Convolutional Neural Network and LSTM Applied to Abnormal Behaviour Detection from Highway Footage

Rafael Marinho de Andrade
Instituto Nacional de Pesquisas Espacias – INPE
São José dos Campos, SP, Brasil
rafael.andrade23@fatec.sp.gov.br

Elcio Hideti Shiguemori
Instituto de Estudos Avançados – IEAv
São José dos Campos, SP, Brasil
elcio@ieav.cta.br

Rafael Duarte Coelho dos Santos
Instituto Nacional de Pesquisas Espaciais – INPE
São José dos Campos, SP, Brasil
rafael.santos@inpe.br

ABSTRACT

Relying on computer vision, many clever things are possible in order to make the world safer and optimized on resource management, especially considering time and attention as manageable resources, once the modern world is very abundant in cameras from inside our pockets to above our heads while crossing the streets. Thus, automated solutions based on computervision techniques to detect, react or even prevent relevant events such as robbery, car crashes and traffic jams can be accomplished and implemented for the sake of both logistical and surveillance improvements. In this paper, we present an approach for vehicles' abnormal behaviours detection from highway footages, in which the vectorial data of the vehicles' displacement are extracted directly from surveillance cameras footage through object detection and tracking with a deep convolutional neural network and inserted into a long-short term memory neural network for behaviour classification. The results show that the classifications of behaviours are consistent and the same principles may be applied on other trackable objects and scenarios as well.

KEYWORDS

Artificial Intelligence; Behaviours Detection; Computer Vision; Convolutional Neural Networks; LSTM; Highway Footage.

1 INTRODUCTION

The remote sensing area has benefited for decades from images and data acquired from above ground level, being since then considered essential for applications in that science area. Theapplications make use of those images for information extraction and then decision making. To do so, several technologies forobtaining those data have been applied and developed [1].

An area that can benefit from these solutions is the highway monitoring area. [2] made use of computer vision techniques for vehicles' speed estimation, in order to detect violations on speed limit laws, and [3] applied others computer vision's techniques on a camera monitoring system to identify and count traffic on highways, and [4] had an approach with vehicles' trajectory extraction from aerial images, something relevant for traffic management.

The integration of sophisticated artificial intelligence techniques have shown to be a promising solution for behaviours detection, and can, for example, be applied for prevention and reaction on harmful events where immediate decisions are crucial, like traffic accidents, robbery, fire and the like [5]. Considering road scenarios, such solutions can be applied for detection and reaction to accidents and traffic jams, and if techniques to detect the behaviours of the vehicles on the road be used, it's evenpossible to mobilize preventive actions to such events on relevant behaviour detections such as erratic behaviours (typical of drunk drivers, for example), reckless driving and slowness [6].

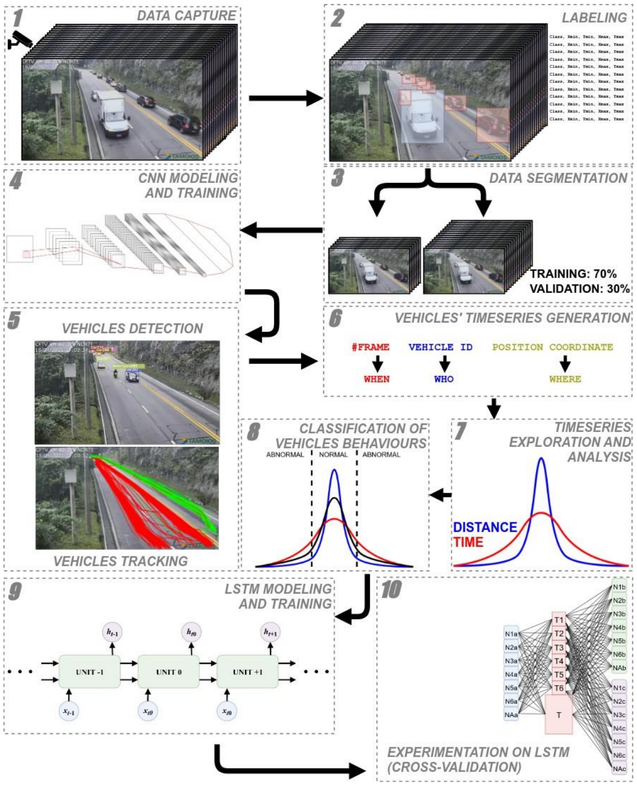


Fig. 1: Methodological process applied in the project.

The development is motivated mainly but not exclusively by the following factors: 1) minimization of quantity and severity of failures on computer vision based security systems by extracting, processing and interpreting data, allowing the identification of more precise informations and then knowledges that allows more appropriate decisions making on these systems; 2) resources optimization through automation, by making the automated systems depend less on monitoring from human operators and by less interventions from them; and 3) the development of computer vision area by the development and practice of skills that allow a better understanding and application of concepts that fit this area.

This project aims to develop an application and method capable of receiving road images from a wide perspective, such asthose made available by surveillance cameras or low-height aerial drones, and detecting occurrences of relevant behaviours of vehicles in them.

2 THEORETICAL BACKGROUND

In this section, we summarize and discuss some theoretical background surrounding the approach applied in this paper.

2.1 Behaviour detection

Considering the idea of "behaviour", there's a wide amount of research around this topic. However, the factual existence of a consolidated and established research area aimed at behaviour detection inside the computer vision area is difficult to notice,once there aren't behaviours' representation standards as the ones that can be found for object detection and classification, for example.

