Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09099-8

Big data platform of traffic violation detection system: identifying the risky behaviors of vehicle drivers

Shiva Asadianfam, et al. *[full author details at the end of the article]*

Received: 20 July 2019 / Revised: 28 March 2020 / Accepted: 22 May 2020
Published online: 23 June 2020

Abstract
Since the traffic data has a high volume, high diversity and high speed of production, the traditional systems cannot process them accurately. In this paper, a big data based system was designed and implemented for identifying the offenders' behaviors. The proposed Traffic Violation Detection System (TVD system) included four main phases which were described using the MAPE methodology. In the monitor phase, unstructured data, such as videos captured by traffic control cameras as well as the images and the descriptions provided by traffic officers, were collected. In the analysis phase, the knowledge base of unsafe driving behaviors was created and classified. In this phase, a standard Work Breakdown Structure for unsafe behaviors was created by the experts in traffic control. In the Plan phase, in order to detect unsafe driving behaviors related to police descriptions for collected images, and to detect unsafe behaviors from video cameras, Behavior-Based Safety process with Map/ Reduce technique and Vector Space Model (VSM) were employed. In the last phase, all types of data, including the structured data and the multimedia/ unstructured data, along with the types and details of violations, were stored on the Hadoop Distributed File System. The prototype of the proposed TVD system was successfully implemented for a few common violations. The results showed that by applying big data technologies, the driving violations could be detected more accurately using the combination of the structured and unstructured data. The results indicate that compared to the sequential program, Hadoop only with a single slave-node decreases the processing time of big data by more than 70%. Also, by increasing the number of slave nodes from 1 to 7 in the police descriptions and images of surveillance cameras, the processing time reduces by 60.87% and 70%, respectively. Thus, the TVD performance increases by more than 75% as the number of data nodes boosts. Based on the results, it can be concluded that by identifying unsafe driving behaviors, it is possible to diminish traffic accidents and the damage caused by them at a satisfactory level. Also, the authors decided to compare these studies in terms of qualitative criteria, such as fieldwork and behavioral identification, and quantitative criteria.

Keywords Big data · Behavior-based safety · Risky behavior · Traffic violation detection · Traffic surveillance cameras

1 Introduction

Dealing with driving offenses and road casualties is deemed as a significant issue in social and public management all over the world [59]. According to World Health Organization (WHO), in 2015, 1.25 million people died in accidents [44]. Therefore, because of the risky behavior of drivers, the high rate of driving incidents leading to financial loss and fatalities cannot be ignored. It is required to pay particular attention to unsafe driving behaviors in order to reduce the social cost of road injuries and casualties [50]. The study conducted by Rakotonirainy et al. indicates that violations and aggressive driving behaviors can be regulated via safety measures taken by drivers themselves [50]. Based on the behavior-based safety (BBS), humans are the primary cause of accidents and 88% of accidents occur due to unsafe human actions. Consequently, identifying and redressing behavioral problems during accidents can be an effective approach in helping people show safe behaviors [25]. Various studies demonstrate that the BBS has been widely used by various industries [11], such as the oil and gas [28], production [43, 62], the nuclear energy industry [15], and the construction industry [35, 36].

In order to overcome behavioral problems the existence of intelligent video surveillance systems is automatically considered a deterrent to malicious behavior. Identifying and analyzing unsafe driving behavior can be a step towards reducing the damages and accidents caused by misbehaving drivers. However, a model can be designed to modify the traffic behavior of drivers. Based on what was mentioned, this study is designed to propose and implement a big data based platform of violation detection system for identifying risky driving behaviors by Behavior-Based Safety process so that using this information, the traffic organization can execute certain strategies and techniques to improve the safety of drivers. In this system, a database of risky behaviors is generated to classify and encode observed behaviors. Then, smart video surveillance systems and related software are utilized to collect visual data from streets. Hence, it is possible to convert large images of behavior observations into specific information. Finally, the data can be stored on the Hadoop Distributed File System.

Reviewing the studies carried out in the field of transportation, especially, the studies conducted on surveillance camera images, shows that there is a dearth of research on big data using the Map/Reduce techniques. Due to the availability of various data sources, including structured data (relational database) and unstructured data (video, images, etc.), a framework is required for processing and storing them. Unfortunately, the work done due to the lack of combination of structured and unstructured data is still not sufficiently accurate and processing time is also important with regard to data composition.

Therefore, this study investigates the existing methods of analyzing urban traffic images, compares them, and extracts patterns, and utilizes those patterns to analyze the images and the data obtained from the urban traffic control centers, which are referred to as big data, based on the BBS.

The main questions that we attempt to address in this paper are:

- Is it possible to design a big data platform for implementing the Traffic Violation Detection System?
- Can the proposed system overcome the high volume and high diversity of the structured and unstructured input data?
- Can the proposed system have sufficient accuracy, precision and recall using a combination of structured and unstructured data?
- Will the Traffic Violation Detection System be effective in speeding up the processing time of big data, using map/reduce technique?

To answer the above questions, we define some objectives. We start by designing the big data platform for TVD System. The system includes four main components: Monitor, Analysis, Plan and Execute (MAPE methodology). A Behavior-Based Safety process is used to generate a violation knowledge base and the corresponding standard Work Breakdown Structure was created and coded. Then, all images and videos captured by traffic cameras and the supervised version collected by police officers along with their description, are gathered. After analyzing these unstructured data using image processing and machine learning techniques, they are ready to be stored in the Hadoop File System. The last component is designed to store all collected data based on the Hadoop Distributed File System. The other objective is to demonstrate that the proposed system can detect the driving violation more accurately than the traditional systems. Finally, the purpose of the TVD system using BBS process by MapReduce technique is to lessen errors considerably, increase speed, and keep maintenance costs very low, all of which make using it in traffic control centers to detect the risky behaviors of drivers beneficial.

In this paper, in Section 2, the literature review on BBS and traffic control are presented. In Section 3, the proposed big data-based traffic violation detection system is explained, and then the way the data function and are classified is explained. Afterwards, the way the multimedia and unstructured data are amassed, along with the software organized for the structured data, is expounded. After that, the steps of storing the obtained data on the Hadoop are described. In Section 4, the implementation of TVD system is described. In Section 5, the experimental result is described, and then a list of dos and don'ts is provided based on the findings. Finally, in Section 6, a general conclusion is drawn.

2 Literature review

With the growing population in big cities, traffic has become an unsolvable problem. Meanwhile, the main factor of street accidents is the risky behavior of drivers. The incremental increase in driving violations and their harmful effects are one of the most important social issues in different countries. To improve the driver's risky behaviors, the type of risky behaviors, the contribution degree of each behavior in an accident and their influential factors by BBS should be identified. Since the traffic data has a high volume, high diversity and high speed of production, the traditional systems cannot process them accurately. So, this section examines the research literature on BBS and traffic control.

2.1 Literature review on BBS

Behavior-Based Safety (BBS) is the "application of science of behavior change to real world safety problems". BBS focuses on what people do, analyzes why they do it, and then applies a research-supported intervention strategy to improve what people do" [25]. The rate of traffic violations is very high and their catastrophic consequences have made traffic violations be regarded as one of the most important social issues in countries. Traffic violations, as a type of behaviors occurring while applicable traffic laws and regulations are violated, are incrementally increasing [1]. Due to the fact that some unseen violations and a majority of observable violations cannot be recorded by traffic officers, it is too difficult to provide accurate statistics regarding the precise number of violations. In this regard, the strict enforcement of rules and the use of the latest technology and scientific achievements can be helpful. Therefore, investigating drivers' unsafe behaviors and identifying the violations committed are effective steps in improving driving behaviors, reducing traffic accidents, and boosting urban community well-being. Traffic accidents occur due to the interactions between the vehicle's driver (the human factor) and the road and surrounding environment [48]. Based on the studies on the contribution degree of each influential factor in the collisions in the industry [41], sea accidents [9], and driving accidents [46], it has been found that the human factor alone or in combination with other factors plays the most important role in causing accidents.

Violations refer to the behaviors that endanger driving safety, including over-speeding or driving without keeping a safe distance from other vehicles [29]. Unintentional violations are the behaviors resulting in the violation of laws inadvertently, such as driving on a narrow two-lane highway slowly. Deliberate violations are the behaviors being intended to harm another and to violate laws and being regarded as a form of sabotaging behavior [51].

Directing attention to driving behaviors is of crucial importance because aggressive driving behaviors have caused enormous physical damage to the structure of countries [5]. Several studies on influential human factors in the accident rate have come to the conclusion that in order for a theoretical framework to provide a proper and precise explanation of accidents, it is required to differentiate between errors and violations. The BBS process consists of three steps as follows [20]: initially, unsafe behaviors are listed; then, drivers' violations are observed and their frequency is recorded; and in the end, the given feedback (penalties) and unsafe driving behaviors change. In this way, it is possible to gain satisfactory results and reduce accidents and injuries caused by unsafe driving behaviors.

2.2 Literature review on big data platform for traffic control

Big data analytics introduces new methods and technologies for collecting, storing and analyzing unstructured data on a scalable basis [12, 18, 55]. The International Data Corporation (IDC) defines big data as the cost-effective extraction of value from high-volume and high-variety data that are received, discovered, or analyzed at high speed [38]. In recent years, with the growth and development of the automotive industry in countries, and due to the growing number of vehicles on streets and roads, problems pertaining to traffic control and to the enforcement of driving laws have increasingly been addressed. Also, owing to the installation of different cameras at jam intersections and on roads, a large amount of data related to road user behavior on roads and streets are obtained every day. This data set, albeit not all of it, is rarely examined for unsafe acts (driving violations). Statistics indicate that after cardiovascular disease and cancer, road accidents are the third leading cause of death in many countries and are considered as a major health, social, and economic risk factor [3, 56]. Investigating the correlation between safe behaviors and workplace safety, Reber et al. concluded that injuries reduced via increasing the percentage of safe behaviors [52]. Consequently, if unsafe behaviors can be controlled through observation processes with effective interventions, the probability of devastating incidents can significantly decrease. This section provides a review of systems for unsafe behavior detection in traffic control.

In article [45], a system is designed and implemented for the real-time detection of traffic violations. This paper presents a diagnostic algorithm that can identify various types of violations occurring on roads and in parking lots. In order to conduct real-time analysis, parallel computing techniques are used. In [32], it is stated that the risky behavior of lane changing may occur before stop lines in the range of red lights. The system used in this article can detect red light violations and lane change violations. The experimental results of this paper show that the above-mentioned DSS recognizes both violations with high precision. In [24], a system identifies unsafe behaviors using RFID. Through RFID technology, vehicles are connected to computer systems, smart lights, and other existing hardware along the way. The article declares that this intelligent control system is capable of tracking all vehicles, controlling and managing crises, managing traffic congestion, and recording road traffic accidents on highways. In article [2], the framework and components of an experimental platform for an advanced driver support system are described. The purpose of this system is to give drivers feedback on their traffic violations. The system is able to identify certain traffic violations and record the details of these violations in a local database, which makes it possible to display the spatial and temporal data of traffic violations on a geographic map using the standard Google Earth tool. In article [4], a new approach to modeling traffic behaviors is proposed for video-based road monitoring. In the proposed system, the Pachinko allocation model (PAM) and the backup vector machine (SVM) are combined to display hierarchies and identify traffic behaviors. Compared to the conventional model of latent Dirichlet allocation (LDA), the proposed model shows more flexibility and greater expressiveness, which enables it to provide a more accurate diagnosis of behavioral classification. In article [8], the focus is on the simulation and evaluation of a system for traffic congestion detection. This system is a combination of communication among vehicles, fixed infrastructure, and infrastructure communications with big data. The system discussed in this article helps drivers to identify the location of traffic jams and consequently change their direction. The results obtained from the validation of the performance of this traffic detection system verify its positive effect on detecting, reporting and redirecting traffic at the time of traffic accidents. In article [42], a system is proposed for the real-time identification of highway lanes. The proposed scheme uses both lane information and vehicle information to improve the reliability of its detection. In this system, initially, lanes are identified, and then vehicles surround the vehicle under study. In article [63], a driver assistant system is proposed based on image processing techniques. A camera is located in the front window of cars to identify lanes and to determine the position of those vehicles relative to the lanes. The consecutive images taken by the proposed system are analyzed and processed, and the lanes are automatically detected. The experimental results of this paper suggest that the system functions well in terms of lane identification. Also, in articles [23, 37] in the big data domain are used the BBS to identify unsafe behaviors of individuals in the construction and illegal construction waste dumping. So, each of which are explained. In article [23], Guo et al. introduce a big data-based platform for identifying workers' safe behaviors in a metro construction project in China. In this study, the Behavior-Based Safety (BBS) is used to observe, analyze and change the behavior of the workers in the construction site. Their paper is a case study to develop a big data platform for classifying, collecting, and storing the data related to the unsafe behaviors of construction workers. The findings reveal that the platform can effectively analyze the semantic information of the images and can automatically extract the unsafe behaviors of the workers. In article [37], Lu presents a big data platform for identifying the unauthorized behaviors of urban drivers in construction waste discharge in Hong Kong in order to manage construction waste and combat urban crimes. The data collected from nine million cases of waste discharge from 2011 to 2017 are analyzed. Using the behavioral indicators and the updates of big data analyses, the potential drivers for illegal evacuation (for example, long loads) are identified. The results reveal a list of 546 waste collection trucks suspicious of being involved in illegal discharges. Therefore, using the big data and BBS, it is possible to monitor drivers' violations via the data gathered and stored through observation.

