ZANCO Journal of Pure and Applied Sciences
The official scientific journal of Salahaddin University-Erbil
https://zancojournals.su.edu.krd/index.php/JPAS

ISSN (print):2218-0230**, ISSN (online):** 2412-3986**, DOI: http://dx.doi.org/10.21271/zjpas**

RESEARCH PAPER

Optimized frame detection technique in vehicle accident using deep learning

Mardin A. Anwer[1], Shareef M. Shareef[1], Abbas M. Ali[1]

[1]Department of Software and Informatics, College of Engineering, Salahaddin University-Erbil, Kurdistan Region, Iraq

A B S T R A C T:

Video processing becomes one of the most popular and needed steps in machine leering. Todays, Cameras are installed in many places for many reasons including government services. One of the most applications for this concern is traffic police services. One of the main problems of using videos in machine learning application is the duration of the video; which is consuming time, paperwork and space in processing. This leads to increase the computation cost through a high number of frames. This paper proposes an algorithm to optimize videos duration using a Gaussian mixture model (GMM) method for real accident video. The Histogram of Gradient (HoG) has been used to extract the features of the video frames, a scratch CNN has been designed and conducted on two common datasets; Stanford Dogs Dataset (SDD) and Vehicle Make and Model Recognition Dataset (VMMRdb) in addition to a local dataset that created for this research. The experimental work is done in two ways, the first is after applying GMM, the finding revealed that the number of frames in the dataset was decreased by nearly 51%. The second is comparing the accuracy and complexity of these datasets has been done. Whereas the experimental results of accuracy illustrated for the proposed CNN, 85% on the local dataset, 85% on SDD Dataset and 86% on VMMRdb Dataset. However, applying GoogleNet and AlexNet on the same datasets achieved 82%, 79%, 80%, 83%, 81%, 83% respectively.

KEY WORDS: Intelligent transportation system, Object detection, video processing technique, video segmentation, Gaussian mixture model, transfer learning, deep learning, GoogleNet, AlexNet.
DOI: http://dx.doi.org/10.21271/ZJPAS.32.4.5
ZJPAS (2020) , 32(4);38-47 .

1.INTRODUCTION:

Many applications on Intelligent Transport Systems (ITS) focus on three missions which are detecting (Minaee et al., 2015), tracking and recognizing vehicles in image sequences. Most of these applications apply their proposed algorithms and techniques on common datasets such as SUN2012 (Clady et al., 2008), 101_Object Categories (Ranganatha and Gowramma, 2018) and CIFAR-10 (Hasanpour et al., 2018) and COLD and IDOL (Salih and Ali, 2019)

Deep learning algorithms perform well when the dataset is large because these algorithms are data dependencies to understand it perfectly (Anwer et al., 2019). However, minimizing the number of frames in each video will increase the number of videos in the dataset. CNN replaces the standard methods of video classification which consist of three stages with a single neural network that is trained end to end from raw pixel values to classifier outputs (Niebles et al.,2010). Currently, there are no video classification benchmarks that coordinate the scale and variety of existing picture datasets since recordings are essentially more troublesome to gather, clarify and store. Converting video to frames is a fundamental step when dealing with video processing. In this research, MATLAB Deep learning tool has been used to accomplish the proposed CNN. Deep and

* **Corresponding Author:**
Mardin A. Anwer
E-mail: mardin.anwer@su.edu.krd
Article History:
Received: 30/12/2019
Accepted: 22/02/2020
Published: 08/09 /2020

transfer learning algorithms depend heavily on high-end machines, contrary to traditional machine learning algorithms, which can work on low-end machines. This is as a result of the necessities of deep learning formula embrace GPUs that are associate degree integral part of its operating. Deep learning algorithms inherently do an oversized quantity of matrix operations. These operations are with efficiency optimized employing a GPU. Hence, GPU is made for working in machine learning. We had many alternatives such as using cloud computing for deep learning or on-premises. However, in this research the implantation environment was a computer with 8 GB RAM, CPU @2.40GH and MATLAB Simulink with deep learning toolbox are used to conclude the results. The paper is organized as follows: section on is an introduction, section two is about the related studies and literature review. Methodology explained in section three. Section four clarify results and discussion. We finish with conclusions and future work.