Despite this, there are papers in recent literature detailing the application of computer vision for the detection and identification of behaviours, and it is important to note that these are *ad hoc* applications, thus distancing them from generalism by definition. For example: 1) [4] presented an example where a method was proposed to autonomously extract vehicles' trajectories from aerial images, thus also allowing the analysis of behavioural deviations; 2) [7] presented a paper where long-range cameras were applied to detect suspicious activity on the Texas-Mexico border; 3) the paper of [8], where they presented a system for detecting irregular behaviour in courts; 4) [9] presented an approach to analyzing and detecting sleep-related movement disorders based on nocturnal behaviours captured using sensorson the Kinect One device.

However, it is important to emphasize that the concept of "behaviour" is differently approached between different sciences, which defines and explores them accordingly, so it's essential to disambiguate the concept of "behaviour". It is defined that the concept of "behaviour" of the detected agents (vehicles) considered in this project concerns the attributes referring to pure vectorial physical quantities (such as modulus and direction), combinatorial (speed) and derivatives (displacement, acceleration), in particular its variations in the observed fields (images that compose the footages). Considering this definition, anomaly detection in road traffic is a suitable field for research, and [10] delivered a survey exploring several approaches, in which [11] stands out by trying an approach applying LSTMs to discriminate behaviours; [12] approach defines outliers as abnormal behaviours, identified by K-Means clustering, and applies a Markov Hidden Model to detect them; [13] uses object'sdimensions and displacement to identify vehicles in sidewalks andpersons in roads; and [14] takes speed information to detect road laws infringements. All these approaches make use of surveillancesecurity cameras as image source.

2.2 Machine learning

The machine learning area consists of technologies and methods modeled in a way that allow the tuning of their features in order to deliver improved results. Several machine learning techniques can be applied under the work context.

We shall consider an emphasis on artificial neural networks, especially convolutional networks (CNN) and recurrent networks (RNN), which have proven to be efficient for learning patterns in images and sequential data, respectively, in addition to being able to be applied in other tasks.

Among the convolutional neural networks, LeNET-5 by [15] started a lineage of CNNs for computer vision, and AlexNet by [16] gave rise to what became known as deep convolutional neural networks (DCNN) and represents a watershed in computer vision approaches for detecting and classifying objects in images. Just ahead, the Spatial Pyramid Pooling approach by [17] introduced relevant performance improvements, and Darknet from[18], with the DCNN YOLO, became a popular solution aggregating resources present in the approaches considered asstate of the art. YOLOv4, a modification of YOLO from [18] developed by [19], remains a popular DCNN resource for object detection and classification in images.

Among recurrent neural networks, the networks with long and short-term memory (LSTM) are a common approach for learning sequential data such as time series, audio and video. The LSTMs were first proposed by [20].

2.3 Highway imaging

The presence of highways is a common feature in actual societies, and the existence of laws for their use, resulting from the inherent risks of vehicular traffic, makes surveillance of these environments desirable. The presence of road surveillance cameras seeks to fulfill this purpose, along with other remotesensing approaches for control and surveillance.

Road imaging surveillance can be performed by virtually any imaging device, from smartphone's cameras to drones, but surveillance camera circuits (administered by highway authorities) are a common solution in these societies. These cameras are ideally positioned in locations that are advantageous to the observed environment in order to acquire images with a wide field of view, and the imaging sensors may provide imagesin the visible spectrum or in other spectra, depending on thesurveillance needs.

3 METHODOLOGY

With the purpose of identifying abnormal behaviours performed by vehicles, one of the ways is to present their behavioural characteristics referring to vector quantities as input. The methodological process comprises a series of interactions. One way is to acquire data. Among the various alternatives, the use of varied artificial neural networks and the analysis of derived data stands out. However, in order to train a neural network to discriminate behaviours, one proposal considers that it receives vector data from each vehicle, and for that purpose these vector data must be extracted directly from the video stream input.

The Figure 1 shows the ordered flowchart of themethodological process applied in this project, characterized by 1)images capture, 2) labeling of vehicles present in the images, 3) data segmentation in a training and validation set, 4) modeling of the CNN for vehicles' detection and training with the prepared data, 5) activation of the CNN for vehicles detection in roadvideos and tracking of the detected objects, 6) time series extraction from tracked vehicles, 7) exploration and analysis of time series sets in order to extract behavioural characteristics, 8) classification the time series of vehicles in normal and abnormal behaviours, 9) modeling of LSTMs and training with the time series, and 10) experimentation of the LSTMs with cross validation methods. Therefore, this approach takes features applied by [12-14], considering the footage from surveillance cameras as well from aerial non-static perspectives of a drone.

The first step consists of acquiring the images consumed in the project. These data can be made up of individual images and videos (from which individual images can also be extracted) presenting road scenes in high perspective where the vehicles'flow can be observed, as well as images from cameras of road surveillance. Having the data, it can be analyzed and evaluated about their quality and suitability for the project purposes.

After acquiring the images, an important step is the selection of data to be presented to the object detectors. A common practice is a labeling process where samples are extracted containing vehicles duly discriminated in their classes (cars, trucks, motorcycles, among others). The technique that has stood out is CNN, in which these samples are inserted during training for the detection and classification of vehicles. After labeling and prior to the development process of this network, a process of segmentation into subsets for training and its validation is important, mainly, to employ strategies that provide adequate training of the CNN.

With the samples set ready, a CNN for object detection and classification in images is modeled and trained. This CNN istrained with the collected samples until it reaches a satisfactory performance in the detection of vehicles in the images applied in the project.

The trained CNN, once validated about its ability to detect vehicles in the images, is applied in an application that also is able to track them between the images sequences (process presented in block 5 of the Figure 1) so that each detected vehicle has its own persistent identifier throughout the entire sequence of images where it is present. This application also stores the data of each vehicle in each frame of the video where it is detected, withstorage of the frame number, vehicle identifier and its positional information; therefore, the application stores time series that allow knowing where and when each vehicle was present in the video sequence (block 6 of the Figure 1).