3 The proposed big data based platform of traffic violation detection system

In order to resolve the problems described in Sections 1 and 2 as well as to overcome the limitations of conventional methods regarding behavior observations, a big data platform has been devised using advanced techniques. MAPE methodology [14] has been used to design the proposed platform in the current study. The MAPE methodology is a model consisting of Monitor, Analyze, Plan and Execute phases, which is used in self-adaptive systems. In principle, the monitor phase is responsible for collecting data. In the analysis phase, the system analyzes the data and argues about the symptoms presented by the monitor phase. When the analyst specifies, the system invokes functions in the plan phase to create a method of actions used to apply a desired match. Finally, in the execute phase, the behavior of the system changes according to the performance recommended by the program [27, 30]. Figure 1 shows the phases of the proposed platform that describes the actions used in each phase in detail.

Phase 1 (monitor)
The data are obtained from the supervisory videos all over the city and the traffic images of the past violations of drivers (along with subtitles of unsafe behavior) provided by traffic operators.

Phase 2 (analysis)
In this step, all the driving violations (unsafe behaviors) using safety driving standards, and experts' experience are amassed. Considering the expertise and experience of experts and the required work breakdown structure (WBS) standards, a driving violation WBS is generated to classify and encode the traffic violations of drivers.

Phase 3 (plan)
In order to detect traffic violations of drivers related to police descriptions for collected images, and also to identify unsafe behaviors in videos taken from traffic cameras, Behavior-Based Safety (BBS) process is used with MapReduce techniques. The images associating with drivers' unsafe behaviors are compared with all traffic violations (unsafe behaviors), which are committed in all parts of the country and stored in the database, using the vector space model (VSM). The VSM is a model which ranks documents based on the relative significance of the terms in them. This model is deemed as a standard technique for information retrieval. Through the VSM, it is possible to identify which documents are similar to each other or to the keywords being searched [54]. Therefore, the keywords in the description of traffic officers can be organized by this software. This phase aims to detect drivers' behavioral risk. Also, in order to analyze the unsafe behavior of images taken by surveillance cameras with BBS and the identification of any of BBS components unsafe driver behavior should be detected. The proposed approach can analyze and detect drivers' violations and their high-risk behaviors automatically. The detection of violations using this method requires the application of new techniques to the stored big data. This phase aims to detect drivers' behavior.

Phase 4 (execute)
Structured and unstructured data are stored on the Hadoop distributed file system [60]. The HDFS is designed for very large files being accessible in a streaming manner and runs on common hardware; in other words, this file system can appropriately handle multi-gigabyte up

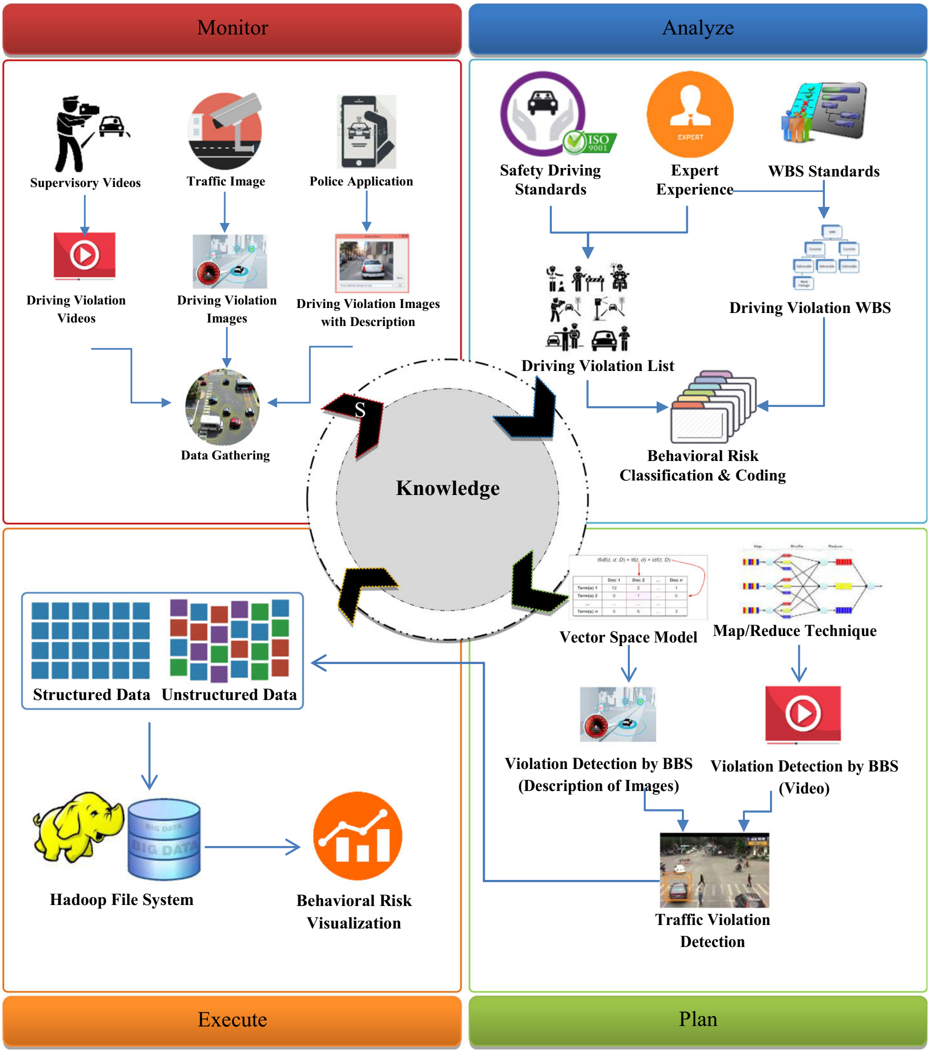


Fig. 1 The big data system to detect drivers' risky behaviors

to multi-terabyte files that are once written and repeatedly read. Since it can run on regular hardware and does not need any specific hardware, this file system itself should be able to overcome any disk or node breakdowns [61]. This phase aims to store and visualize drivers' behavioral risk.

3.1 Monitor

The data are obtained from the supervisory videos all over the city and the traffic images of the past violations of drivers (along with subtitles of unsafe behavior) provided by traffic operators. In order to investigate and test the proposed TVD system, the information provided by the cameras in the city is utilized. In the present study, this information was obtained from the

Traffic Control Center in Yazd, Iran. The traffic control center in Yazd Province requires adopting sophisticated approaches and strategies due to its vast area and the large number of vehicles in it. In order to enhance urban management's control over all important events and incidents, it is critical to improve traffic control systems and traffic violation detection systems. In the current study, the data recorded by several connected cameras being at the specified height were collected (Between December 3 and 20, 2018). The data were provided by the traffic control center. Unsafe behavior detection may include a consideration of the following factors.

3.1.1 Traffic cameras

Nowadays, given the incessant decrease in security levels in communities, security authorities carry a special burden of responsibility compared to the past. Security measures are always supposed to guarantee the provision of safe and peaceful bedrock for more effective and faster management of various affairs and for the provision of more comfort and well-being in working and living environments; as a result, video surveillance is deemed as an integral part of modern security systems. The term 'Video surveillance system' is an umbrella term used for any closed-circuit camera system. This technology usually involves object motion/detection, object classification, object tracking, person identification, and information integration [6].

In general, a majority of scenes are captured; however, due to the lack of time, not all of them can be watched and analyzed. As a result, the recordings of incidents and activities are lost and suspicious behaviors are not detected while being occurred; hence, almost no prevention measure is adopted. Based on what was stated, video surveillance is considered as an effective tool for revealing mysteries behind illegal actions and as a deterrent to unsafe behaviors.

Smart video surveillance systems automatically analyze the recordings taken by cameras and use the information obtained from the analysis. In addition, smart video surveillance systems allow users to check risky situations. The application that makes such an analysis possible is "Video Content Analysis" (VCA) or video analysis (VA) [49].

Drivers' unsafe behaviors can automatically be identified through video surveillance cameras. In this video surveillance analysis of behaviors, various types of functions, such as road detection, sidewalk detection, vehicle detection, traffic sign detection, pedestrian crossing detection, and the identification of violation types are implemented. Therefore, traffic violation detection system is automatically organized.

3.1.2 The appropriate software for the images taken at different locations (police application)

Traffic Fines and penalties refer to notices issued by law enforcement officers because of the violations of traffic laws. Police officers are sent to locations where the probability of accidents is high in order to control and deter violators. Moreover, so as to record violations, they identify and fine violators. Proved by Traffic Control Center, permanent traffic officers, who are selected and receive the necessary training in how to identify traffic and transportation violations, are authorized to detect violations based on traffic laws and to fine violators. Currently, officers subtly control traffic from the Traffic Control Center using cameras and penalize violators; therefore, it is not required for traffic officers to be present in streets. Based on the law of the management of unsafe behaviors of offending drivers, while violators are fined in a surrendered manner or through cameras, traffic officers send the reports of violations in a predefined table, including the drivers' number and type of driving license, to the relevant traffic control office.

However, an officer may inadvertently or deliberately make a mistake and may fine a vehicle erroneously, which can only be the result of a human error in observation. Therefore, it is required to exploit the penalties and images associated with violations along with smart video surveillance. In this way, in order to determine drivers' violations and unsafe behaviors based on captured images and given penalties, certain approaches should be adopted so that the extracted meanings can be verified, and consequently, inappropriate behaviors would be identified.

3.2 Analyze

In order for drivers' unsafe behaviors to be detected by the traffic control cameras automatically, drivers' unsafe behaviors are classified into different categories based on the list of all the driving violations (inappropriate behaviors) using safety driving standards, and experts' experience. In a meeting held on September 25, 2011 at the suggestion of the Ministry of the Interior, the Ministry of Justice, and the Ministry of roads and urban development, the ministers of the Commission on Social Affairs and the Government of the Republic of Iran verified the violations on the basis of Article (21) of the Law on Driving Misdemeanors approved in 2010, and in compliance with the Approval of Letter No. 164082/T373H on December 31, 2007. These violations pertained to traffic and transportation and were committed in different parts of the country and free trade-industrial zones. They were listed in a table (in 171 rows), approved by the "Government Office" seal, and were used as rules in this study. In addition, to identify risky behaviors, several thousand incidents that happened during the year were investigated.

A total of 171 unsafe behaviors were identified and classified as follows: speeding (01), overtaking (02), crossing prohibited places or red lights (03), displaying spiral and theatrical movements (04), being drunk while driving (05), moving with rear gear (06), stopping in prohibited places or causing blockage (07), (not) installing plaques or similar equipment (08), not fastening safety belts (09), talking on mobile phones (10),carrying corpses with an unauthorized vehicle (11), hugging children or animals while driving (12), not having driving license or insurance (official documents) (13), eating and drinking (14), not having safety equipment and a spare tire (15), being inattentive to the forward roadway (16), violating the passenger loading rules (17), paying no attention to police orders (18), neglecting the rules related to automotive components (19), and driving with full beam headlights or with open doors (20).

However, we need to convert unsafe behaviors to the Work Breakdown Structure (WBS). In order to discover unsafe behaviors, the traffic system is divided into four sectors as follows: (1) unit work, (2) subunit work, (3) division work, (4) subdivision work. A part of the WBS for unsafe driving behaviors is presented in Fig. 2. Unsafe driving behaviors include six parts of subunit work, and in particular, automobiles encompass 20 parts of division work, including speeding, parking, overtaking, talking on mobile phones, and paying no attention to police orders. In the WBS designed for this study, one code was defined for each unsafe behavior (or violation) and all of them were provided in the form of a database.

An example of the partial unsafe behavior of 'stopping in prohibited places or causing blockage' classification is presented in Table 1.

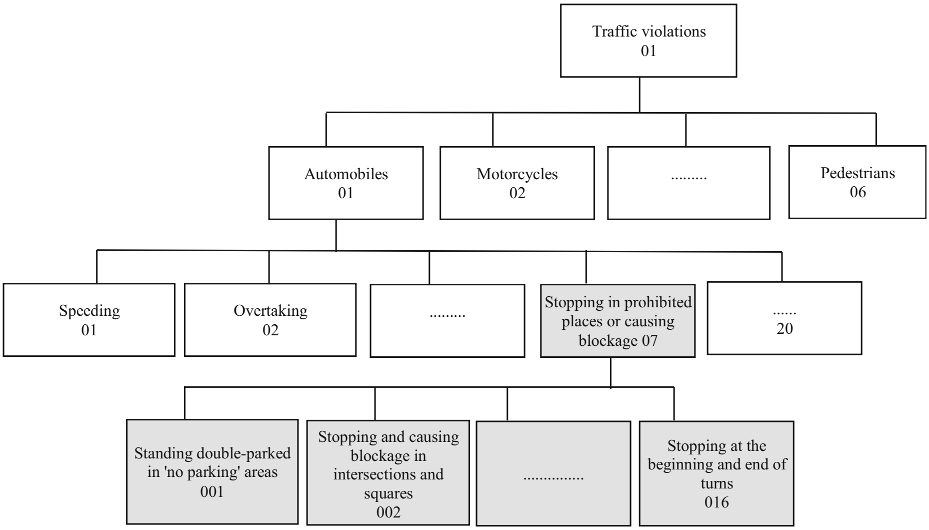


Fig. 2 A part of the WBS

Stopping in prohibited areas includes 16 parts of subdivision work, including standing double-parked in 'no parking' areas, stopping and causing blockage in intersections and squares, and stopping at the beginning and end of turns. In this way, job duties (regarding traffic violation detection) are divided up. As you can see in Table 1, the factors related to each unsafe behavior were extracted using BBS [22]. In the BBS, six factors are considered, such as determining the type of drivers' behaviors, objects (vehicles), resources (including traffic signs), locations, and nearby objects and actions. These factors have been extracted with the help of traffic specialists.

