have been provided.

3. MATERIALS AND METHODS

This section provides various methods and researches on the MOT(multiple object tracking) and parallel systems.

3.1 Multiple Object Tracking

MOT (Multiple object tracking) is basically a task of automatically locating the object of need and tracking their trajectories in a sequence of a given video .[1] explains that There are two steps in MOT first one is object detection stage and the next one is the data association stage. In object detection stage objects are detected in each frame of a video, here no inter-frame dependencies are there. The output of this stage is a set detections .D={di} where each di = {xi,yi,wi,hi,fi} is an object detected in frame fi. Location of frame is identified by (xi,yi) which are co-ordinates for the left top corner of a rectangular box and (wi,hi) is the width and height. Object detection stage is followed by data association stage which associates the output of detected stage it aims to form the possible object trajectories , this stage gives the Tracklets(object trajectories) S= {tj} where tj = {dj1,dj2.....} which is the sequence of detections across the frame which is partivular for one particular object. Data association stage has inter-frame dependencies because it is formulated as global optimization problem for complete sequence of a video file[5,6].

3.2 Parallel Systems

Since the production of large video data around the world is increasing day by day and at some point of time many algorithms becomes slow for analysis purpose it may take days in some cases . To overcome this problem distributed and parallel platforms are being developed so that analytics work will be distributed. In [8,9,10] these systems are illustrated. In [8] an extensible video processing framework in Apache Hadoop is parallelized a video processing tasks in a cloud environment. Except for video transcoding systems. In[9] video stream acquisition is presented , processing and analytics framework is in the clouds for addressing some of the traffic monitoring challenges This framework provides an end-to-end solution for video stream capture, storage and analysis using a cloud based GPU cluster. In[10] paper proposes an approach for fast and parallel video processing on MapReduce-based clusters such as Apache Hadoop. By utilizing clusters, the approach is able to handle large-scale of video data and the processing time can be significantly reduced.

3.3 Hadoop

Hadoop is a collection of open-source software utilities that uses network of many computers to solve problems involving massive amounts of any kind of data and computation. It provides a framework for distributed storage and processing of huge amount of data using the MapReduce Model. Initially made for computer clusters built from commodity hardware— till the common use—it has also found use on clusters of higher-end hardware.All the modules in Hadoop are designed with an assumption that hardware failures are commonly occur and should be automatically handled by the framework. The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits the files into large blocks and distributes them across nodes in a cluster. It then transfers the code that is packaged into nodes to process the data in parallel. This way of dealing takes the advantage of data locality where nodes manipulate the data they have access to. This permits the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking.

4. MAPREDUCE

MapReduce based techniques implementing for MOT(multiple opbject tracking) on multiple nodes of clusters focuses on processing single video file in parallel simultaniously . In[1] the video file is split into multiple number of chunks and these chunks are processed independently during the Map phase of the MapReduce program. The intermediate results that are obtained from all of these chunks are then combined to produce the continuous final object trajectories in the Reduce phase. In[1] there are two parallel techniques that were proposed which handles the time dependencies by performing data associate techniques either sequencially or in a parallel way .

4.1 Partially parallel Technique

The Partially parallel techniques divides the object detection stage with data association stage .Here time independent which is object detection stage run in parallel and time dependent process which is data association run in a sequence[1]. Here video get split into chunks CS={ci} where during map phase input is a key-value pair<i,ci>, where I is the unique identity. A map task produces a set of object detections di by invoking object detection method that is Detector.getDetections(ci) which is the intermediate output for video chunk ci the output from this task is in the form of key-value pair <s,di> where s is some constant here constants are used so that when these will combine in reduce phase they will be in a proper sequencial manner. the sequential implementation of the data association stage is also a bottleneck for the speedup. The faster the data association method is, the lesser is its effect on the speedup achieved by using multiple processing nodes. Thus, this technique is more useful for greedy data association methods such as [14, 15] which are computationally fast. Algorithm for partially parallel technique is shown in Table 1

Table1. Algorithm for partially parallel Technique

Input = a a video sequence V ,N: number of nodes
1: CS = splitVideo(V, N) 2: Upload CS to HDFS and start Hadoop Job 3: Mapper (< i, c i >): 4: D i = Detector.getDetections(c i) 5: return (ς,D i) 6: end Mapper 7: Reducer (<ς,{D i)>): 8: D=sortAndAppend({D i }) 9:S=Tracker.getTracklets(D)

10: return (‘Trajectories’,S)

11: end Reducer

12: Copy S to local-file-system of Master node

4.1.1 Fully Parallel Technique

In fully parallel technique time independent state and data association state both are parallelized[1] .The map phase implements the full MOT method . Similar to partially parallel technique input for the map task is the key-value pair <i,ci> each chunk that is ci processed in an independent way on single map task set of object detections will be obtained by the command Detector.getDetections(ci) and next stage that is data association state is obtained by invoking the command Tracker.getTraklets(Di) following table shows the Table 2