In possession of the time series, they undergo a restructuring, exploration and analysis to extract derived information from the data then collected, which comprise what is defined as the behaviour of the vehicles referring to them. Among the relevant derivative information that can be extracted, the speed and direction of displacement of each vehicle stand out, as well as how straightforward are such displacements. After the exploration process of the collected data, the extracted information is analyzed together to define the thresholds that separate behaviours considered normal from those abnormal (block 8 of the Figure 1).

Then the LSTMs are trained, a process that starts with the classification of the time series based on the aforementioned criteria and then their segmentation in training and validation subsets. The LSTM networks then receive these properly structured and labeled data in order to learn with them how to discriminate the so-called normal and abnormal behaviours. Several LSTM networks can be modeled and trained, thus comprising experiments in search of the best results.

Finally, the LSTM networks must have their performance validated and evaluated. For this, a cross validation process is applied, comprising several distinct subsets and also combined in the various LSTM networks that have been developed (block 10 of the Figure 1).

4 DEVELOPMENT AND EXPERIMENTS

In this session, we detail the development process, in which the aforementioned methodology was applied.

4.1 Data capture

The data captured for the project are several videos presenting road scenarios, with a high perspective to highways and vehicles from different angles of sight and focal length, in different places and weather conditions. The Figure 2 shows some pictures from the carried out footages. All videos used for behaviours discrimination were captured by surveillance road cameras andare under public domain. Data were obtained from a highway in the region of Vale do Paraíba, in the State of São Paulo, Brazil (Figure 3).

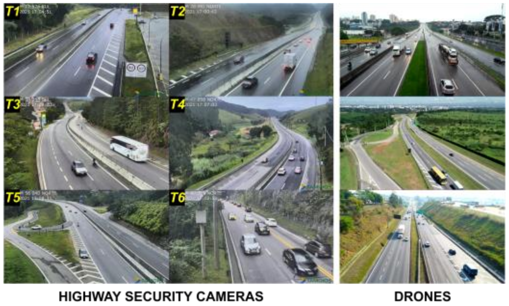


Fig. 2: Excerpts from the footages used in the project, where the six images on the left refer to six road surveillance cameras (datasets *T1* to *T6*) and the three images on the right refer to some footages taken by drone. Each surveillance camera is located between about 11 km from each other.

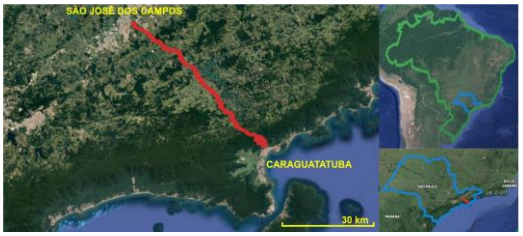
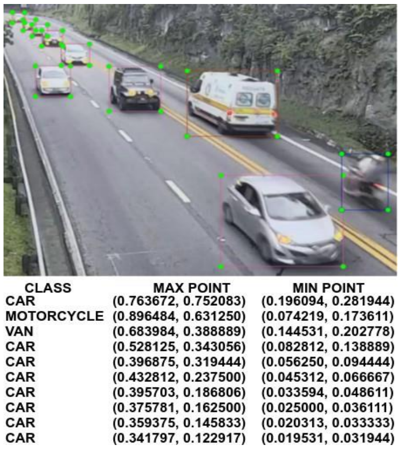


Fig. 3: Map highlighting the Tamoios Highway (line in red), where the images were captured, between the cities of São José dos Campos and Caraguatatuba. On the right side, there are perspectives showing the state of São Paulo (bluesilhouette) and Brazil (green silhouette).

For the formation of training sets, all videos were analyzed and some arbitrarily selected frames were extracted under the criteria of having at least one vehicle noticeable at glance, with the objective of building a dataset with wide variety of vehicles. Only the videos captured by the road cameras were used in such format in the final application; there are six videos with 10 minutes each, adding up to exactly 1 hour of video. In all, 2 hours, 6 minutes and 29 seconds of video were captured, from which 411 images were extracted for the process of training set generation.

4.2 Training set

The images extracted from the videos may present up to a few dozen vehicles of different discriminatory classes. Each of these images was then submitted to a labeling process where the instances of these vehicles are marked and classified (Figure 4), thus resulting in the extraction of samples to compose the training set; this labeling process was performed with LabelImg [21]. Vehicles labeled are classified as car, motorcycle, truck, pickup truck, van, bus, bicycle, tractor or airplane (although these last two aren't present in the videos where the network is finally applied).



CLASS	MAX POINT	MIN POINT
CAR	(0.763672, 0.752083)	(0.196094, 0.281944)
MOTORCYCLE	(0.896484, 0.631250)	(0.074219, 0.173611)
VAN	(0.683984, 0.388889)	(0.144531, 0.202778)
CAR	(0.528125, 0.343056)	(0.082812, 0.138889)
CAR	(0.396875, 0.319444)	(0.056250, 0.094444)
CAR	(0.432812, 0.237500)	(0.045312, 0.066667)
CAR	(0.395703, 0.186806)	(0.033594, 0.048611)
CAR	(0.375781, 0.162500)	(0.025000, 0.036111)
CAR	(0.359375, 0.145833)	(0.020313, 0.033333)
CAR	(0.341797, 0.122917)	(0.019531, 0.031944)

Fig. 4: Labeling example, which consists of delimiting image samples between a pair of coordinates (maximum and minimum point of the bounding box) and assigning classes to them. Coordinates are stored in relative values (between 0 and 1), making the sampling invariant to scales and reusable in cases of image resizing.