The unsafe behaviors being classified are encoded based on the following rule: XX-XX-XX-XXX (unit work, subunit work, division work, subdivision work, behavior types, and unsafe behaviors). For example, the unsafe behavior of "stopping and causing blockage in intersections, squares, or checkered surfaces" is encoded as '01–01–07-002'. The risky Behavior Database is comprised of the encoded items for the behaviors being observed. Thus, the output of the Analyze phase was created by following the steps described above.

3.3 Plan

Drivers' violations occur in various driving situations. By using the BBS, the following six points should be extracted from the images taken by traffic cameras in order to confirm a violation [21]: Determining the type of drivers' behaviors, objects (vehicles), resources (including traffic signs), locations, and nearby objects and actions. The type of drivers' behaviors refers to forbidden actions, such as stopping or crossing prohibited areas, done by light or heavy vehicles or motorcycles on streets, roads, or highways. Moreover, nearby objects and actions refer to those objects and actions that prompt violations or high-risk behaviors of drivers. Figure 3 shows a violation of 'standing double-parked' next to another parked car in the street. In Fig. 3, six influential factors can be analyzed: considering the parked vehicle (source) in the street (location), the vehicle's (object) standing double-parked; the driver's behavior is identified as a violation.

Table 1 Example of stopping in prohibited places or causing blockage classification as unsafe behavior

Unsafe behaviors	Behavior Code	Potential Damage	Unsafe behavior-related factors
Standing double-parked in 'no parking' zones	001	Traffic jams	**Action**: prohibiting double-parking **Object**: vehicles **Location**: streets **Sources**: Traffic signs **Nearby object**: Vehicles **Nearby action**: parked vehicles
Stopping and causing blockage in the city, intersections, squares, or checkered surfaces	002	Accidents	**Action**: causing blockage **object**: vehicles **Location**: Intersections and squares or checkered surfaces **Sources**: Intersections and squares or checkered surfaces **Nearby object**: no **Nearby action**: no
Not stopping vehicles when facing school vehicles picking up and dropping off students	003	Accidents with students	**Action**: Not stopping **Object**: vehicles **Location**: streets **Sources**: School vehicles **Nearby object**: Students and School Vehicles **Nearby action**: stopping school vehicles
Stopping in 'no parking' zones (stopping absolutely prohibited)	004	Congestion and blockage	**Action**: Stopping or parking in prohibited areas **Object**: vehicles **Location**: streets **Sources**: parking prohibited road signs **Nearby object**: no **Nearby action**: no
Stopping public passenger or cargo vehicles outside the designated stations	005	Traffic jams	**Action**: Stopping in prohibited areas **Object**: vehicles **Location**: streets **Sources**: lack of traffic signs **Nearby object**: no **Nearby action**: no
Stopping other vehicles at public stations	006	Blockage	**Action**: Stopping in prohibited areas **Object**: vehicles **Location**: streets **Sources**: traffic signs **Nearby object**: no **Nearby action**: no
Stopping vehicles on the way of vehicles in motion and on roadside shoulders	007	Traffic jams	**Action**: Stopping on the way of vehicles in motion **Object**: vehicles **Location**: Roads

Table 1 (continued)

Unsafe behaviors	Behavior Code	Potential Damage	Unsafe behavior-related factors
			Sources: Roadside shoulders **Nearby object**: no **Nearby action**: no
Standing double-parked in passageways	008	Blockage	**Action**: Standing double-parked **Object**: vehicles **Location**: Passageways **Sources**: Stop lines **Nearby object**: Vehicles **Nearby action**: parked vehicles
Stopping/parking vehicles on the edge of roads for selling goods	009	Accidents Congestion	**Action**: Stopping illegally **Object**: vehicles **Location**: Roads **Sources**: Roadside Lines **Nearby object**: no **Nearby action**: no
Repairing or washing vehicles on public roads	010	Blockage	**Action**: washing vehicles **Object**: vehicles **Location**: Public roads **Sources**: no **Nearby object**: no **Nearby action**: no
Any types of stopping that lead to blockage or traffic interruptions	011	Blockage	**Action**: traffic interruptions **Object**: vehicles **Location**: Streets and Sidewalks **Sources**: no **Nearby object**: no **Nearby action**: no
Stopping vehicles on sidewalks	012	Blockage	**Action**: stopping on sidewalks **Object**: vehicles **Location**: sidewalks **Sources**: no **Nearby object**: pedestrians **Nearby action**: no
Crossing or stopping on passages at times when stopping/crossing is prohibited	013	Congestion	**Action**: Stopping or crossing areas included in traffic plans **Object**: vehicles **Location**: streets in the designated range of traffic plans **Sources**: traffic control panel **Nearby object**: no **Nearby action**: no
Stopping in 'no parking' (parking prohibited) areas	014	Blockage	**Action**: parking in prohibited areas **Object**: vehicles **Location**: streets

Table 1 (continued)

Unsafe behaviors	Behavior Code	Potential Damage	Unsafe behavior-related factors
			Sources: parking prohibited signs **Nearby object**: no **Nearby action**: no
Parking vehicles on roads with their engine running and no driver inside	015	Air pollution Vehicle theft	**Action**: Parking while keeping the engine running **Object**: vehicles **Location**: streets **Sources**: no **Nearby object**: no **Nearby action**: the running engine
Stopping at the beginning or end of turns, on bridges, within the tunnel, and in the intersections of railways	016	Accidents	**Action**: Stopping at turns **Object**: vehicles **Location**: at the beginning and end of turns, bridges, and tunnels **Sources**: turns, bridges, and tunnels **Nearby object**: no **Nearby action**: no

Based upon traffic surveillance cameras, and police descriptions, unsafe behaviors or violations are identified.

This phase consists of two parts. (A) Identify unsafe behavior by using a police application and determine the similarity of police descriptions to unsafe behaviors in the Traffic Violations Handbook (in section 3.2**)** as detailed in Section 3.3.1. (B) Identifying unsafe behavior by considering surveillance camera video and images as detailed in Section 3.3.2.

3.3.1 Big data analysis and identification of unsafe behavior by using police application

Observing a driver's BBS points (Fig. 3). The descriptions provided by traffic officers in terms of influential components are used to extract effective factors.

In order to determine to which of the 171 behaviors each image with descriptions is similar and to identify the exact type of unsafe behaviors, the following steps are taken in the VSM:

Action	Standing double-parked
Object	Marked vehicle
Sources	nearby vehicle
Location	Street
Nearby object	Nearby vehicle
Nearby action	Nearby vehicle's standing double-parked

BBS Components

Fig. 3 Extracting the influential factors through the analysis of the captured image

(1) All violations are described in terms of the aforementioned structure (Fig. 3). Therefore, it is possible to utilize the vector space model (VSM) to display traffic officers' descriptions using the terms in them. In this way, all the terms related to unsafe behaviors can be identified as a general vector (Table 2).

The vector space model of unsafe behaviors = {factors of violation$_1$, factors of violation$_2$,…, factors of violation$_N$}.

(2) The weight for each term is estimated using the Term-Frequency method and the cosine similarity measure(Table 3[39, 64].

$$\mathrm{COS}(\text{police description, unsafe behavior}) = \frac{<\ \rightarrow\!\textit{police description}\rightarrow\!\textit{unsafe behavior}\ >}{|\rightarrow\!\text{police description}||\rightarrow\!\text{unsafe behavior}|}$$

$$= \frac{(\rightarrow\!\text{police description}) \times (\rightarrow\!\text{unsafe behavior})}{\text{length}(\rightarrow\!\text{police description})*\text{length}(\rightarrow\!\text{unsafe behavior})}$$

(3) The influential factors provided in the descriptions are considered as a query.

(4) The interval between the undertaken query (step 3) and all relevant unsafe behaviors, along with the influential factors in their occurrence, is computed using the database. In this way, the responses are ranked based on the cosine similarity value.

Consequently, the query is checked and compared with the terms in the database. Similar and relevant items to the query are retrieved. The output is ranked based on the calculated similarity. The largest output is identified as the driver's violation. As a result, the image with this structure is saved on the HDFS.

3.3.2 Big data analysis and identification of unsafe behavior by considering surveillance camera video and images

Video data processing is utilized to extract semantic information which can be used to monitor and identify abnormal behaviors. Monitoring and controlling are the leading issues in technology and are undertaken based on the growing needs of communities

Table 2 All unsafe behaviors convert to terms

Documents	convert	Terms
Unsafe Behavior1	→	Term1, Term2, Term3,…,Term6
Unsafe Behavior2	→	Term1, Term2, Term3,…,Term6
...	→	...
Unsafe Behavior N	→	Term1, Term2, Term3,…,Term6

Table 3 Term frequency vector of all unsafe behaviors

		Term Frequency Vector				
		Term1	Term2	Term3	....	Term N
Unsafe Behaviors	Unsafe Behavior1	x	x	x	....	x
	Unsafe Behavior2	x	x	x	....	x
	Unsafe Behavior3	x	x	x	....	x
	....	....	....	....	....	....
	Unsafe Behavior N	x	x	x	....	x

to maintain security and enhance the quality of life. Various studies show how to use video cameras in traffic control [10, 34, 40]. A functional of explanation and the application of cameras in the traffic control can be seen in Table 4.

In this study, the functions listed in Table 4 have been used to extract the effective factors in BBS, which are available in a wide range of researches. For example, 'the identification of traffic signs' and 'the detection of road range' functions were used to identify resources (including traffic signs) and locations (street) factors, respectively. Therefore, the output of the BBS functions as a vector of the officer descriptions is in section 3.3.1, and so the type of driver violation can also be extracted from surveillance camera images.

Camera technology has been used in the automatic analysis of human behavior [22], the transportation industry [26], and the construction industry [13]. One of the problems of designing efficient systems for the semantic analysis of videos is semantic gaps [16]. Semantic gaps are the discrepancy between the concepts extracted by the semantic analysis system of videos and the concepts being understood by humans.

In order to extract incidents and concepts from videos, it is necessary to carry out information processing at different semantic levels. The nature of data processing varies from one level to another, and as a result, the information extracted from each level has a different level in terms of semantics. Based on a general classification, the processes undertaken to extract semantic information from videos can be categorized into three levels as follows: low processes (including border detection between shots, key frame extraction, and type-view detection), intermediate processes (including the disclosure and detection of objects, the disclosure and identification of individuals, the analysis of the motor behavior of objects) and high processes (including incident disclosure and identification, the semantic scrutiny and retrieval of videos). In order to eliminate such semantic gaps between low-level features and high-level concepts, it is essential to extract features based on the introduced levels; in other words, high-level concepts should be extracted from low-level features in several steps [58].