Table2. Algorithm for fully parallel technique

Input = a a video sequence V ,N: number of nodes
1: CS = splitVideo(V,N)
2.: Upload CS to HDFS and start Hadoop Job
3: Mapper (< i, c i >):
4:D i = Detector.getDetections(c i)
5:S i = Tracker.getTracklets(D i)
6:return(ς,S i)
7: end Mapper
8: Reducer (<ς,{S i } >):
9:sort({S i })
10: S = S 1
11:for i = 2 to N Do:
12.Tracker.combine(S, S i)
13:end for
14:return (‘Tracklets’,S)
15: end Reducer
16: copy S to local-File-System of Master node

5. MAPREDUCE BASED TECHNIQUE FOR MOT

Fig 1 represents mapreduced based technique .[1]Here a video file V is split into set of chunks of multiple videos and uploaded to Hadoop distributed file system . Here f is the total number of frames in a video V then video chunk ci will have fi number of frames .Here each of the map task operates on one of the chunks independently. The intermediate outputs from these map tasks are thencombined by the reduce task. Although an object trajectory detected by a map tasks remains continuous for the respective chunk a discontinuity will be introduced when the trajectory spans across subsequent chunks and each of these chunks is processed independently. So to generate continuous output trajectories for the same object in the video file V, a single reduce task is to be used for combining the intermediate object trajectories. Also, note that higher the number of chunks a video file is split into, more is the overhead of combining the intermediate results of these chunks in the reduce task. It is assumed that initially, the video file V is available on the master node of the Hadoop cluster.

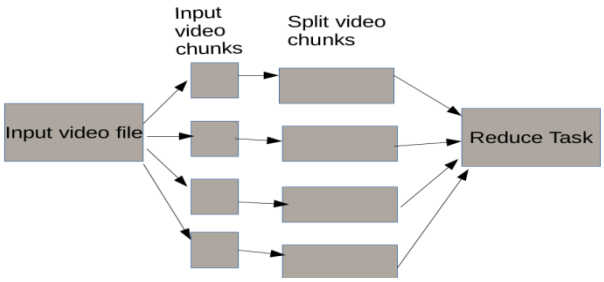

Fig1. MOT using MapReduce

6. RESULTS AND DISCUSSIONS

In [1] the qualitative performance of the video datasets that reflected the accuracy of tracking was analyzed by using performance matrices presented in [16] that will be presented next:

Number of False Positives (FP): The occurrence of a target object in the output of a tracking algorithm when it is not available in the ground truth is considered a false positive and increments FP by 1.

Number of False Negatives (FN): If the target object is available in the ground truth but the tracking algorithm misses to locate it, then it is considered a false negative and FN is incremented by 1.

Number of Identity Switches (IDS) : IDS is the total number of discontinuities in the object trajectories generated by a tracking algorithm in comparison to continuous object trajectories in the ground truth. The object trajectory is discontinuous if the same object is assigned with two or more different identities.•# Multiple Object Tracking Accuracy (MOTA): It is a measure that combines FP, FN, and IDS and is given by:where t is a frame index (a sequential ID of a frame in a video sequence) and GT t is a total number of objects in frame t of ground truth [16]. MOTA provides the overall accuracy of tracking algorithms. MOTA = 1-∑t(FNt+FPt+IDSt)/∑tGTt .

Table 3: Instructions obeyed during annotations in [16].

Instruction What? Targets: All upright people including
+ walking, standing, running pedestrians
+ cyclists, skaters
Distractors: Static people or representations

+ people not in upright position (sitting, lying down) + reflections, drawings or photographs of people + human-like objects like dolls, mannequins
When? Start as early as possible. End as late as possible. Keep ID as long as the person is inside the field of view and its path can be determined unambiguously.
How? The bounding box should contain all pixels belonging to that person and at the same time be as tight as possible.
Occlusions Always annotate during occlusions if the position can be determined unambiguously. If the occlusion is very long and it is not possible to determine the path of the object using simple reasoning (e.g. constant velocity assumption), the object will be assigned a new ID once it reappears.

In[16]to represent the any of the entire person and to estimate the level of cropping. If an occluding object cannot be accurately enclosed in one box (e.g. a tree with branches or an escalator may need a large bounding box where most of the area does not belong to the original object), then many boxes may be used to better approximate the extent of that object. Persons on vehicles will only be annotated separately from the vehicle if clearly visible. For example, children inside carriage or people in the cars will not be annotated, while motorcyclists or bikers will be.

6.1 Detections

In paper[16] it is being tested various state-of-the-art detectors on our bench- mark, obtaining the Precision-Recall curves out-of-the- box R-CNN outperforms DPM in detecting all object classes except for the class "person", which is why they supplied DPM detections with the benchmark. they used the already trained model with a low threshold of −1 in order to maintain relatively high recall. It was noted that the recall did nor reach 100% because of the non-maximum suppression applied. A detailed breakdown of detection bounding boxes on individual sequences is provided in Table. 2.