1.1. Gaussian mixture model

A Gaussian mixture (GM) is characterized as a raised combination of Gaussian densities (Jakob et al. ,2003). A Gaussian mixture model (GMM) is valuable for modelling data that comes from one of several groups: the groups could be diverse from each other, but data points within the same group can be well-modelled by a Gaussian distribution (Kolekar, 2010). Image is divided up into a mixture of pixel cells. The esteem of the pixel may be a number that appears the intensity or color of the picture. Let X be a irregular variable that takes these values. For a probability model determination, we assumed to have a mixture of Gaussian distribution as the following:

$$f\,(x) = \sum_{i=1}^{k} p_i \; N(x|\,\mu_i, \sigma_i^2) \qquad (1)$$

Where k is the number of components or regions and p_i>0

weight such that $\sum_{i=1}^{k} p_i = 1$,

$$N(\mu_i, \sigma_i^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_i)^2}{2\sigma_i^2}\right) \qquad (2)$$

where μ_i, σ_i^2 are mean and standard deviation of class i. For a given image X, the lattice data are the values of pixels and MoG is our pixel-based model. However, the parameters are $\theta = (p_1, \dots, p_k, \dots, \mu_1, \dots, \mu_k, \dots, \sigma_1^2, \dots, \sigma_k^2)$ and we can guess the number of regions in MoG by the histogram of lattice data.

1.2 The Histogram of Gradient

HOG features provide a concise but powerful image representation for general object classification. They have found significant application in the field of pedestrian identification (Navneet and Bill, 2005), where they have continued to provide one of the more robust feature extraction techniques even in recent analyses. HOG features are based on gradient angle and magnitude distributions, and in visual data, they are robust due to the gradient's natural invariance to slight changes in ambient lighting and colour variations. Equation number 3 is used to find the magnitude and direction of the gradient.

$$g = \sqrt{g_x^2 + g_y^2} \qquad (3)$$

$$\theta = arctan\frac{g_y}{g_x} \qquad (4)$$

1.3 Google net Framework

GoogleNet described as a Networks with Parallel Concatenations. It proposed an architecture that associates the strengths of the NiN and repeated blocks paradigms [9]. Their architecture involved twenty-two deep CNN layers with fifty-six million parameters less than AlexNet (Christian et al.,2015). Figure 1 shows the construction of the GoogleNet.

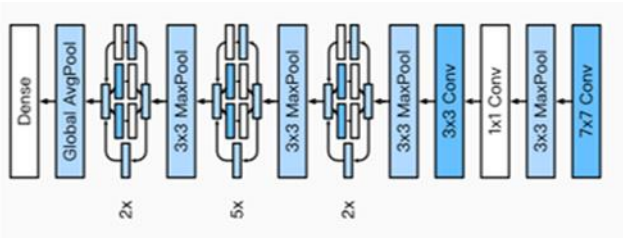


Figure 1: Full GoogLeNet Model (Das., 2017)

1.4 AlexNet framework

AlexNet is one of the most widely-used architecture as CNN models (Krizhevsky et

al.,2012). It consists of eight layers which are: convolution, max pooling and fully connected. AlexNet used non-linearity (Relu) activation function because of better training performance than tanh and sigmoid. Figure 2 shows the construction of AlexNet. It is important to highlight that AlexNet needed six days to be trained using two Nvidia Geforce GTX 580 GPUs which is the reason for why their network is split into two pipelines (Siddharth ,2017).