3.4 Execute

In this study, in order to store a large amount of data, including the data obtained from the surveillance cameras in the city and the organized software, the unsafe behaviors of the drivers, and the data pertaining to the image analysis, and to manage them, the HDFS was employed. The HDFS is designed to store very large files across vehicles in large clusters reliably. The

Table 4 Camera function and its application in traffic control

Functions	Description of functions	Applications in Traffic
Accident prediction	Checking current road conditions and the probability of accidents using current data	Providing solutions to avoid accidents
The detection of road range	Installing cameras in new positions, extending the range of the created surface, and approaching them to road structure	Having an automatic traffic monitoring system using mobile video cameras
The identification of the left and right lines of roads	Identifying the range of streets in the monitored area	Having an automatic traffic monitoring system using mobile video cameras
Traffic identification	Identifying traffic conditions in the monitored area	Controlling traffic congestion, identifying less congested pathways
The identification of traffic signs	Identifying traffic signs in the monitored area	Creating and expanding smart systems in the realm of transportation
The identification and count of vehicles	Detecting cars based on the images provided by surveillance cameras, blocking vehicles, and estimating the number of vehicles per unit of length and time	Providing traffic indicators
The identification, tracking, and classification of vehicles	Identifying vehicle types (light vehicles, heavy vehicles, motorcycles) in the monitored area	Detecting the type of vehicles
The identification of traffic signs based on the intersection of crossing lines	Identifying traffic signs in the monitored area	Being used in self-driving vehicles
The detection of critical accident-prompting behaviors	Having a driver tracking system to track upper body movements, eye movements, and external objects using multiple cameras in vehicles for the purpose of determining driving risks (sending short messages, drinking, or grabbing objects while driving)	Detecting drivers' distractions, which plays a major role in accidents
The analysis of urban traffic	Analyzing urban traffic using videos located at signalized intersections	Setting the traffic light timer
The detection of drivers' drowsiness	Measuring the percentage of eyelids' approaching each other and comparing it with the usual distance of them, and if there is a difference between them, activating the warning system via the feedback defined in the system.	Warning about one of the main causes of road accidents (the warning system may warn the driver through a voice message, seat vibrations, or the steering wheel).
The identification of vehicle plaques	Determining the type of plaques	Controlling traffic in specific areas (the range of traffic plans, emergency range, urban areas, etc.)
The identification of getaway people's behavior	Creating a Bayesian model to detect the behavior of getaway people	Detecting abnormal conditions in unexpected accidents promptly
The identification of roles	Using multiple cameras to identify people with specific roles using a cause and effect network	Identifying people in the role of traffic officers
The identification of the background in crowded scenes	Using the RGB component and colored channels to neutralize the effect of uniform lighting, removing	Nothing

Table 4 (continued)

Functions	Description of functions	Applications in Traffic
	shadows, and identifying objects using slow motion and far objects	
Fire detection	Using RGB color spaces	Having Security Supervision
The detection of incidents and abnormal behaviors	Determining dynamic objects in videos	Detecting unusual behavior of people
The identification of traffic signs' meanings	Monitoring vehicles in motion against traffic rules	Being used in self-driving vehicles
Path clustering	Extracting paths in the images taken	Teaching legal behavior and detecting illegal practices

Hadoop distributed file system or HDFS [57] is a distributed file system based on Java, storing big data on multiple nodes in a Hadoop cluster. This file system can store the images taken by the cameras in the city and it is able to generate a database which can be used effectively.

Due to the availability of various data sources, including structured data (time and location, etc.) and unstructured data (texts, videos, images, etc.), a framework is required for their processing and storage.

The driving violation knowledge base is comprised of a list of drivers' unsafe behavior (violation) and the WBS of traffic laws. Application management information includes structured data. The data obtained from the analysis of the images include the aforementioned structured data, such as dangerous behavior (violation), the type of behavior, damage caused by unsafe behaviors, and the behavior code, which are abstracted from the database of risky behaviors. (Fig. 4).

3.4.1 Data storage processes

The HDFS architecture in the Apache Hadoop follows Master/Slave architecture. An HDFS cluster consists of a 'NameNode' and multiple 'DataNodes'. The 'NameNode' is a server node that manages file system spaces and controls clients' access to files. The 'NameNode' is a repository of all of the HDFS metadata and includes no user data. 'DataNodes' store information in files in your local file system. Metadata information includes information features (the image file name, the date, the city, the location of the violation, the amount of the violation fine, etc.) and the file directory obtained from the database of risky behaviors. Without including 'NameNodes', the file system is not able to function properly. The HDFS stores each file as sequential blocks. A file is divided into one or more blocks and these blocks are stored in a set of 'DataNodes'. All blocks except the last block have the same size. Block sizes are 64 or 128 MB, being 64 MB by default. The related data obtained from the analysis of the violation images are stored in 64-megabyte data blocks. The 'DataNode' is in charge of storing and retrieving the blocks summoned by a client or a 'NameNode' and sending a report, along with a list of blocks that are being stored to the 'NameNode'.

Clients include uploading and retrieving ports. When a client, after collecting the images of unsafe behaviors, sends the structured and unstructured data, along with their thorough semantic information, to be written on the HDFS, the data is initially stored in a temporary local file via the client application until its size reaches 64 MB, which is the default block size. Then, the client sends a request to the 'NameNode' and receives a 'DataNode' name. When the file system starts, the 'DataNode' generates a list of all the HDFS files and sends this list to

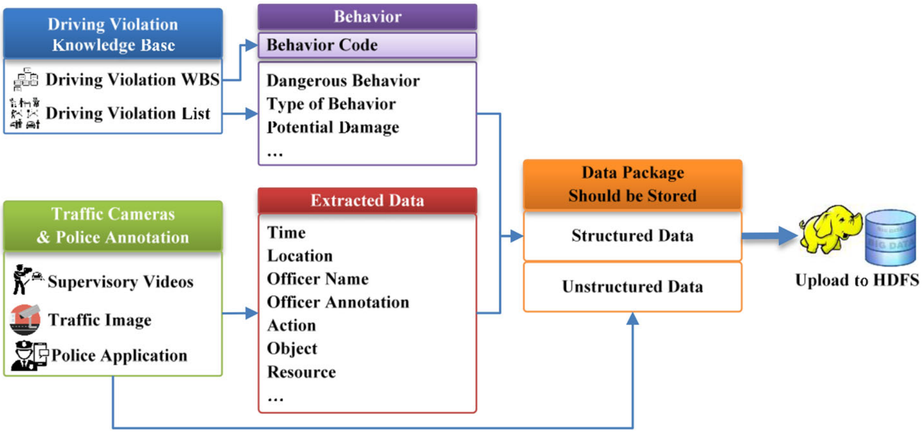


Fig. 4 The data sources and categorized information ready to be stored on the HDFS

the 'NameNode'. In the present study, the user utilized queries to extract useful information from a large amount of the data collected. In the first step, the query metadata in the 'NameNode' of the port were retrieved; then, 'NameNode' read the data achieved from the analysis of the images in the data blocks of the 'DataNode' and maintained the mapping information between the 'DataNodes' and data blocks. The data were duplicated in the cluster environment through the 'NameNode'. By default, each block of the data was copied three times (Fig. 5).

4 Implementation

According to the initial designers and legislators, all the puzzle pieces of any enacted rules must be enforced in the society in order for the rules to function effectively; in other words, if due to a variety of reasons or inadequate supervision, some pieces of this puzzle are implemented and some pieces are not enforced, the desired result is not achieved; therefore, this would create a loophole enabling drivers to escape rules. On the other hand, most drivers are afraid of being penalized or avoid even the slightest violations in the presence of police officers. So, if more precise monitoring is automatically done, it can act as a deterrent to drivers' violations, prompt changes in drivers' behavior, and prevent drivers from displaying

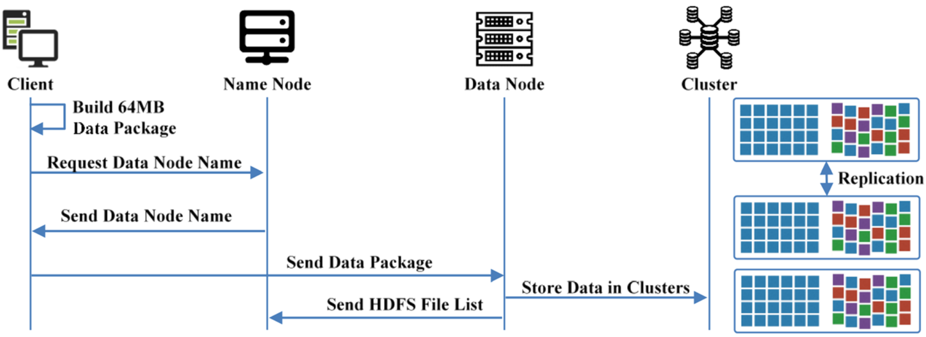


Fig. 5 Storage on the HDFS

unusual behaviors in the community. Therefore, if the traffic violator is identified immediately after committing a violation, and is punished corresponding to that violation, and treated in a decisive and patient manner, violations definitely reduce in the community. This section describes the sequential process of TVD system for better understanding of the reader. Then an example is described step by step.

4.1 The sequential process of TVD system

As seen in Fig. 6, the procedures for implementing TVD system for unsafe behavior detection are presented in a sequential diagram.

In this diagram, various phases of the procedure (objects) are represented as boxes at the top of the vertical cut-off lines. At the top of the vertical lines, there are certain items, including, user, monitoring, knowledge base, analysis, design, and implementation. The message sent from one phase to another is shown by a directional line between the lifelines of the two phases. Each message signifies a request to perform an action or a statement to announce an event. This message can include a set of parameters and control information. In Fig. 6, the videos of the surveillance cameras and the images (subtitled by the traffic officers) are sent by the user (the police), and then, the data is sent to the monitor phase. The collected data are stored in the knowledge base for further analysis. Moreover, the WBS created by traffic experts is loaded on the knowledge base. In this way, a list of unsafe driving behaviors along

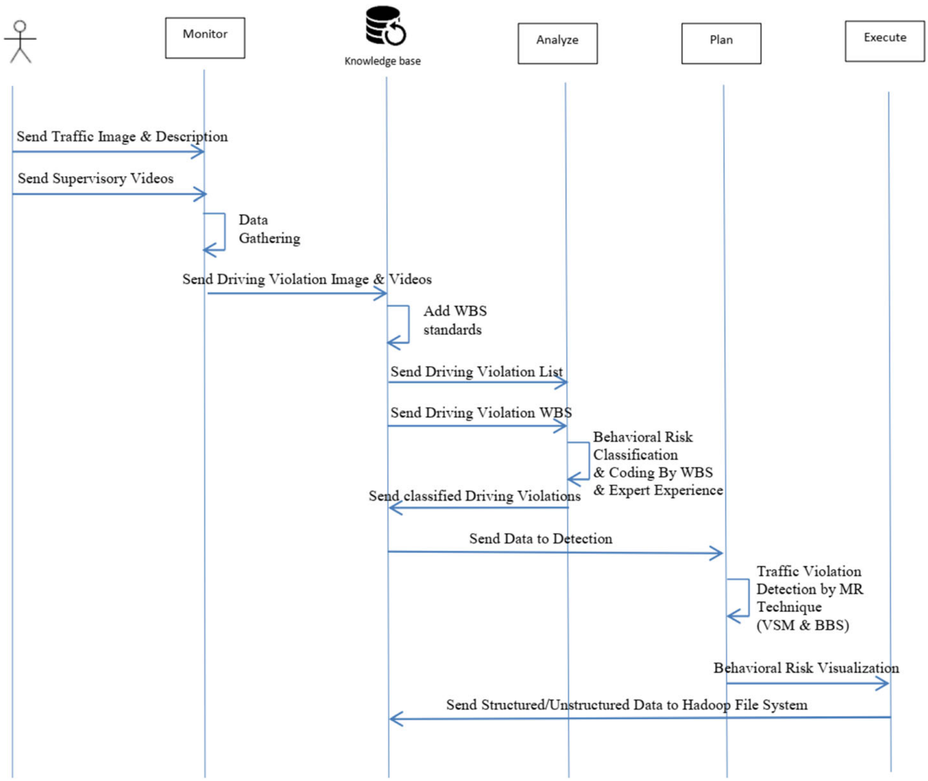


Fig. 6 The sequential diagram of TVD system

with the WBS and driving standards is sent to the analysis phase. In the analysis phase, each unsafe behavior is classified and coded (by experts), and the categorized unsafe behaviors are sent to the knowledge base. Then, the data in the knowledge base along with the WBS obtained from the previous phase are sent to the plan phase for discovering unsafe behaviors and validating them. In the plan phase, the drivers' unsafe behaviors are identified using the Map/Reduce technique; in other words, during the implementation phase, behavioral risk is visualized and the data are sent to the HDFS to be stored.

4.2 Pseudo-code of TVD system

Evaluating and analyzing the similarities of the police descriptions to all the unsafe behaviors is done by the MapReduce function in the Similarity-Description and Behaviors algorithm.