In the labeling process, the delimitation of each instance tightly covers all the pixels referring to it; each sample is therefore characterized by its class and maximum and minimum point coordinates of its bounding box in the image where it is present. The labeling also followed some definition criteria, where vehicle drivers are also marked as part of the instance (important in cases of motorcycles and bicycles) and the trunks and payloads of trucks and pickup trucks are also considered part of the vehicle rather than separate instances. For greater rigor, this entire processwas carried out by just one person who delimited and labeled eachinstance of vehicle that could be perceived in the images.

The set of labeled images comprises the training set for CNN training. Finally, the entire set of samples was segmented into a training and validation subset with a ratio of 70% and 30%, respectively.

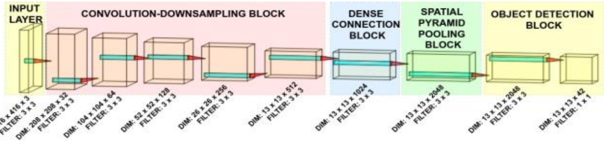

INPUT LAYER
CONVOLUTION-DOWNSAMPLING BLOCK
DENSE CONNECTION BLOCK
SPATIAL PYRAMID POOLING BLOCK
OBJECT DETECTION BLOCK
DIM: 416 x 416 x 3 FILTER: 3 x 3
DIM: 208 x 208 x 32 FILTER: 3 x 3
DIM: 104 x 104 x 64 FILTER: 3 x 3
DIM: 52 x 52 x 128 FILTER: 3 x 3
DIM: 26 x 26 x 256 FILTER: 3 x 3
DIM: 13 x 13 x 512 FILTER: 3 x 3
DIM: 13 x 13 x 1024 FILTER: 3 x 3
DIM: 13 x 13 x 2048 FILTER: 3 x 3
DIM: 13 x 13 x 2048 FILTER: 3 x 3
DIM: 13 x 13 x 42 FILTER: 1 x 1

Fig. 5: Diagram with YOLOv4 architecture.

4.3 CNN modeling and training

The chosen CNN for detection of the instances of vehicles present in the images is the YOLOv4, a deep convolutional neural network with 137 layers and whose architecture is represented in the diagram in the Figure 5. Prior to the training, the network architecture and training parameters were adjusted to best fit the training set formed, as detailed in Table I.

The network training was carried out with the Darknet framework, and the network was trained for 18000 epoches (as set by max batches), finally reaching a mAP (mean Average Precision) of 90.4% and a loss of 0.4942. Throughout the training process, the Darknet algorithm also applies Cross mini-Batch Normalization, which contributes to increasing the training performance, as explained in [19].

Parameter	Value
Input layer dimensions	416 x 416, with 3 color
Batches	64
Subdivisions	16
Momentum	0.949
Decay	0.0005
Learning rate	0.0013
Burn-in	1000
Max batches	18000
Steps	14400 to 16200

Table 1: Parameters used in the trained CNN.

After the training, the YOLOv4 CNN was applied to an algorithm for vehicles' detection in the captured videos, in order to evaluate its performance in such a task. As a result, the trained CNN has proven itself capable of detecting all vehicles in the field of view within several meters of the camera, delimiting themtightly and consistently between the video frames.

4.4 Vehicle detection and tracking

For this paper, an application for detecting and tracking objects in videos was developed. This application has as input the video where the objects must be detected and tracked and the CNN trained to detect such objects. With that, the application reads the video frame by frame and applies the CNN in these frames to detect and classify the vehicles presented in them, storing, therefore, the class and coordinates of the maximum andminimum points of each detected instance.

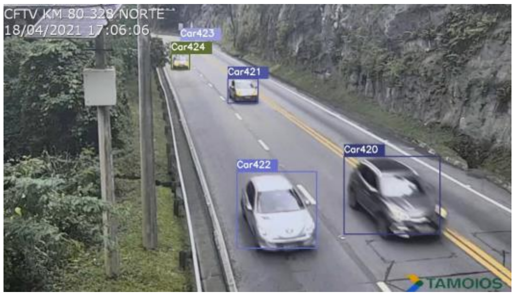


Fig. 6: An example frame from a output video generated by the algorithm, showing instances of detected vehicles with bounding boxes, class discrimination and assignment of identifiers.

More than just detecting, the application is also able to track objects between frames. For this, the Deep SORT framework by [22] is used, which assigns the same identifier to instances between different frames based on the displacement distance and visual similarities. The distance is calculated using the Mahalanobis distance equation, which has the advantage of being invariant to scales, and the comparison of the visual characteristics of the instances is performed by a deep association metric model, which applies a Kalman filter for error minimization and a Hungarian method for optimization in theassociation. Using Deep SORT also makes the classification moreconsistent.

The algorithm also has graphical computation functions to generate an output video where the detected instances are highlighted (their bounding boxes, classes and identifiers), as can be seen in the Figure 6, and a time series with the detections is generated. Therefore, the algorithm works as the following:

```
1. Loads video;
2. Loads CNN;
3. Initialize tracker;
4. Initialize video and time series writers;
5. Reads video, frame by frame, in which each frame:
    5.1. Detect the vehicles;
    5.2. Compares the detected instances in the actual frame with the
         instances in previous frames:
        5.2.1. If it's an instance of a new vehicle, assigns a new identifier
               to it;
        5.2.2. If it's an instance of a previously detected vehicle, assigns
               the same known identifier to it.
    5.3. Draws the bounding box of each instance in the image with a label
         with its class and identifier;
    5.4. Stores the detected instances in a time series.
6. Closes the writers.
```

4.5 Time series formation

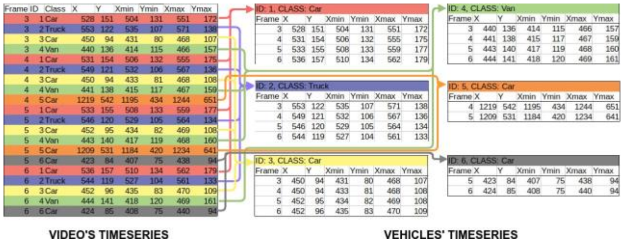
As evidenced in the aforementioned algorithm, each instance of vehicle detected in the video is inserted into a time series. These time series are two-dimensional structures where each line is an instance detected in the video and the columns store information of the frame where it was detected, the vehicle identifier and the coordinates of the maximum, minimum point and centroid of the bounding box. Therefore, with this the time series contains information to know where and when each vehicle was detected in the video, also with a notion of vehicle size based on the dimensions of the bounding box as also [13] did.