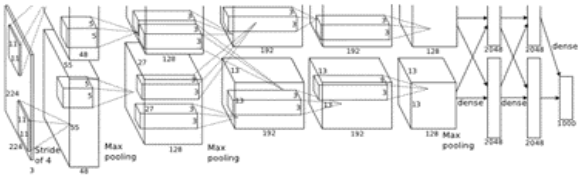


Figure 2: An illustration of the architecture of AlexNet (Alex et al,2012)

2. LITURETURE REVIEW

Video indexing ended up one of the foremost critical domains in video applications since 1990, specifically change of videos into a format appropriate for retrieval frameworks. In such scenery, the approach is repeatedly as follows: to begin with videos are fragmented into shots, at that point keyframes are extracted from each shot. At this point, the issue of looking through videos can be performed through picture recovery, i.e. by recovering keyframes comparable to a picture submitted as inquiry (Zhang et al.,1997)(Chang et al.,1999)(Pickering and Ru¨ger,2003)(Sze et al.,2005). Alternately nowadays the same applications are substantial but utilizing speech processing (Dey et al. ,2018), signal processing (Sreenu et al.,2019), deep learning (Padalkar, 2010) and neural network. Figure 3 illuminates the classified construction of video. (Karpathy et al.,2014) run a wide empirical evaluation of CNNs on large-scale video classification using a new dataset of one million YouTube videos belonging to 487 classes. They retain the top layers on the UCF-101 Action Recognition dataset and observe significant performance improvements compared to the UCF-101 baseline model (63.3% up from 43.9%). (Sochor et al.,2016) used images from video frames and CNN to boost the recognition performance using 3D vehicle bounding box. Some researchers created car accident dataset from scratch and develop custom software to process on it (Mustaffa and HOKAO, 2013). In their investigation (Xinchen et al., 2017) classified the sorts of vehicles utilizing Faster R-CNN (Region-Convolutional Neural Networks). They demonstrated that utilizing deep learning is more strength and higher precision than conventional machine learning strategies. Thus, (Sun et al.,2019) improved Faster R-CNN algorithm by adding Regional Proposal Network (RPN) to the convolution layer 5 in Faster R-CNN. Recently, academics are attentive on pre-trained and fine-grained CNN in their research. (Yang et al.,2015) distributed a huge dataset for fine-grained vehicle sort recognition they proposed a CNN-based strategy for fine-grained vehicle model recognition which accomplished of recognizing vehicles from diverse perspectives. Extra fine-grained vehicle type recognition technique was done by (Fang et al., 2017) based on a combination of local and global features extracted from the CNN model which is afterwards named pre-trained AlexNet. More details about Alexnet will be discussed in section 3.3.2.

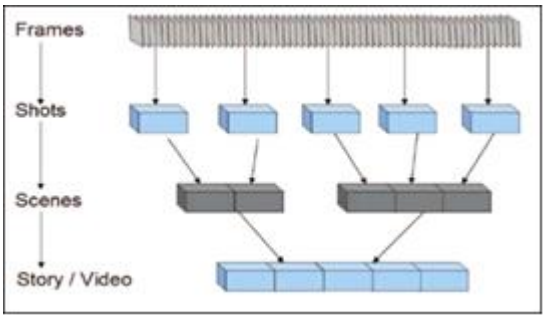


Figure 3: Hierarchical Structure of Video (Milind,2010)

3. THE PROPOSED METHOD

This research consists of three stages which are building a raw dataset, building new dataset and complexity comparison. Figure 4 demonstrates the main methodology of this research. Each stage will be detailed in separate sections.

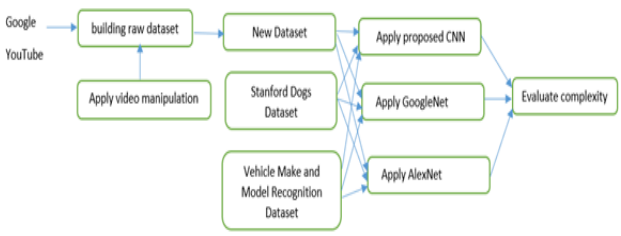


Figure 4: The main methodology of the research

3.1 Building new dataset

The first stage toward executing this research is to build a dataset. This considered as a changeling issue. YouTube and Google were the main sources of these videos. Some sources were from some countries official sites that publish vehicle crash accident in their country and no copyright need it for using them. The dataset that we built contained 10 videos. These videos are recorded at different traffic light positions in different countries. Also, two of the videos were taken from animation website.