As the Similarity-Description and Behaviors algorithm indicate, in order to detect the similarity, TF-IDF, as the similarity criterion [64], is used,as shown in Eq. (1). In Eq. (1), $d_1$and d_2 are the vectors for police descriptions and for each of the predefined driving behaviors, respectively, and θ is the degree of their similarities. In line 2, the algorithm begins with two input parameters, i.e., all the predefined driving behaviors and the preprocessed police descriptions. In the Map function, in lines 3 to 8, the frequency vector of the police statements and the predefined behavioral data are extracted and in lines 9 to 16, their similarity is calculated using the cosine criterion in Eq. 1. In line 17, the similarity between the police descriptions and each of the predefined unsafe driving behaviors is determined and shown as output. The Reduce function, in the first line, receives all the calculated values along with the degree of their similarities and in lines 3 to 7, calculates the largest value, and ultimately in line

Algorithm1. Similarity-Description and Behaviors

```
1: Class Mapper
2:     Method Map (doc Behaviours, doc Police-Desc)
3:             for all terms t1 ∈ array(line-of- Behaviors) do
4:                     If does not Exist in array(line-of- Behaviors)
5:                             Frequency[t1] ← 1
6:                     else
7:                             Frequency[t1]++
8:                     end if
9:             pow1 ←0
10:            pow2 ←0
11:            sum ← array(Police-Desc)* array(Frequency)
12:            for all terms t1 ∈ array(Police-Desc) do
13:                    pow1 ← pow1+ pow(t1,2)
14:            for all terms t2 ∈ array(Frequency) do
15:                    pow2 ← pow2+ pow(t2,2)
16:            distance ←sum/(sqrt(pow1)*sqrt(pow2))
17:            Emit(doc line-of- Behaviors, Count distance)

1: Class Reducer
2:     Method Reduce (doc line-of- Behaviors, Count [distance1,distance2,…])
3:             Max← distance1
4:             for all Count distance ∈ Count[distance2,distance3,…] do
5:                     Max← Find-Max[Max,distance]
6:                     Max-S← Line-Max[distance]
7:             end for
8:             Emit(doc Max-S, Count Max)
```

Algorithm2. Mechanized diagnosis of driver behavior

```
Class Mapper
    method Map (Initial Images)
        x←pic.height()
        y←pic.width()
        Creat-Empty-Image(x,y)
        GrayScale (Initial Images)
        Lane←Lane-Detection(pic)
        sign←Traffic-Sign-Detection(pic)
        for all Element e ∈ arysign do
            If (sign == arysign) then
                Text1← name-of-TrafficSign
            end if
        end for
        Light ←Traffic-Light-Detection(pic)
        If (Background- of-light == Red) then
            Text2← "RED"
        else If (Background- of-light == Green) then
            Text2← "GREEN"
        else If (Background- of-light == Yellow) then
            Text2← "YELLOW"
        end If
        Line ←Pedestrain-Lines(Lane-pic)
        Grid ←Grid-Lines(Lane-pic)
        If (Text2 == "RED") OR (Text2=="YELLOW") then
            Vehicle1←Vehicle-Detection(Line-pic)
        end If
        if (vehicle1 is not empty ) AND (Text2=="RED") then
            Text3← "crossing prohibited places or red lights"
        end If
        x1←Lane.height()
        y1←Lane.width()
        Creat-Empty-Image(x1,y1)
        Lane1← ((x1)/4,y1)
        Vehicle2←Vehicle-Detection(Lane1)
        If ((Text1 == "No Stop")OR(Text1=="Carrying by crane "))AND(vehicle2 is not empty) then
            Text4← "stopping in prohibited places"
        end If
        Lane1← ((x1)/2,y1)
        Vehicle2←Vehicle-Detection(Lane1)
        If (Text1 == "No entry ") AND (vehicle2 is not empty) then
            Text5← "No entry "
        end If
        If (Text1 == "Stop absolutely prohibited") AND (vehicle2 is not empty) then
            Text6← "Stop absolutely prohibited"
        end If
        Emit(Final Picture, Text1, Text2, Text3, Text4, Text5, Text6)

Class Reducer
    Method Reduce (Final Picture, Text1, Text2, Text3, Text4, Text5, Text6)
        If (Text3 is not empty) then
            Emit(Text2+Text3, Final Picture)
        end If
        If (Text4 is not empty) then
            Emit(Text1+Text4, Final Picture)
        end If
        If (Text5 is not empty) then
            Emit(Text1+Text5, Final Picture)
        end If
        If (Text6 is not empty) then
            Emit(Text1+Text6, Final Picture)
        end If
```

8, provides the output of the Reduce function as the type of the driver's behavior and violation. In this way, human error, such as visual errors and the exact diagnosis of behavior types, is realized.

$$\cos(\theta) = \frac{d_1.d_2}{|d_1||d_2|} \tag{1}$$

The Pseudo-code related to the mechanized diagnosis of the driver behavior using videos is explained in the Mechanized diagnosis of driver behavior algorithm.

As seen in the mechanized diagnosis of driver behavior algorithm, in line 2, the Map function receives the image or frame of the captured video, and the length and width of the image are extracted in lines 3 and 4, respectively. In line 5, using the length and width of the initial image, an empty frame of the same size is generated. In line 6, the preprocessing operations, such as image graying, are applied to the image and the created gray image is placed in the empty frame. Then, in line 7 of the Map function, the street area is extracted (i.e., lane detection) using the method employed in papers [42, 63]. In lines 8 to12, to determine the type of traffic signs, the traffic sign detection function is called using the method utilized in papers [7, 17, 31, 33]. In lines 14 to 21, the traffic lights and their color are identified using the functions introduced in papers [19, 53] and the color is stored in a variable. Afterwards, in line 22, crosswalks are detected in the street area extracted in line 7 using the method proposed in paper [47]. In line 24, if the traffic light is red or yellow, the vehicle detection function in the area of crosswalks is called using the method introduced in paper [31]. Then, by using several conditions written in the next lines of the Mechanized diagnosis of driver behavior algorithm, it is possible to identify the violation type for each behavior. For example, in line 27, if the vehicle detection function within the specified range identifies a vehicle and the color of the traffic light is red, a violation, labeled as "red-light crossing", is stored in a variable. This case is an example of a "red-light crossing" violation. In lines 30 to 34 of this algorithm, the vehicle detection function in the street area is called. In lines 35 and 36, if the type of the extracted traffic signs is "no parking" or anything related to parking, a violation labeled "parking" is stored in a variable. In the remaining lines of the Map function (lines 37 to 44), given the aforementioned conditions, two other violations, namely, "no entry" and "no stopping" are also identified. It's worth noting that just a few prevalent violations are mentioned. Finally, all the functions extracted from the called functions are used as the input of the Reduce function.

The obtained results from each function in the Map function, including the last extracted image, the type of traffic signs, the color of the traffic light, and the violation type (red-light crossing, no parking, no entry, no stopping, etc.), are used as the input (line 2) of the Reduce function. Then, via several sequential conditions in lines 3 to 13, the outputs of different functions are compared with each other, and ultimately the violation type and the driver behavior are determined for the generated frame. For instance, in line 3, if a violation associated with "traffic lights" is detected in the frame under analysis, the output of the Reduce function is the final extracted image along with the color of the traffic light and the type of the committed violation. In line 6, if a violation related to "parking" is done in the frame under analysis, the output of the Reduce function is the final extracted image along with the type of the traffic signs and the violation type, i.e., "parking in the area of traffic signs." In line 9, if a violation pertaining to "no entry" is committed in the frame, the output of the Reduce function is the final extracted image along with the type of the traffic signs and the violation type, i.e., "no entry in the area of traffic signs". Similarly, in line 12, the violation of "no stopping" along with the image in which this violation is committed is shown as the output.

4.3 An example step by step

According to the Head of the Traffic Police, a majority of the violations committed last year were related to parking on pedestrian crossings, and heavy vehicles' entering and crossing urban roads. Therefore, in this study, the unsafe behaviors associated with 'no parking' and 'no entry' were investigated.

The procedure for entering the police officer's descriptions of the driver's unsafe traffic behaviors and the image taken at the location, used in the current study, is illustrated in Fig. 7. The description of the image is as follows: " **The car is parked near a stop sign in the street** ", and the influential factors include parking, the car, the stop sign, the street. As you can see, if the police description is not complete, all six BBS factors cannot be extracted, and as in Fig. 7 only four factors can be extracted from the six factors related to the police description.

The vector pertaining to the descriptions of traffic officers is calculated via the term-frequency method. The vectors of traffic violations are associated with the unsafe behavior of stopping in prohibited areas or creating a blockage, which includes 16 unsafe behaviors each of which is comprised of six influential factors, as seen in Table 1. It should be noted that there is zero in the vector for the police descriptions instead of non-listed factors (insufficient descriptions of traffic police).

Then, based on the cosine measure, the vector related to the police descriptions is compared to each of these 16 behaviors, and their values are calculated, as Table 5 indicates. The value of 0.99 is reported as the highest value. Therefore, the exact type of violation based on the police descriptions is related to behavior No. 14 namely, "**Stopping in 'no parking' areas (parking prohibited)**".

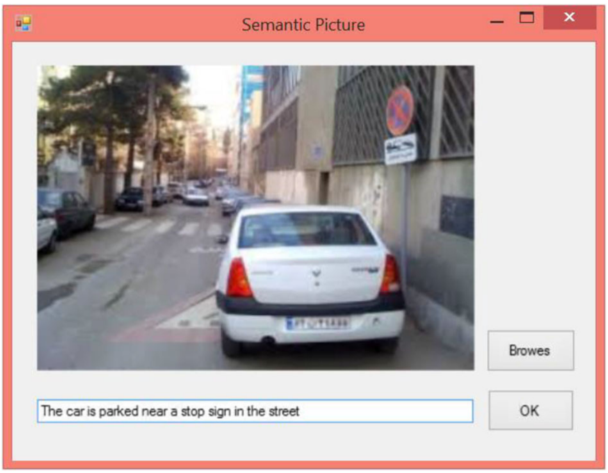


Fig. 7 The software interface for the police officer's descriptions

The images accompanied by the information on the exact type of violations are stored on the Hadoop distributed file system (HDFS). Therefore, the information can be stored and retrieved at a high speed. Figure 8 displays a diagram of the interface of the automatic video surveillance system related to the HDFS. In Fig. 8 (a), the interface is designed to apply restrictions in terms of the dates, traffic violation types, etc. and to analyze them via surveillance cameras or organized software. Figure 8 (b) indicates the results of imposing restrictions, including stopping in prohibited areas or crossing areas at certain times when crossing is prohibited, along with certain details, such as the code, the date, the image, the type of the analysis of the police descriptions, the ramifications of the violation, and the exact name of the violation. Therefore, this interface makes it possible to display analyzed images and videos.

5 Experimental result

In this section, the efficiency of the proposed approach is examined and the results are presented. In this experimental study, the proposed approach is implemented via Java language along with OpenCV library. The tested computer system is implemented in the VMware environment using two clusters. The first cluster has only one node and the second cluster contains a master node and seven slave nodes. The Linux Ubuntu operating system is installed and run on all the nodes. Software and hardware specifications of the proposed system are shown in Table 6. To evaluate the performance of TVD, two datasets in Table 7 are employed.

To consider various criteria in analyzing the results of TVD, four scenarios are used as follows: performance and efficiency, scalability, accuracy of error detection, and communication overhead. In what follows, these scenarios are discussed in detail.

5.1 Performance scenario

To compare the performance of TVD, two computational experiments were carried out. In these experiments, the sequential program with the Hadoop cluster in stand-alone mode was investigated and the performance of the TVD, sequential program, and traditional system were compared. These comparisons are explained separately in the following sections.

Table 5 The Sample output of Calculated Weights and Cosine Similarity values

	Term1	Term2	Term3	Term4	Term5	Term6	....	Term N
Unsafe Behavior1	0	2	1	0	0	1	....	1
Unsafe Behavior2	0	0	0	0	0	0	....	0
Unsafe Behavior3	0	1	0	0	2	2	....	2
Unsafe Behavior4	1	1	1	1	2	1	....	0
Unsafe Behavior5	0	1	1	0	0	0	....	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Unsafe Behavior16	0	0	0	0	3		....	0

	Unsafe Behavior1	Unsafe Behavior2	Unsafe Behavior3	Unsafe Behavior4	. . . . . .	Unsafe Behavior14	Unsafe Behavior15	Unsafe Behavior16
Value	0.62	0.24	0.74	0.74		0.99	0.44	0.44

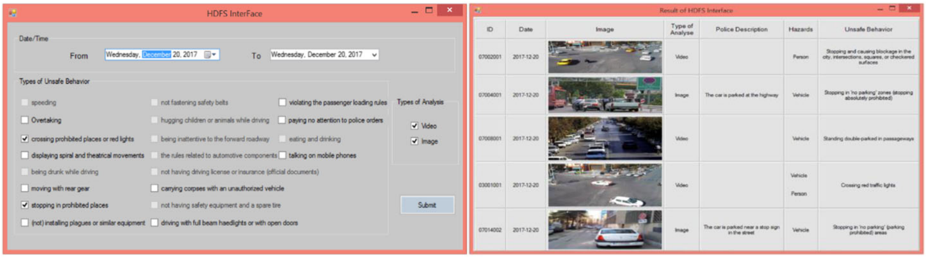


Fig. 8 A diagram of the interface of TVD system related to the HDFS

5.1.1 Scenario: The comparison between the performance of the sequential program and Hadoop in stand-alone mode

To compare the performance of TVD in the sequential program without the Map/Reduce technique and in Hadoop in stand-alone mode, the processing time (CPU time) of a number of police descriptions and the images of the surveillance cameras is examined in the range of 10,000 to 100,000 and then 1000 to 10,000. The processing time is calculated from the time the police descriptions and the images are received to the time the output of TVD is obtained. The results are shown in Figs. 9 and Fig. 10.

In Fig. 9, the blue line and the red line, as the processing time of the sequential program and that of the Hadoop cluster in stand-alone mode, respectively, are used to analyze a different number of unstructured data (police descriptions). Clearly, Hadoop in stand-alone mode requires more processing time due to high overhead. In addition, the difference in their processing time increases as the number of textual data (police descriptions) boosts. For example, the processing time of 10,000 textual data using Hadoop in stand-alone mode is 2400 units while the processing time of 100,000 data increases to 23,000 units. Therefore, the performance of the sequential program is much better than Hadoop in stand-alone mode since in Hadoop in stand-alone mode, there is overhead when the data are loaded into the Hadoop cluster.

In Fig. 10, the blue line and the red line, as the processing time of the sequential program and that of Hadoop in stand-alone mode, are employed to analyze a different number of unstructured data (the surveillance camera images). Like the analysis of the processing time for the unstructured data (textual data), the processing times increases as the number of multimedia data (the images) augments. For example, in the processing of 100 images via Hadoop in stand-alone mode, the processing time is 3300 units while in the processing of 10,000 images, it increases to 33,000 units.