The time series generated by the application, however, are formed with the instances of all vehicles, while it is desirable that each vehicle has its own time series and so that it be possible to do an analysis. Then, after the formation of the time series of each video, the data is segmented by identifier, so that each vehicle detected in the video has its own time series. A schematic of this process can be seen in Figure 7.

As each detected vehicle had its own time series, one of the first things noticed was a varying length of the resulting time series, as a consequence of the fact that the vehicles appeared in different amounts of frames. So that all time series had the same dimensions, the lines formed by the positional coordinates were interpolated in 1000 values, under the criteria of cubic interpolation.

It is important to take into account that the amount of values for the interpolation must not be less than the number of frames present in the videos, thus guaranteeing that even a vehicle that was present in all frames wouldn't provide a larger amount of values than the interpolation, thus avoiding loss of vectorial information. Still, one of the side effects of interpolation is theloss of temporal information and therefore speed information for each vehicle (here treated simply as pixels per frame rather than real world units). Therefore, before the interpolation, the vehicle speeds were extracted and stored as a new column.

Finally, all time series were gathered in a single data structure for analysis of behavioural profiles as a whole.



Frame	ID	Class	X	Y	Xmin	Ymin	Xmax	Ymax
3	1	Car	528	151	504	131	551	172
3	2	Truck	553	122	535	107	571	138
3	3	Car	450	94	431	80	468	107
3	4	Van	440	136	414	115	466	157
4	1	Car	531	154	506	132	555	175
4	2	Truck	549	121	532	106	567	136
4	3	Car	450	94	433	81	468	108
4	4	Van	441	138	415	117	467	159
4	5	Car	1219	542	1195	434	1244	651
5	1	Car	533	155	508	133	559	177
5	2	Truck	546	120	529	105	564	134
5	3	Car	452	95	434	82	469	108
5	4	Van	443	140	417	119	468	160
5	5	Car	1209	531	1184	420	1234	641
5	6	Car	423	84	407	75	438	94
6	1	Car	536	157	510	134	562	179
6	2	Truck	544	119	527	104	561	133
6	3	Car	452	96	435	83	470	109
6	4	Van	444	141	418	120	469	161
6	6	Car	424	85	408	75	440	94

VIDEO'S TIMESERIES

ID: 1, CLASS: Car

Frame	X	Y	Xmin	Ymin	Xmax	Ymax
3	528	151	504	131	551	172
4	531	154	506	132	555	175
5	533	155	508	133	559	177
6	536	157	510	134	562	179

ID: 2, CLASS: Truck

Frame	X	Y	Xmin	Ymin	Xmax	Ymax
3	553	122	535	107	571	138
4	549	121	532	106	567	136
5	546	120	529	105	564	134
6	544	119	527	104	561	133

ID: 3, CLASS: Car

Frame	X	Y	Xmin	Ymin	Xmax	Ymax
3	450	94	431	80	468	107
4	450	94	433	81	468	108
5	452	95	434	82	469	108
6	452	96	435	83	470	109

ID: 4, CLASS: Van

Frame	X	Y	Xmin	Ymin	Xmax	Ymax
3	440	136	414	115	466	157
4	441	138	415	117	467	159
5	443	140	417	119	468	160
6	444	141	418	120	469	161

ID: 5, CLASS: Car

Frame	X	Y	Xmin	Ymin	Xmax	Ymax
4	1219	542	1195	434	1244	651
5	1209	531	1184	420	1234	641

ID: 6, CLASS: Car

Frame	X	Y	Xmin	Ymin	Xmax	Ymax
5	423	84	407	75	438	94
6	424	85	408	75	440	94

VEHICLES' TIMESERIES

Fig. 7: Schematic of the process of segmenting video timeseries into different vehicle time series. It can be also noticed that in this segmentation occurs the discretization of vehicle identifiers and their classes. The scheme also shows that the time series dimensions vary after this process, as vehicles appear in varying amounts of frames.

4.6 Time series analysis

Assuming that most drivers perform behaviours considered normal, the time series were analyzed with the main intent of finding outliers, like [12] once did. Based on this, some metrics based on vectorial physical characteristics that comprise the vehicle displacements were considered for the analysis, whichwere then weighted for a final thresholding.

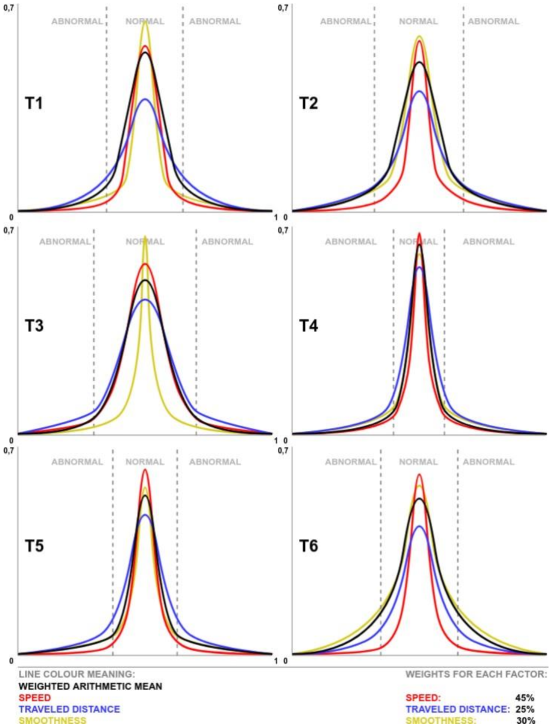


Fig. 8: Gaussian distribution of the observed factors in each dataset: speed, traveled distance and displacement smoothness (lines in red, blue and yellow, respectively), which exert an influence of 45%, 25% and 30% on the weighted arithmetic means (black lines), respectively. The dashed lines over the distributions are the classification thresholds, where the central values are defined as normal behaviour and the lateralvalues are defined as abnormal behaviour.