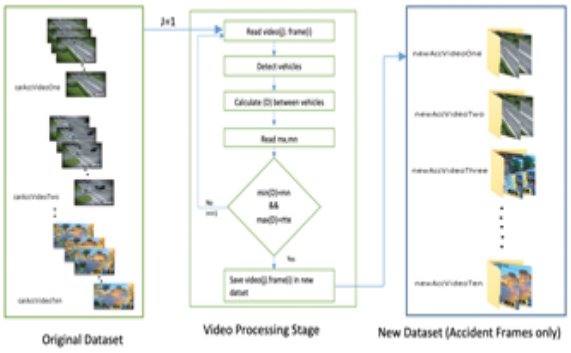


Figure 5: Methodology of creating a new dataset

3.2 Optimized dataset

By using video manipulation process only accident frames will be saved to create a new dataset. The process is applied for all videos in the raw dataset. This means the new dataset will contain 10 new folders with fewer frames number. The number of frames in each folder is different than the other because the time length of each video is different and the accident frames in some videos are more than the other. In some videos, the accident starts in the second 50 and last to minute 2. While other videos start in minute 1 and last until minute 7. The number of frames compromised depending on specific parameters. Table 1 shows the number of frames before and after stage one. The end of this process is a new data set that contains ten folders.

3.2.1 Detecting accident Frames

This section presents the main techniques for accident frames detection in videos. Each video consists of sequence of frames f, f= {f1, f2, f3, …., fn}. The function G, computer Gaussian mixture model to calculate centroids (C). Distance (D) measures the difference between frames fi and fi+1 in the sequence at each time (t). Calculating maximum and minimum distance is an adjustment calculation between the frames. When the vehicles are very far from each other than max(distance) will be saved. The opposed is true when the vehicles are very close. While applying this algorithm on different videos we discovered that the main problem of this approach is calculating min and max. However, even if they were not calculated properly, it will increase 5-8 frames which do not increase the number of frames in the new dataset. The algorithm is written below. For vehicles detection Gaussian mixture model is used. Figure 5 shows the methodology used in the proposed work. The algorithm explains how frames inside the video analyzed to be processed for recognizing the accident. The following is a segment of pseudo-code for the proposed algorithm used in this research.

Algorithm one: Accident frames detection

Procedure Video handling (vcarAccVideoTwo, C,d)

Repeat for video file

```
for each f in carAccVideoTwo  do
  if(size(C,1)>1) then
      if(size(C,1)==2) then
          call find the distance(d)
       elseif(size(C,1)==3) then
          call find the distance(d)
       elseif(size(C,1)==4)  then
          call find the distance(d)
          elseif(size(C,1)==5)  then
          call find the distance(d)
      end if
  mx <-- max(d)
```

mn <-- min(d)

end for

until no frame in the video

end procedure

Figure 6: The original frame number 49 from carAccVideoOne (b) binary image of frame number 49 (c) detecting the car object from carAccVideoTwo (d) foreground segmentation of image frame number 254. (e) consider the vehicles as accidents cars for reporting. (f) label the vehicles that are close to making accidents.

3.3 Network Configuration

The last stage is the process of CNN configuration. It involves 3 phases; design, train and checking the accuracy. Figure 7 shows the steps needed for this stage. Network configuration.

3.3.1 Features Extraction and classification

Extracting salient features is the major critical phase in most research in object recognition and computer vision tasks. Hence, some works have concentrated on extracting strong features for a variety of image classification tasks [1]. Bag-of-words model is a common method used for image classification due it gave a promise results for a huge of research works. However, BoW is containing some drawbacks like leads a high dimensional feature vector and in-order visual words for the represented images; this leads to giving similarity of images with different content (Liu et al.,2015).

Table (1) Number of frames in each video before and after applying a video optimization step.