Table 4: Detection bounding box statistics in paper[16]

Seq	nDet.	nDet./fr.	min height	max height
MOT16-01	3,775	8.39	19.00	258.92
MOT16-02	7,267	12.11	19.00	341.97
MOT16-03	85,854	57.24	19.00	297.57
MOT16-04	39,437	37.56	19.00	341.97
MOT16-05	4,333	5.18	19.00	225.27
MOT16-06	7,851	6.58	19.00	210.12
MOT16-07	11,309	22.62	19.00	319.00
MOT16-08	10,042	16.07	19.00	518.84
MOT16-09	5,976	11.38	19.00	451.55
MOT16-10	8,832	13.50	19.00	366.58
MOT16-11	8,590	9.54	19.00	518.84
MOT16-12	7,764	8.63	19.00	556.15
MOT16-13	5,355	7.14	19.00	210.12
MOT16-14	8,781	11.71	19.00	258.92
total	215,166	19.15	19.00	556.15

6.2 Data Format

In[16]a ll images were converted to JPEG and named sequentially to a 6-digit file name . Detection and annotation files were simple CSV files. Where each line is represented by one object instance and contains 9 values as shown in . The first number indicates in which frame the object appears, while the second number identifies that object as belonging to a trajectory by assigning a unique ID (set to −1 in a detection file, as no ID is assigned yet). Each object can be assigned to only one trajectory. The next four numbers indicatd the position of the bounding box of the pedestrian in 2D image coordinates. The position was indicated by the top-left corner as well as width and height of the bounding box. This was followed by a single number, which in case of detections denotes their confidence score. The last two numbers for detection files are ignored (set to -1).

7. CONCLUSION AND FUTURE WORK

In [1] efficiency of the partially parallel . fully parallel and fully parallel with overlapping are demonstrated through prototyping and measurement of performance which was done in Amazon EC2 cloud. Following are the performance evaluation for that:

- Tracking accuracy: No matter what's value of N and association method being used pp technique gave best tracking accuracy.
- Fast data association: partially parallel technique was better because by increasing the number of nodes accuracy did not deteriorate. But there was significant difference in fully partial technique with overlapping and there was small change in fully parallel technique and partially parallel technique for datasets used in[1].
- Slow data association with small video files: fully parallel system outperforms better for small video files and data association done for that in comparison to partially parallel technique and full parallel technique in [1].
- slow association of data and large videos: fully parallel technique gave more computational speed in comparison to partially parallel technique used and fully parallel technique with overlapping gave more qualitative performance then fully parallel technique.

Overall results gathered by experiments in[1] says that fully parallel technique performs better then partially parallel technique when slow data association method is being used for MOT similarly the case is reverse in case fast data association here partially parallel technique outperforms fully parallel technique. Basically fully parallel technique and fully parallel technique with overlapping are better for slow data association methods and for large video files whereas partially parallel techniques are better for fast data association methods and small video files.

Paper[1] says that future includes automatic technique for configuring the number of nodes of hadoop cluster also for handling live stream videos. Future work can done to automate the diving work for splitting the videos according to number of nodes in the hadoop cluster.

8. REFERENCES

[1] Gurinderbeer Singh, Shikharesh Majumdar, Sreeraman Rajan,"MapReducebased Techniques For Multiple Object Tracking in Video Analytics",Systems and Computer Engineering Department.

[2] P. F. Felzenszwalb et al., "Object Detection with Discriminatively Trained Part-Based Models," IEEE Trans. Pattern Anal. Mach. Intell., vol. 32, no. 9, pp. 1627-1645, 2010.

[3] R. Girshick, "From Rigid Templates to Grammars: Object Detection with Structured Models." Ph. D dissertation, The University of Chicago, 2012.

[4] R. Girshick, "Discriminatively Trained Deformable Part Models,"2012.[Online].Available: http://people.cs.uchicago.edu/~rbg/latent-release5/.

[5] J. Berclaz et al., "Multiple Object Tracking Using K-Shortest Paths Optimization," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 9, pp. 1806-1819, 2011.

[6] S. Tang et al., "Subgraph Decomposition for Multi-Target Tracking," in IEEE CVPR, 2015.

[7] A. Milan et al., "Continuous Energy Minimization for Multitarget Tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 1, pp. 58-72, 2014.

[8] C. Ryu et al., "Extensible Video Processing Framework in Apache Hadoop," in IEEE International Conference on Cloud Computing Technology and Science, 2013.

[9] T. Abdullah et al., "Traffic Monitoring Using Video Analytics in Clouds," in IEEE Conference on Utility and Cloud Computing, 2014.

[10] H. Tan and L. Chen, "An Approach for Fast and Parallel Video Processing on Apache Hadoop Clusters," in IEEE International Conference on Multimedia and Expo, 2014.

[11] R. Girshick et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," in IEEE CVPR, 2014.

[12] R. Girshick, "Fast R-CNN," in IEEE ICCV, 2015.

[13] A. Milan et al., "MOTCHALLENGE," Available: [Online]. https://motchallenge.net/results/MOT16/.

[14] G. Singh et al., "A Greedy Data Association Technique for Multiple Object Tracking," in IEEE International Conference on Multimedia Big Data, 2017.

[15] H. Pirsiavash et al., "Globally-Optimal Greedy Algorithms for Tracking a Variable Number of Objects," in IEEE CVPR, 2011.

[16] A. Milan et al., "MOT16: A Benchmark for Multi-Object Tracking," in arXiv:1603.00831 [cs.CV], 2016.