These metrics were distributed in Gaussian models and, based on that, the thresholds that separate normal and abnormal behaviours (which can be seen in the Figure 8) were defined. Differently of [12], however, the vectorial characteristics considered are the speed, the distance traveled and the smoothness of the displacement, which exerted an influence on the weighting in the arithmetic mean of 45%, 25% and 30%, respectively. Therefore, the premise considers vehicles that stand out from the others when they travel much slower or much faster, travel much more or much less space to move between the limits of the field of view or move with much more or much less smoothness are performing abnormal behaviours.

Since each video is considered a different dataset, as it presents varied perspectives and slightly different road characteristics (being some straights and some with corners, also as some with multiple ways, as can be seen in the Figure 2), the thresholds that separate normal from abnormal behaviours differ according to the video where the time series were extracted. Still, the thresholding criteria remains the same regardless of the distribution profile. Finally, once the thresholds that separate the behavioural profiles considered normal and abnormal weredefined, the analyzed time series were labeled according to this and, as well as the data set for CNN training, were segmented into training and validation in proportions of 70% and 30%, respectively. A mixed dataset was also generated, containing a balanced mixture of all six prepared datasets and which was also segmented with the same proportion for the same purposes.

4.7 LSTMs modelling and training

Following the premise of discriminating behaviours considered normal and abnormal in road images, neural networks designed for this purpose were modeled and trained. In order to experiment the approach, different neural network architectures were selected and implemented, all of them being LSTMs. The use of LSTM for classification is motivated by its premise of specially processing sequential data (such as time series) and anomalies detection, and by its sensibility to discrete features rather than the only defined by a human being.

Three different LSTM architectures were modeled, all of which receive time series containing vectorial information (speed and maximum, minimum and center point coordinates of the instance) as input and process them to return a binary value.where 1 refers to normal behaviour and 0 refers to abnormalbehaviour. The first architecture has four layers, with 16 units in the first layer, 8 in the second, 4 in the third and 2 in the output layer; the second architecture has five layers, with 32 units in the first layer, 16 in the second, 8 in the third, 4 in the fourth, and 2 inthe last layer; and the third architecture has four layers, with 128 units in the first layer, 32 in the second, 8 in the third, and 2 in the fourth. These architectures were respectively named as *16-8-4-2*, *32-16-8-4-2* and *128-32-8-2*. Having three different input sizes allow us to check if more or less timespan is needed to deliver a better discrimination, and graphic representations of these architectures can be seen in Figure 9.

A network of each architecture was modeled for each dataset (including the mixed set). Therefore, 21 LSTM networks were trained, in training sessions with up to 1000 epochs under the condition of stopping if the loss didn't decay for 10 epochs straight.

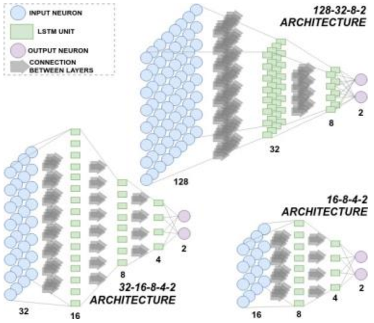


Fig. 9: Graphical representation of the modeled architectures, showing the difference in the complexity (amount of input data and its funneling) of each one, being theupper *128-32-8-2*, the bottom-left *32-16-8-4-2* and the bottom- right 1*6-8-4-2*.

5 RESULTS ANALYSIS

The 21 LSTM networks were extensively tested following permutations with the available test sets, in order to classify the networks on their ability to discriminate behavioural profiles as normal and abnormal. Therefore, there were 147 executions of LSTM networks, whose results, in addition to choosing the most appropriate network to discriminate behaviours in the discussed scenarios, allow to measure both the quality of the implemented architectures and the representativeness of each training set over the full set.

Since the networks perform only binary classifications, the metrics of the results are discretized to only the accuracy in the discrimination of each class (normal and abnormal) and the general accuracy (the mean between both). The results, however, vary considerably between high and low values among the applied training sets, so the generalized results hide cases of both good and bad performance.

5.1 Architecture analysis

Considering all training sets, we theoretically can infer which of the modeled architectures achieve the best performance. The results indicate, however, an overall accuracy among all architectures of 66.38%, ranging between 67.87% and 63.84%; these results can be seen in the Table 2.

Architecture	Normal	Abnormal	General
General	68.19	62.28	66.38
16-8-4-2	**70.31**	50.96	63.84
32-16-8-4-2	69.62	60.38	**67.87**
128-32-8-2	64.63	**75.48**	67.44

Table 2: Mean accuracy results for each architecture on all training sets.

Comparing the networks in their overall ability ofdiscriminating normal behaviours, the distance is slightly greater, with an overall accuracy of 68.19%, ranging between 70.31% and 64.63%. The ability of discriminating abnormal behaviours showed the most accentuated variation, with an overall accuracyof 62.27% and a variation between 75.38% and 50.96%.

Thus, we can assume that, among the three modeled architectures, the choice of architecture is a less deterministic factor in the final performance, as the amount of input data and its funneling. The LSTM *32-16-8-4-2* network achieved slightly better results despite to this.