Folder Name	Number of frame (original videos)	Duration	Number of frames after manipulating videos
carAccVideoOne	238	00:03:07	150
carAccVideoTwo	240	00:04:08	198
carAccVideoThree	302	00:03:30	169
carAccVideoFour	304	00:03:50	115
carAccVideoFive	345	00:04:15	146
carAccVideoSix	302	00:04:10	175
carAccVideoSeven	398	00:05:14	111
arAccVideoEight	398	00:05:14	130
carAccVideoNine	238	00:03:20	180
carAccVideoTen	402	00:10:01	200

Figure 7: Network outline phases

In CNN, convolution layers, max-pooling layers and activation function are responsible for extracting features while the fully connected layer is used for classification. It collects the final convoluted feature and returns a column vector where each row focuses towards a class. More accurately, components of the output vector each represent the likelihood estimation of each class and the sum of the components is one as shown in figure 8. The extracted features of each video have been used for SVM (Support Vector Machine) classifier.

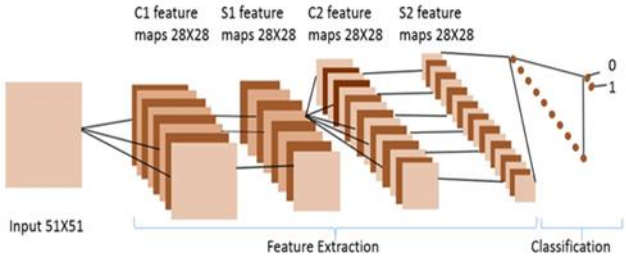


Figure 8: CNN layers for feature extraction and classification [28]

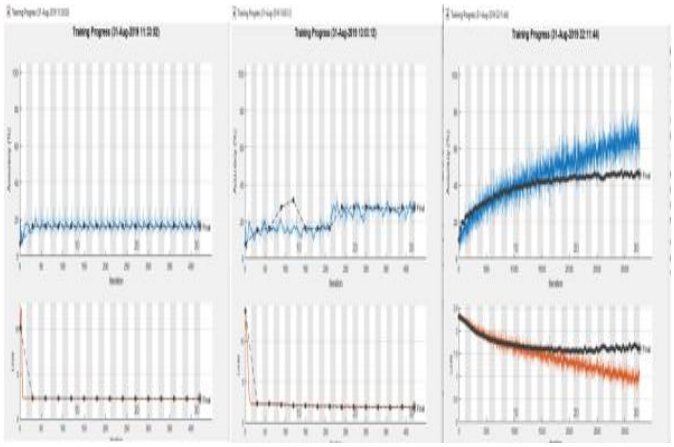
3.3.2 The proposed architecture of CNN

The proposed construction of the CNN contains input layer, numerous convolutional layers (C), max-pooling layers (M), and fully connected layers (F). Pooling layer is mostly used immediately after the convolutional layer to decrease the spatial size. As a result, the number of parameters and computation will be decreased to avoid overfitting. The experimental work has been started with twelve basic layers of architecture, then the number of layers has been tuned increasingly to match the planed accuracy. Table 2 illustrates one of the configurations of structured CNN. However, the results were not as expected, and the accuracy was 0.1619 which is 16.19%. After adding additional layers, the accuracy increased and became 27.26% then 49% as shown in figure 9. More experiments have been done by adding more layers, led to increasing the accuracy to be more than 98%. The final architecture of the proposed CNN demonstrates in table 3.

Table (2) One of the intended CNN structure in our research

Layer No.	Layer Type
1	Input layer
2	Convolutional
3	Relu
4	Max pooling
5	Cross Chanel
6	Convolutional
7	Relu
8	Max pooling
9	Fully connected
10	Fully connected
11	Soft max
12	Classification

In each design, Different parameters have been applied for the entire layers regularly until match the accuracy needed. Figure ten illustrates the required accuracy that we were seeking.



(a) (b) (c)

Figure 9: Validation accuracy, training and loss cycles. (a) for proposed CNN in table 3. (b),(c) after adding more layers.