Table 6 Software and hardware specifications

SW/HW	Version
Operating System	Linux Ubuntu 14.04 LTS
Hadoop distributed file system	Hadoop 2.7.1
Java	JDK 1.8.0_152
OpenCV	OpenCV_3.2.0
Virtual tool	VMware Workstation 12
CPU	Intel (R) Core(TM) i7-4710HQ CPU 2.50GHz
Memory	16 GB

Table 7 Dataset descriptions

No.	Dataset	#Records	#Dimensions	#Clusters
I	Unstructured Dataset (Police Description)	1,000,000	6	–
II	Multimedia Dataset (Images captured by the traffic control center)	100,000	–	–
III	Predefined behavioral data of traffic	171	6	20

Therefore, regarding the textual data, the efficiency of the driver behavior detection system using Hadoop in stand-alone mode reduces due to excessive overhead.

5.1.2 Scenario: The comparison of the performance of TVD and traditional system

To detect the reason for the superiority of TVD over the traditional system and to compare them, the best approach is to compare their performance against each other, and in this way, the effectiveness of this method in identifying drivers' violations (unsafe behaviors) becomes more evident. It is worth noting that the traditional system of detecting unsafe behaviors refers to the presence of an operator in traffic control centers or traffic managers (trained personnel) at the city level. This operator knows traffic laws and identifies drivers' unsafe behaviors based on those laws. To evaluate the performance of the proposed approach with seven slave nodes, the processing time of a number of police descriptions and surveillance cameras with 2000, 5000, and 10,000 data are examined via a sequential program without Map /Reduce and a traditional system (Job). The results related to their performance are shown in Fig. 11.

In Fig. 8, the red and blue bars represent the processing time of the cameras and police explanations, respectively. In this diagram, the processing time of three types of systems, which are called TS (traditional system), sequential program (SP), and MR (TVD), with seven

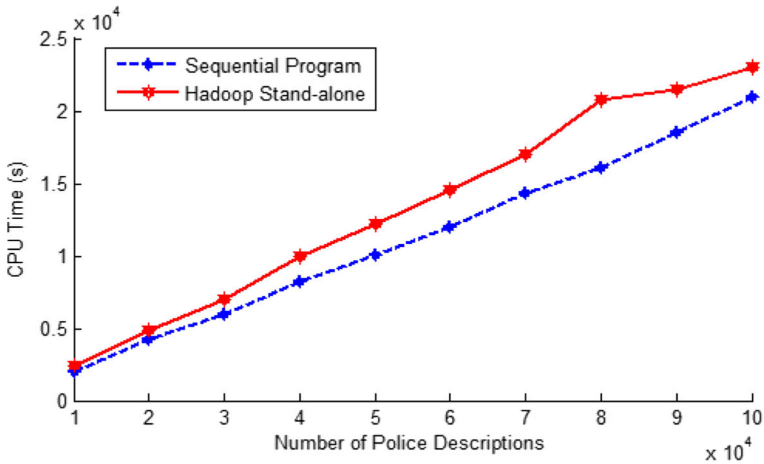


Fig. 9 The processing time of the police descriptions using TVD in the sequential program and Hadoop in Stand-alone mode

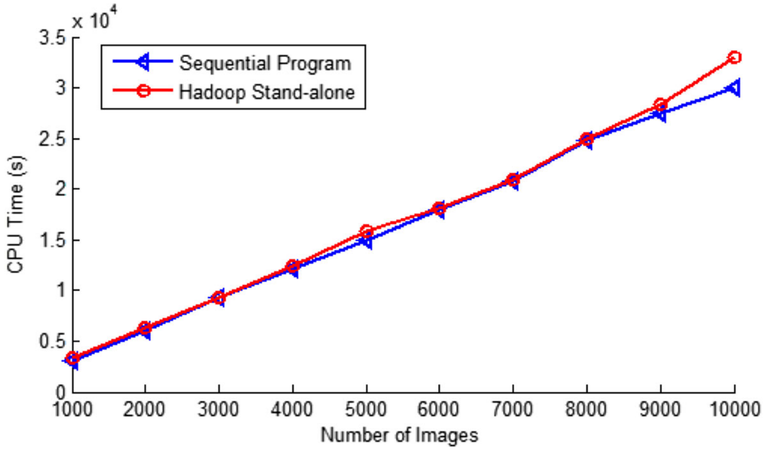


Fig. 10 The processing time using TVD for the surveillance camera images in the sequential program and Hadoop in stand-alone mode

slave nodes is compared. It is clear that the Map/Reduce program (Hadoop cluster) consumes less time in processing multiple data due to the presence of seven nodes. In addition, the difference in their processing time boosts with increasing the number of images and extracting unsafe behaviors. For example, when the program with seven slave nodes in the Hadoop cluster processes 5000 images and police descriptions, its processing time is 5740 and 544, respectively, while this processing time for the same data in the sequential program is equal to 15,000 and 1000, respectively, and that in the traditional system is 24,000 and 1500, respectively. Therefore, the performance of the program in the Hadoop cluster with seven slave nodes is far better than that in the traditional system and sequential program.

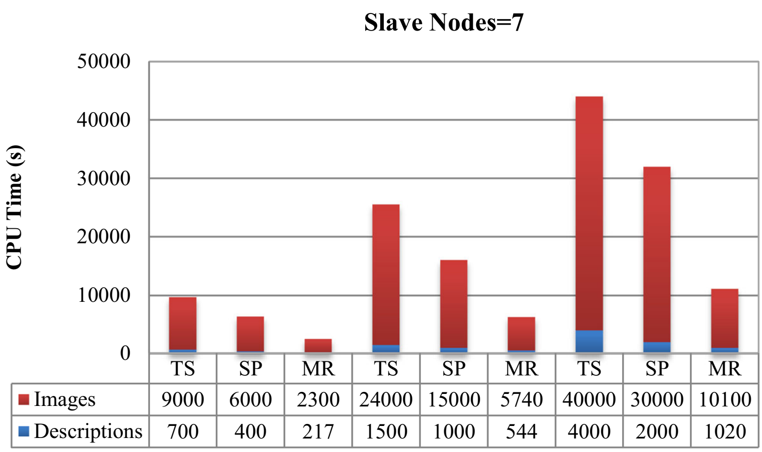


Fig. 11 Evaluating the performance of TVD, sequential program, and traditional system

5.2 Scalability scenario

In order to estimate the scalability of TVD, three computational tests have been performed, which are separately explained in detail in the following sections.

5.2.1 Scenario: Scalability regarding the number of slave nodes

In addition, in order to improve the performance of TVD, the number of nodes in each Hadoop cluster is increased; then, the processing time of 100,000 unstructured /textual data and that of 1000 multimedia/image data are compared with a cluster containing 1 to 4 slave nodes shown in Fig. 12. As clear, the processing time reduces as the slave nodes (data nodes) increase. Interestingly, the processing time required for detecting the drivers' behavior using the unstructured/ textual data via a Hadoop cluster with 1 slave node is 23,000 s, while running this operation on a cluster with 4 slave nodes takes about 12,500 s. Also, regarding the unstructured data, the processing times in the Hadoop cluster with 1 slave node and in the Hadoop cluster with 4 slave nodes are 33,000 units and 21,500 units, respectively, indicating a considerable reduction from the former to the latter. Therefore, both textual/image data are distributed all over the created data nodes, reducing the processing time. In other words, Fig. 12 shows that the efficiency of TVD increases linearly by increasing the number of slaves in the Hadoop cluster.

As a result, with an increase in the number of slave nodes, the processing time of the police descriptions and that of the images of the control centers reduces by over 45.7% and 35%, respectively. The reason for the reduction in processing time is that the Hadoop breaks the files into large blocks and distributes them across the nodes of one cluster. For data processing, the Map/Reduce section sends code packets to the nodes to make them process the data in parallel. This approach exploits data locality, meaning that, nodes

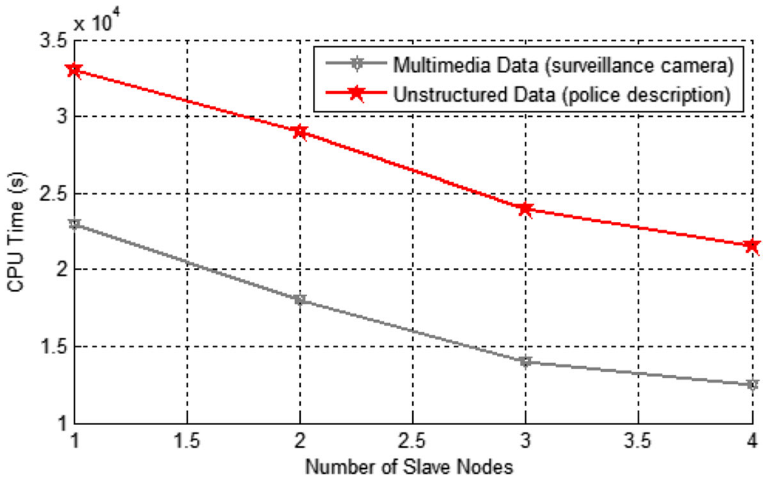


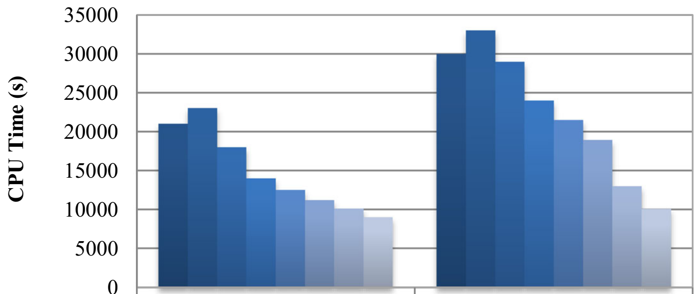
Fig. 12 The processing time using TVD for the unstructured/textual and multimedia/image data with regard to different slave nodes

work on the parts of the data they have access to. In this way, the data are processed faster and more efficiently.

5.2.2 Scenario: The comparison of the efficiency and scalability of the Hadoop cluster with one master node and several slave nodes

The purpose of the first evaluation is to compare the scalability of implementing the TVD algorithms by increasing the number of nodes. During these experiments, two datasets shown in Table 7 are processed using a number of nodes.

Based on the results presented in Fig. 13, apart from the single node cluster (Hadoop in stand-alone mode), the system has near-linear scalability. However, in the analysis of the police descriptions, compared to the Hadoop cluster with the seven slave nodes, the Hadoop cluster with one slave node leads to an almost 60.87% reduction in the processing time. The reason for this improvement is that in the Map/Reduce program, the system automatically performs calculations among large clusters of machines in parallel, manages system failures, and schedules within-system communication, and in this way, it makes the use of network and disks easier and more effective. Thus, through the balanced distribution of computing load across nodes, the parallel execution is performed without interference and the processing time decreases. Besides, in the analysis of the surveillance camera images, the processing time reduces by 70%. As stated in Section 5.1, in comparison with the sequential implementation, a cluster with one node reveals poorer performance, which is due to Hadoop overhead.



	Police Description	Surveillance Camera
Sequential program	21000	30000
1 Node	23000	33000
2 Nodes	18000	29000
3 Nodes	14000	24000
4 Nodes	12500	21500
5 Nodes	11200	18913
6 Nodes	10128	12975
7 Nodes	9001	10100

Fig. 13 The processing time using TVD with regard to the number of nodes

5.2.3 Scenario: The comparison of the scalability of TVD and traditional system

The purpose of this scenario is to examine the scalability of TVD, the sequential program, and the traditional system for unsafe behavior detection considering the size of the data and the number of different slave nodes. To this end, the number of slave nodes varies between one to seven, and the processing time is computed for 2000, 4000, 6000, 8000 and 10,000 data derived from the police descriptions and the images of the cameras (Job). The results are shown in Fig. 14.