5.2 Architecture analysis

Considering all architectures, we can infer, also in theory, which training sets best represent the integral set. The mixed set was also evaluated, with the aim of showing those representations and how universal training sets tend to be in broader scenarios.

Dataset	Normal	Abnormal	General
T1	26.90	**87.96**	46.72
T2	**83.83**	68.87	**79.86**
T3	58.70	57.86	55.59
T4	63.59	55.36	60.14
T5	80.64	35.70	66.16
T6	65.96	53.78	64.12
Mixed	**97.69**	76.37	**92.09**

Table 3: Mean accuracy results for each training set in all architectures.

As can be seen in the Table 3, the results in the mixed set level the overall accuracy at 92.09%, well above the best result in the isolated sets, with 79.86% accuracy. The worst result is evenlower, with 46.72% accuracy.

Comparing by the ability of discriminating normal behaviours, the accuracy in the mixed set reached a high 97.69%, above the 83.83% of the best isolated set and far from the 26.90% of the worst set. As for abnormal behaviours, the mixed set levels the accuracy at 76.37% and the best result surpasses this mark, withan accuracy of 87.96%, while the worst result is at 35.70%.

Anyway, no isolated set seems to represent the integral set as the mixed set does. Despite this, networks trained only with the training set from the *T2* scenario has the best overall accuracy, but with a disparity when considering only the ability to discriminate between normal and abnormal behaviours, where the sets of the *T2* and *T1* scenarios achieved the best results, respectively.

5.3 Training sets analysis by architecture

Looking at the results more closely allows us to better understand their nuances, so that the influence of the architecture for each training set becomes more evident. The evaluation of the results also considers the leveling of performance from the results in the mixed set in comparison with the isolated sets.

Dataset	Normal	Abnormal	General
T1	38.42	**67.93**	48.16
T2	**91.85**	60.77	**83.36**
T3	54.99	63.43	52.87
T4	56.10	65.31	57.82
T5	84.00	0.71	57.23
T6	70.59	32.94	59.53
Mixed	**96.25**	65.60	**87.90**

Table 4: Accuracy results for each training set in *16-8-4-2* architecture.

With the *16-8-4-2* architecture, whose results can be seen in the Table 4, the overall accuracy in the mixed set was 87.90%, while the best result in a isolated set was close, with 83.36%, and far from the worst result, of 48.16%.

Based only on the ability of discriminating normal behaviours, the mixed set reached 96.25% of accuracy, slightly above the 91.85% of the best result in a isolated set and far from the 38.42% of the worst result. Regarding abnormal behaviours, the accuracy in the mixed set reached 65.60%, slightly below the 67.93% of the best set and well above the 0.71% of the worst set.

With the *16-8-4-2* architecture, the mixed set is what seems to best represent the integral set, being slightly better at this than the *T2* scenario set (which is also the one that best performs the discrimination of normal behaviours). The *T1* scenario had the worst overall accuracy and, mainly, in the discrimination of normal behaviours, despite the better performance in the discrimination of abnormal behaviours.

Dataset	Normal	Abnormal	General
T1	20.47	**96.23**	45.18
T2	**95.03**	52.11	**84.22**
T3	70.99	50.73	63.51
T4	78.35	41.67	67.01
T5	66.36	42.62	59.02
T6	58.18	70.59	65.67
Mixed	**97.98**	68.69	**90.50**

Table 5: Accuracy results for each training set in *32-16-8-4-2* architecture.

Available in the Table 5, the results with the *32-16-8-4-2* architecture present an overall accuracy in the mixed set of 90.50%, while the best results in a isolated set were close, at 84.22%. The worst result was 45.18%.

In the discrimination of normal behaviours, the mixed set reached 97.98% of accuracy, slightly above the 95.03% of the best result in a isolated set and well above the 20.46% of the worst result. Regarding abnormal behaviours, the accuracy in the mixed set reached only 68.69%, well below the 96.23% for the best set halfway up to 41.67% for the worst set.

As with the *16-8-4-2* architecture, the results with the *32-16-8-4-2* architecture also point to the mixed set as what seems to best represent the integral set, also slightly better than the set of the *T2* scenario (which is also the one with the best results in detecting normal behaviours). The *T1* scenario had the worst overall accuracy, again mainly in the discrimination of normal behaviours, despite the better performance in the discrimination ofabnormal behaviours.

The results with the *128-32-8-2* architecture, whose can beseen in the Table 6, present an overall accuracy in the mixed setof 97.87%, with the best and worst results in the isolated reaching 82.24% and 46.82%, respectively.

Limiting the analysis to the discrimination of normal behaviours, the mixed set reached an accuracy of 98.84%, slightly above the 91.57% of the best result in a isolated set and far from the 21.46% of the worst result. With the abnormal behaviours, the accuracy in the mixed set reached 94.83%, slightly below the 99.71% for the best set and still far from 57.80% for the worst set.

Dataset	Normal	Abnormal	General
T1	21.81	**99.71**	46.82
T2	64.62	93.72	71.98
T3	50.11	59.43	50.38
T4	56.33	59.11	55.59
T5	**91.57**	63.77	**82.24**
T6	69.12	57.80	67.16
Mixed	**98.84**	94.83	**97.87**

Table 6: Accuracy results for each training set in *128-32-8-2* architecture.

The *128-32-8-2* architecture also presents the mixed set as the most representative of the integral set, with the set of the *T5* scenario being the closest (and also with the best results in the detection of normal behaviours). The *T1* scenario again presented the worst overall accuracy, also mainly due to the worst result in the discrimination of normal behaviours, even though it reached almost 100% in the discrimination of abnormal behaviours.