As seen in Fig. 14, the blue bars represent the processing time of the police descriptions and the red bars represent the processing time of the images obtained from the surveillance cameras. In this diagram, the processing time of three types of systems, called TS (traditional system), SP (sequential program) and MR (TVD), with different numbers of slave nodes are compared. As observed in this figure, for a given amount of data, the processing time in the traditional system and the sequential program with different slave nodes is the same. For example, for 4000 textual and visual data with different numbers of slave nodes (varying from one to seven), the processing time of the traditional system is similar to that of the sequential program. The reason is that there is just one node in the sequential and traditional systems. However, the processing time

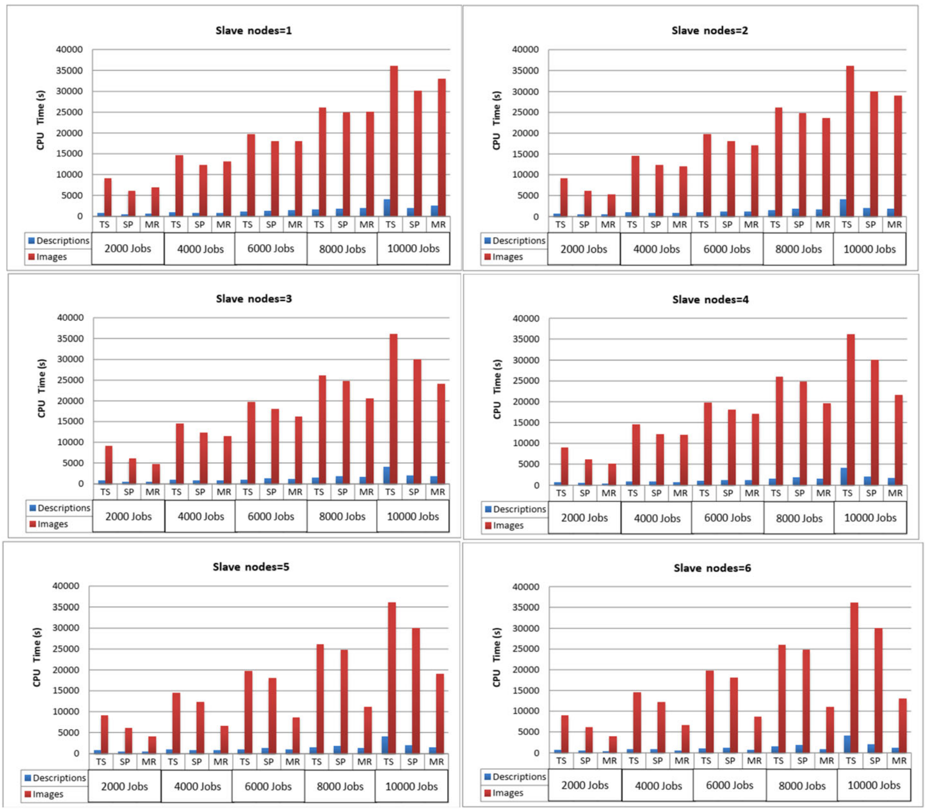


Fig. 14 The scalability of TVD, sequential program and traditional system for unsafe behavior detection in terms of the volume of data and different numbers of slave nodes

differs for various jobs when different numbers of slave nodes are used; in other words, in each graph, as the number of slave nodes increases with regard to the volume of data, the processing time decreases.

5.3 Scenario: The comparison of the accuracy, precision, and recall of TVD and traditional system in terms of data volume

Assuming that our goal is to predict the "stopping absolutely prohibited" action. If our prediction is positive (Yes), it means that the drivers' behavior is a "stopping absolutely prohibited" action, and if our prediction is negative (No), the driver's behavior is not a "stopping absolutely prohibited" action. Regarding the analysis of the cells of the confusion matrix for this case study, the following points must be considered:

- True Positive (TP): If the driver's action is a "stopping absolutely prohibited" action in reality, and the predicted value indicates a "stopping absolutely prohibited" action.
- False Positive (FP): If the driver's action is not a "stopping absolutely prohibited" action in reality, and the predicted value shows a "stopping absolutely prohibited" action.
- False Negative (FN): If the driver's action is a "stopping absolutely prohibited" action in reality, and the prediction does not indicate a "stopping absolutely prohibited" action.
- True Negative (TN): If the driver's action is not a "stopping absolutely prohibited" action in reality, and the prediction does not indicate a "stopping absolutely prohibited" action.

We review the precision scenario for 200 records (the textual data obtained from the police descriptions and the visual data from surveillance cameras). As clear, in the ideal situation, the negative case (FN & TN) are zero; however, in reality, it is far-fetched. Thus, certain mechanisms and criteria are required to check the accuracy, precision, and performance of the model generated from the data. Our expectation is that true cases (T) are much higher than negative cases (N). Therefore, certain criteria are needed to be able to measure the performance and accuracy of these models. In what follows, these criteria or measures are explicated in detail.

N = 200		**Predicted values**		
		Positive	Negative	
Actual values	Positive	TP = 105	FN = 16	121
	Negative	FP = 21	TN = 58	79
		126	74	

The first measure or criterion is accuracy or the model's capability to diagnose actions correctly; in other words, accuracy is the ratio of the correct diagnoses (TP + TN) to the total data.

$$\text{Accuracy} = \frac{TP + TN}{N} = \frac{105 + 58}{200} = 0.81$$

Precision means the ratio of the number of true positive diagnostic samples to the total number of positive diagnostic samples (TP + FP):

$$Precision = \frac{TP}{TP + FP} = \frac{105}{105 + 21} = 0.83$$

When the obtained accuracy value is close to one, another criterion is required. This criterion is called Recall. It should be noted that the recall formula of all the true positive samples includes the samples which are correctly detected as positive samples (TP) and the samples which are positive but are mistakenly identified as negative samples (FN).

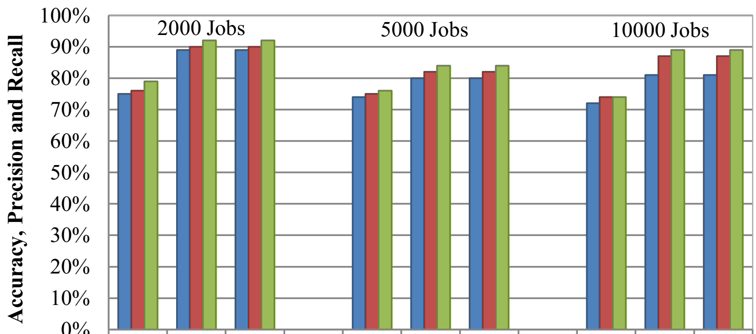
$$\text{Recall} = \frac{TP}{TP + FN} = \frac{105}{105 + 16} = 0.87$$

The purpose of this scenario is to explore the criteria for the detection of unsafe behaviors in different volumes of data in TVD, sequential program (SP), and traditional system (TS). To explore the accuracy, precision, and recall of the proposed approach and compare them with those of the traditional system, different numbers of police descriptions and images of 2000, 5000, and 10,000 data (Job) are shown in Fig. 15.

In Fig. 15, the blue, red, and green bars represent the criteria of accuracy, precision, and recall, respectively. In this diagram, the traditional system is compared to the sequential program and MR program (TVD). It is quite obvious that the values of these criteria in the sequential program and MR are the same due to the mechanization of the system. Moreover, the difference between the sequential and MR programs is only related to processing time and the reduction of processing time. As seen in this figure, the values of all three criteria, namely, Accuracy, Precision, and Recall, in the MR system are far higher than those of the traditional ones.

5.4 Scenario: Communication overhead with regard to an increase in the number of nodes in Hadoop

When the Hadoop is installed, SSH activation makes it possible for the Hadoop to link different nodes using RCP without any passwords. This abstraction in the TCP protocol is officially called the client protocol and the data node protocol. The data nodes (slave) send



	TS	SP	MR		TS	SP	MR		TS	SP	MR
Accuracy	0.75	0.89	0.89		0.74	0.8	0.8		0.72	0.81	0.81
Precision	0.76	0.9	0.9		0.75	0.82	0.82		0.74	0.87	0.87
Recall	0.79	0.92	0.92		0.76	0.84	0.84		0.74	0.89	0.89

Fig. 15 Accuracy, Precision, and Recall of TVD, sequential program, and traditional system for unsafe behavior detection in terms of data volume

heartbeats (every three seconds) to the name of the node to make it realize that it is still active. As the number of nodes is large, approximately more slave and master nodes communicate with each other.

Increasing the number of nodes in a large volume of data can improve the performance of the system, which is deemed as its main advantage. However, the large number of nodes can increase communication costs (communication overhead time) and there is the probability of the failure of a number of slave nodes, which has a profound effect on system performance.

Figure 16 depicts the communication time (communication overhead-delay) and computational time (execution) for different slave nodes. The selected dataset is 100,000 unstructured textual data and 1000 multimedia data with a cluster of 1 to 7 slave nodes. Figure 16 shows that the execution time decreases with increasing the number of slave nodes while the communication time is low and almost constant. As observed in this figure, communication delay and information overhead are somehow noticeable for a greater number of slave nodes. Therefore, based on Eq. (2), it can be stated that with increasing the number of nodes, there is a significant reduction in execution time and an increase in communication overhead, and vice versa.

$$\text{Number of nodes}\uparrow \Rightarrow (\text{Communication overhead})\uparrow \wedge (\text{Execution Time})\downarrow \tag{2}$$

5.5 Scenario: The comparison of three big data platforms

The review of the articles provided in Section 2 and [23, 37] undertaken on unsafe behavior detection in the field of construction using a large volume of data and in comparison with the detection of drivers' unsafe behaviors shows that there are a few quantitative criteria in these articles; therefore, it was decided to compare these two studies with the present study in terms

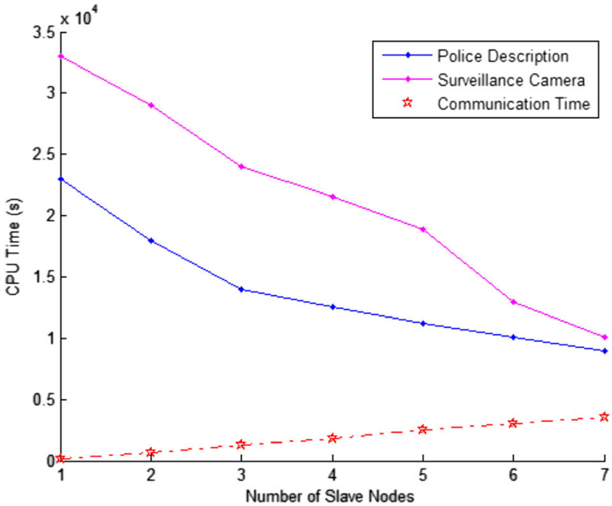


Fig. 16 Communication time (communication delay) and computation time obtained from TVD for unstructured textual and multimedia data with respect to different numbers of slave nodes

of qualitative criteria. Table 8 compares [23, 37] in terms of certain criteria, such as fieldwork, behavioral identification, and so on, through the proposed TVD system. Table 8 suggests that not only is the useful information of the articles on the behavior detection of big data utilized but also it is attempted to introduce, discuss, and compare several quantitative criteria, including, Accuracy, Precision, Recall, in the form of different scenarios.

Therefore, this study aimed at eliminating human errors by using a big data-based design. In this design, the following processes are conducted:

A. Generating a database of driver behaviors is a new approach in the realm of traffic control. It can be used as an information source to detect drivers' behaviors and

Table 8 The Comparison of two big data platforms via TVD system

		System proposed in [23]	System proposed in [37]	TVD system
Qualitative criteria	Big data platform	✓	✓	✓
	Fieldwork	Metro Construction in China (WUHAN)	Illegal construction waste dumping (Hong Kong)	Iranian Drivers (Yazd)
	behavioral identification	Construction workers	Urban drivers	Urban drivers
	Implementation tools (Hadoop)	✓	✓	✓
	Identification of camera functions in the relevant field	✓	–	✓
	Methodology	–	–	MAPE
	Identification of incorrect behavioral activities (actions)	✓	✓	✓
	Action classification and coding	✓	✓	✓
	Identification of influential factors in each action	✓	✓	✓
	Use of WBS in action division	✓	–	✓
	Use of BBS in identifying individuals' behaviors	✓	–	✓
	Use of VSM (TF-IDF) in calculating similarities	✓	–	✓
	Evaluation of the MapReduce technique	–	–	✓
	Extraction of images using semantic information	✓	✓	✓
	Analysis of textual and visual data	✓	✓	✓
	Use of interface design for recording unsafe behaviors	✓	–	✓
	Use of interface design for displaying extracted information	✓	✓	✓
Quantitative criteria	Investigation of Performance	–	–	✓
	Investigation of Scalability	–	–	✓
	Investigation of error detection accuracy, precision, Recall	–	–	✓
	Investigation of Communication overhead	–	–	✓
	Comparison with other systems	–	–	✓

violations. Also, if combined with a work breakdown structure, it can be used to categorize driver behaviors in dynamic environments.

B. In order to analyze unstructured data (such as textual data) and multimedia data (including video files and images), data collection is organized based on two applications. The applications are designed to observe drivers' violations and analyze the textual data obtained from traffic officers' descriptions. The method of determining vectors for each law is predefined and the violations of drivers and the frequency-based cosine similarity measure are used for comparison purposes. The analysis of multimedia data is also based on various functions and the identification of the semantic structure of images. Therefore, focusing on the identification of the mechanism, this TVD system would not be under the influence of people's feelings.

C. In the data storage phase, the mechanism of data storage on the HDFS is utilized since it is error-resistant and prevents the possible risks of duplication on blocks. The emphasis of the HDFS is on reducing the time of data accessibility, which expedites data retrieval. Therefore, driver violations can be retrieved and checked at a desired amount of time.

Despite all the advantages pointed out, this study has several limitations, which may hopefully be eliminated in future studies. Due to the different skills of people, various descriptions for the same concepts may be provided by different traffic officers. It is required to consider various synonyms of different terms. The vehicle under study was the ordinary car and the tests were conducted on it. Moreover, all tests were conducted in the city and on the streets. So, such tests can also be undertaken on highways and roads.

6 Conclusion

Drivers' behaviors are one of the most influential and primary sources of extensive damage, casualties, and injuries to a large number of people. Therefore, employing the BBS is a very effective approach for reducing the incidence of traffic violations and the number of accidents and casualties in countries. In this paper, an approach to the detection of accident-prompting conditions is propounded, resulting in a significant decrease in traffic losses. The main novelty of this paper is the use of big data to analyze officers' descriptions of unsafe behaviors and the exploration of urban surveillance cameras to identify drivers' violations. Given the need for traffic officers or queries, using the organized software, the recovery system expounded in this article examines the entire database of traffic rules with regard to the extracted terms and presents the most similar one as the violation committed. On the other hand, in order to analyze the videos recorded by surveillance cameras, drivers' violations can be identified automatically using the structure described in this article, and then, saved and retrieved through the Hadoop distributed file system (HDFS). Thus, almost all the unsafe traffic behaviors, including structured and unstructured data, along with their semantic information can be saved in this TVD system in full details. Since a major bulk of the studies conducted on this topic solely devoted themselves to qualitative criteria, the present study considered both qualitative criteria such as fieldwork and behavioral identification and quantitative criteria such as accuracy, precision, and recall in its investigation of the issue.

Acknowledgements The authors would like to express their gratitude to the Traffic Control Center and the traffic officers for their contribution to realizing this study.

References

1. Ahmadi, S., Investigating the relationship between gender and driving disorder. Quarterly J Traffic Manag Stud, (2011): p. 13–24
2. Aliane N, Fernandez J, Mata M, Bemposta S (2014) A system for traffic violation detection. Sensors 14(11): 22113–22127
3. Bakhtiyari M, Delpisheh A, Riahi SM, Latifi A, Zayeri F, Salehi M, Soori H (2012) Epidemiology of occupational accidents among Iranian insured workers. Saf Sci 50(7):1480–1484
4. Banos O et al (2015) Traffic behavior recognition using the pachinko allocation model. Sensors 15(7): 16040–16059
5. Bener A, Lajunen T, Özkan T, Haigney D (2006) The effect of mobile phone use on driving style and driving skills. Int J Crashworthiness 11(5):459–465
6. Brémond F, Thonnat M, Zúniga M (2006) Video-understanding framework for automatic behavior recognition. Behav Res Methods 38(3):416–426
7. Bui-Minh, T, et al. Two algorithms for detection of mutually occluding traffic signs. In Control, Automation and Information Sciences (ICCAIS), 2012 International Conference on. (2012). IEEE.
8. Cárdenas-Benítez N, Aquino-Santos R, Magaña-Espinoza P, Aguilar-Velazco J, Edwards-Block A, Medina Cass A (2016) Traffic congestion detection system through connected vehicles and big data. Sensors 16(5): 599
9. Chauvin C, Lardjane S, Morel G, Clostermann JP, Langard B (2013) Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS. Accid Anal Prev 59:26–37
10. Chen T, Lu S (2016) Accurate and efficient traffic sign detection using discriminative adaboost and support vector regression. IEEE Trans Veh Technol 65(6):4006–4015
11. Chen D, Tian H (2012) Behavior based safety for accidents prevention and positive study in China construction project. Procedia Eng 43:528–534
12. Chen M, Mao S, Liu Y (2014) Big data: A survey. Mobile Net Appl 19(2):171–209
13. Chi S, Caldas CH (2011) Automated object identification using optical video cameras on construction sites. Comput-Aided Civil Infrastruct Eng 26(5):368–380
14. Computing, I.A., White paper: An architectural blueprint for autonomic computing. (2005).
15. Cox S, Jones B, Rycraft H (2004) Behavioural approaches to safety management within UK reactor plants. Saf Sci 42(9):825–839
16. Datta R et al (2008) Image retrieval: Ideas, influences, and trends of the new age. ACM Comput Surveys (Csur) 40(2):5
17. De La Escalera A et al (2004) Visual sign information extraction and identification by deformable models for intelligent vehicles. IEEE Trans Intell Transp Syst 5(2):57–68
18. De Mauro A et al (2018) Human resources for big data professions: A systematic classification of job roles and required skill sets. Inf Process Manag 54(5):807–817
19. Elotmani, S and M El Hitmy. A light traffic signs recognition system. In Multimedia Computing and Systems (ICMCS), 2014 International Conference on. (2014). IEEE.
20. Geller, E.S., Keys to behavior-based safety. (2001): Government Institutes.
21. Goh YM, Chua D (2009) Case-based reasoning for construction hazard identification: Case representation and retrieval. J Constr Eng Manag 135(11):1181–1189
22. Gowsikhaa D, Abirami S, Baskaran R (2014) Automated human behavior analysis from surveillance videos: A survey. Artif Intell Rev 42(4):747–765
23. Guo S et al (2016) A big-data-based platform of workers' behavior: Observations from the field. Accid Anal Prev 93:299–309
24. Hajeb, S., et al., Traffic violation detection system based on RFID. Int J Sci Eng Investigat, (2013). 1.
25. Hartford, T, About Behavior-Based Safety Management, in The Hartford Loss Control Department, Technical Information Paper Series. (2002).
26. Hu H-Y, Qu Z-W, Li Z-H (2014) Multi-level trajectory learning for traffic behavior detection and analysis. J Chin Inst Eng 37(8):995–1006
27. Iglesia D (2014) Mape-k formal templates for self-adaptive systems: Specifications and descriptions. Linnaeus University, Smaland

28. Ismail F, Hashim AE, Zuriea W, Ismail W, Kamarudin H, Baharom ZA (2012) Behaviour based approach for quality and safety environment improvement: Malaysian experience in the oil and gas industry. Procedia Soc Behav Sci 35:586–594
29. James, L., Road rage and aggressive driving: Steering clear of highway warfare. (2000): Prometheus Books.
30. Jin, Z., Environment modeling-based requirements engineering for software intensive systems. (2018): Morgan Kaufmann.
31. Kasaei, SHM and SMM Kasaei. Extraction and recognition of the vehicle license plate for passing under outside environment. In Intelligence and Security Informatics Conference (EISIC), 2011 European. (2011). IEEE.
32. Klubsuwan, K., W. Koodtalang, and S. Mungsing. Traffic violation detection using multiple trajectories evaluation of vehicles. In 2013 4th International Conference on Intelligent Systems, Modelling and Simulation. (2013). IEEE.
33. Krishnan, A, C Lewis, and D Day. Vision system for identifying road signs using triangulation and bundle adjustment. In Intelligent Transportation Systems, 2009. ITSC'09. 12th International IEEE Conference on. (2009). IEEE.
34. Kunfeng, W, et al. An automated vehicle counting system for traffic surveillance. in Vehicular Electronics and Safety ICVES, IEEE International Conference. (2007).
35. Li H, Lu M, Hsu SC, Gray M, Huang T (2015) Proactive behavior-based safety management for construction safety improvement. Saf Sci 75:107–117
36. Lingard H, Rowlinson S (1997) Behavior-based safety management in Hong Kong's construction industry. J Saf Res 28(4):243–256
37. Lu W (2019) Big data analytics to identify illegal construction waste dumping: A Hong Kong study. Resour Conserv Recycl 141:264–272
38. Manyika, J., et al., Big data: The next frontier for innovation, competition, and productivity. (2011).
39. Mehta P, Majumder P (2018) Effective aggregation of various summarization techniques. Inf Process Manag 54(2):145–158
40. Metari S, Prel F, Moszkowicz T, Laurendeau D, Teasdale N, Beauchemin S, Simoneau M (2013) A computer vision framework for the analysis and interpretation of the cephalo-ocular behavior of drivers. Mach Vis Appl 24(1):159–173
41. Moura R, Beer M, Patelli E, Lewis J, Knoll F (2017) Learning from accidents: Interactions between human factors, technology and organisations as a central element to validate risk studies. Saf Sci 99:196–214
42. Nguyen V, Kim H, Jun SC, Boo K (2018) A study on real-time detection method of lane and vehicle for lane change assistant system using vision system on highway. Eng Sci Technol, Int J 21(5):822–833
43. Nielsen KJ, Kines P, Pedersen LM, Andersen LP, Andersen DR (2015) A multi-case study of the implementation of an integrated approach to safety in small enterprises. Saf Sci 71:142–150
44. Organization, W.H, Global status report on road safety (2015). 2015: World Health Organization.
45. Ou, G., Y. Gao, and Y. Liu. Real-time vehicular traffic violation detection in traffic monitoring stream. In Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03. (2012). IEEE Computer Society
46. Pakgohar A, Tabrizi RS, Khalili M, Esmaeili A (2011) The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach. Procedia Comput Sci 3:764–769
47. Phung SL, Le MC, Bouzerdoum A (2016) Pedestrian lane detection in unstructured scenes for assistive navigation. Comput Vis Image Underst 149:186–196
48. Pourmahalam N, Ghorbani M, Zakeri H (2010) Assessment of management and safety engineering in driving safety improvement in Iran. J Traffic Manag Stud 19:13–26
49. Qian, H, X Wu, and Y Xu, Intelligent surveillance systems. Vol. 51. (2011): Springer Science & Business Media.
50. Rakotonirainy A, Schroeter R, Soro A (2014) Three social car visions to improve driver behaviour. Pervasive Mobile Comput 14:147–160
51. Reason J et al (1990) Errors and violations on the roads: A real distinction? Ergonomics 33(10–11):1315–1332
52. Reber RA, Wallin JA, Chhokar JS (1984) Reducing industrial accidents: A behavioral experiment. Industrial Relations: J Econ Society 23(1):119–125
53. Sallah, M, et al., Road sign detection and recognition system for real-time embedded applications. (2011).
54. Salton G, Wong A, Yang C-S (1975) A vector space model for automatic indexing. Commun ACM 18(11): 613–620
55. Secundo G, del Vecchio P, Dumay J, Passiante G (2017) Intellectual capital in the age of big data: Establishing a research agenda. J Intellect Cap 18(2):242–261
56. Shalini RT (2009) Economic cost of occupational accidents: Evidence from a small island economy. Saf Sci 47(7):973–979
57. Shvachko, K, et al. The hadoop distributed file system. In Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on. (2010). Ieee.

58. Sigarari MH, Nazadeh HS, Pourreza H (2013) A review of the content and semantic analysis systems of the video from the point of view of the semantic hierarchy structure in film production. J Visual Machine Image Process 1(2)
59. Tao D, Zhang R, Qu X (2017) The role of personality traits and driving experience in self-reported risky driving behaviors and accident risk among Chinese drivers. Accid Anal Prev 99:228–235
60. Verma S, Bhattacharyya SS, Kumar S (2018) An extension of the technology acceptance model in the big data analytics system implementation environment. Inf Process Manag 54:791–806
61. White, T, Hadoop: The definitive guide. (2015): " O'Reilly Media, Inc.".
62. Yeow PH, Goomas DT (2014) Outcome-and-behavior-based safety incentive program to reduce accidents: A case study of a fluid manufacturing plant. Saf Sci 70:429–437
63. Yi S-C, Chen Y-C, Chang C-H (2015) A lane detection approach based on intelligent vision. Comput Electr Eng 42:23–29
64. Zhang W, Yoshida T, Tang X (2011) A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Syst Appl 38(3):2758–2765

Shiva Asadianfam received her B.Sc. degree in Computer Engineering from Urmia University, Urmia, Iran, and received her M.Sc. in computer engineering (software) from Islamic Azad University, Zanjan Branch, Iran. Currently she is PhD Candidate at Faculty of Computer Engineering in Islamic Azad University, Qom Branch, Iran. Her major research interests are Machine Learning, Data Mining, Statistics, Big Data Analysis and Advanced Product and Process Development. She has also interest in Image Processing and Medical Imaging.

Mahboubeh Shamsi received her B.Sc. degree in applied mathematics from Esfahan University, Iran, and received her M.Sc. in computer engineering (software) from Najafabad University, Isfahan, Iran, and received her Ph.D. degree in computer engineering (software) from Malaysian University of Technology, Malaysia. Currently she is working as an Assistant Professor of Electrical and Computer Engineering Department in Qom University of Technology, Qom, Iran. Her main researches focus on Image Processing and Internet of Things and specifically Cloud Computing and its applications. In addition, she has also interest in Distributed Database Management.

Abdolreza Rasouli Kenari received his B.Sc. degree in applied mathematics from Esfahan University, Iran, and received his M.Sc. in computer engineering (software) from Najafabad University, Isfahan, Iran, and received his Ph.D. degree in computer engineering (software) from Malaysian University of Technology, Malaysia. Currently he is working as an Assistant Professor of Electrical and Computer Engineering Department in Qom University of Technology, Qom, Iran. His research interests include Cloud Computing, Internet of Things and Security. In addition, he has also interest in Database Management and Distributed Database.

Affiliations

Shiva Asadianfam[1] **· Mahboubeh Shamsi**[2] **· Abdolreza Rasouli Kenari**[2]

✉ Mahboubeh Shamsi
shamsi@qut.ac.ir

Shiva Asadianfam
sh_asadianfam_stu@qom-iau.ac.ir

Abdolreza Rasouli Kenari
rasouli@qut.ac.ir

[1] Department of Computer Engineering, Qom Branch, Islamic Azad University, Qom, Iran

[2] Faculty of Electrical & Computer Engineering, Qom University of Technology, Qom, Iran