Finally, analyzing the results more closely, we can assume that the set of scenario *T2* is the most representative of the integral set, followed by the set of scenario *T5*. There is a disparity, however, when considering that the best results for the discrimination of abnormal behaviours was with the *T1* set, which presented results with low accuracy for the discrimination of normal behaviours. The other sets showed more balanced accuracy levels for both classes, although without showing the highest accuracy values.

Regardless of the evident accuracy values in the results, it is notable that they are consistent across training sets and architectures, which means that LSTM networks have indeed learned behavioural profiles consistently.

Finally, considering the search for the network with the best results, the 128-32-8-4 network trained with the mixed set reached an overall accuracy of 97.87%, with an accuracy of 98.84% in the discrimination of normal behaviours and 94.83% in the discrimination of abnormal behaviours.

6 CONCLUSION

In this paper, a set of LSTM neural networks were trained to discriminate vehicle behaviours. The behaviours are represented as time series with vehicle vectorial displacement information, formed from their tracking in road surveillance footage. Following the presented methodological process, it was possibleto detect abnormal behaviours in road images, as well as normal ones. However, we can also assume that the datasets are notbalanced in order to ensure the best learning of LSTM networks, regardless of the implemented architecture. Nevertheless, the results are consistent, and the accuracy over 50%, achieving a general accuracy around 66%, makes the LSTMs more reliable than just chance, so we can also assume that LSTM networks are able to learn behavioural profiles consistently.

The developed system can be enhanced with more extensive and diverse datasets in order to make the trained LSTM networks more robust. The project can also be expanded with differentapproaches or other types of scenarios and objects of interest, and may be also applied to detect collective behaviors.

ACKNOWLEDGEMENTS

We would like to thank Concessionária Tamoios for having provided the recordings of the surveillance camera circuit of Rodovia dos Tamoios, which were used throughout the project. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 88887.483839/2020-00.

REFERENCES

[1] Jeziorska, J. Uas for wetland mapping and hydrological modeling. Remote Sensing, v. 11, p. 1997, 08 2019.

[2] Krishna; Poddar, M.; Giridhar M K; Prabhu, A. S.; Umadevi V. Automated traffic monitoring system using computer vision. In: 2016 International Conference on ICT in Business Industry Government (ICTBIG). [S.l.: s.n.], 2016. p. 1–5.

[3] Opatha, R.; Peiris, A.; Gamini, D.; Edirisuriya, A.; Athuraliya, C.; Jayasooriya, I. Automated traffic monitoring for complex road conditions. IDRC Grant/ Subvention du CRDI: 108008-001-Leveraging Mobile Network Big Data for Developmental Policy, 03 2018.

[4] Lima, A. C.; Costeira, J. P.; Marques, M.; Ben-Akiva; E, M. Automatic vehicle trajectory extraction by aerial remote sensing. Procedia - Social and Behavioral Sciences, v. 111, p. 849–858, 2014. ISSN 1877-0428.

[5] Kishi, R.; Yamamoto, K.; Huy, P. T.; Masuda, M. Abnormal Behaviour Detection using Image Sensing. [S.l.: s.n.], 05 2019. v. 86, n. 1.

[6] Shahverdy, M.; Fathy, M.; Berangi, R.; Sabokrou, M. Driver behavior detection and classification using deep convolutional neural networks, Expert Systems with Applications, Volume 149, 2020, 113240, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113240.

[7] Wei, H.; Laszewski, M.; Kehtarnavaz, N. Deep learning-based person detection and classification for far field video surveillance. In: 2018 IEEE 13th Dallas Circuits and Systems Conference (DCAS).

[8] Liu, S.; He, Q.; Wang, Z.; Pu, Y.; Zhang, Y. Irregular action recognition in court with 3d residual network. In: 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA).

[9] Gall, M.; Garn, H.; Kohn, B.; Bajic, K.; Coronel, C.; Seidel, S.; Mandl, M.; Kaniusas, E. Automated detection of movements during sleep using a 3d time-of-flight camera: Design and experimental evaluation. IEEE

[10] Santhosh, K. K.; Dogra, D. P.; Roy, P. P. Anomaly Detection in Road Traffic Using Visual Surveillance. ACM Computing Surveys, v. 53, n. 6, p. 1-26, 29 dez. 2020. Available in: https://doi.org/10.1145/3417989.

[11] Medel, J. R.; Savakis, A. Anomaly detection in video using predictive convolutional long short-term memory networks.

[12] Cai, Y. et al. Trajectory-based anomalous behaviour detection for intelligent traffic surveillance. IET Intelligent Transport Systems, v. 9, n. 8, p. 810-816, 1 out. 2015.

[13] Basharat, A.; Gritai, A.; Shah, M. Learning object motion patterns for anomaly detection and improved object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[14] Giannakeris, P. et al. Speed Estimation and Abnormality Detection from Surveillance Cameras. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[15] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, v. 86, p. 2278 – 2324, 12 1998.

[16] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet classification with deep convolutional neural networks. In: PEREIRA, F.; BURGES, C. J. C.; BOTTOU, L.; WEINBERGER, K. Q. (Ed.). Advances in Neural Information Processing Systems. Curran Associates, Inc., 2012. v. 25.

[17] He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. Lecture Notes in Computer Science, Springer International Publishing, 2014, Pages 346–361, ISSN 1611-3349, http://dx.doi.org/10.1007/978-3-319-10578-9_23.

[18] Redmon, J.; Divala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. 06 2015.

[19] Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. 2020.

[20] Hochreiter, S.; Schmidhuber, J. (1997). Long Short-term Memory. Neural computation. 9. 1735-80. 10.1162/neco.1997.9.8.1735.

[21] Tzutalin. LabelImg. 2015. Git code. Available at <https://github.com/tzutalin/labelImg>.

[22] Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In: . [S.l.: s.n.], 2017. p. 3645–3649.