Table (3) The final version of the CNN structure in our research

Layer No.	Layer Type
1	Input layer
2	Convolutional
3	Relu
4	Cross Chanel
5	Max pooling
6	Convolutional
7	Relu
8	Cross Chanel
9	Convolutional
10	Relu
11	Convolutional
12	Relu
13	Max pooling
14	Fully connected
15	Relu
16	Drop out
17	Fully connected
18	Relu
19	Drop out
20	Fully connected
21	Soft max
22	Classification

4. RESULTS AND DISCUSSION

In order to evaluate the proposed CNN, raw, Stanford Dogs and Vehicle Make and Model Recognition Datasets have been used. Ten videos from Google and YouTube have been collected for constructing a raw dataset. In order to apply 10 folders that contain the video's frames in the experimental work. After detecting accident

vehicles in the frames, the proposed technique will save the accident frames in a new folder to become a new dataset, the frames will be selected based on Gaussien mixture model. The proposed work decreased the number of frames from 2,769 to 1,300. This, in turn, reducing the consumed time for further analyzing the videos for E-government purpose due to fewer frames within each folder

The first dataset has been collected from open sources website. While the duration and the size of the videos were different, the optimization technique using Gaussian mixture model has been used to detect the vehicle accidents frames and save it in another folder. Table 4 shows the dataset used in this research and the number of frames in each dataset. The total number of images used in this research was 10765 with .jpg extension but variance dimensions in each dataset. For the input layer, we started training the proposed CNN with the original dimensions of the frames which were 219X415X3 coloured images. While training, we discover that it took long times and the structure needed extra layers for feature computation. We also test 38X38X3 dimensions and check the differences in computation cost. Because the datasets will be tested in AlexNet and GoogleNet, the input images should be close to the input layer dimension images of these networks which are 227X227X3 and 224X224X3 respectively. Consequently, we changed the dimensions of all frames patch to be 227X227X3. Figure 11 shows a random image of each folder in a raw dataset with 227x227x3 dimension. Recording feature extraction stage, we enhance the features produced by CNN by feeding them to SVM.

Table (4) Dataset used for testing the proposed CNN

Data set	No. of images
Optimized	1564
Stanford Dogs	4560
Vehicle Make and Model Recognition	4300

We have multiple fully connected layers in our recommended structure. This is because it learns higher-level features and the result assistance the final layer especially when a sigmoid function is used as an activation function. Table 5 indicates the accuracy outcomes of the CNNs applied to the datasets. Our proposed model involves 22 layers. It reduces the complexity and number of parameters to be used in a laptop with 8GB RAM. Figure 12 shows the probability accuracy when applying our suggested CNN model on different datasets.

Table (5) The recognition accuracy when adopting different datasets

Dataset / Accuracy	The proposed CNN	GoogleNet	AlexNet
optimized dataset	85%	82%	83%
Stanford Dogs dataset	85%	79%	81%
Vehicle Make and Model Recognition Dataset (VMMRdb)	86%	80%	83%

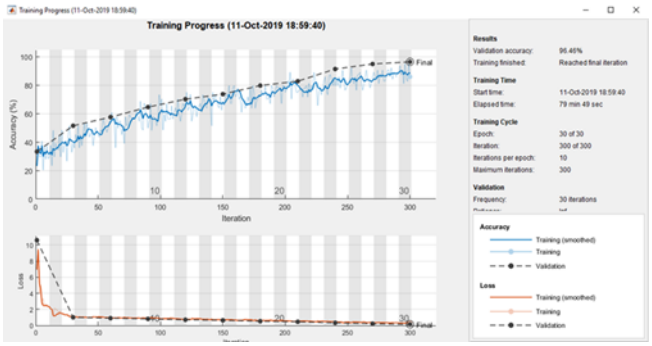


Figure 10: Training process for CNN structure in the table 4.

Figure 11: Images of car accident videos with 227X227X3 dimension**.**

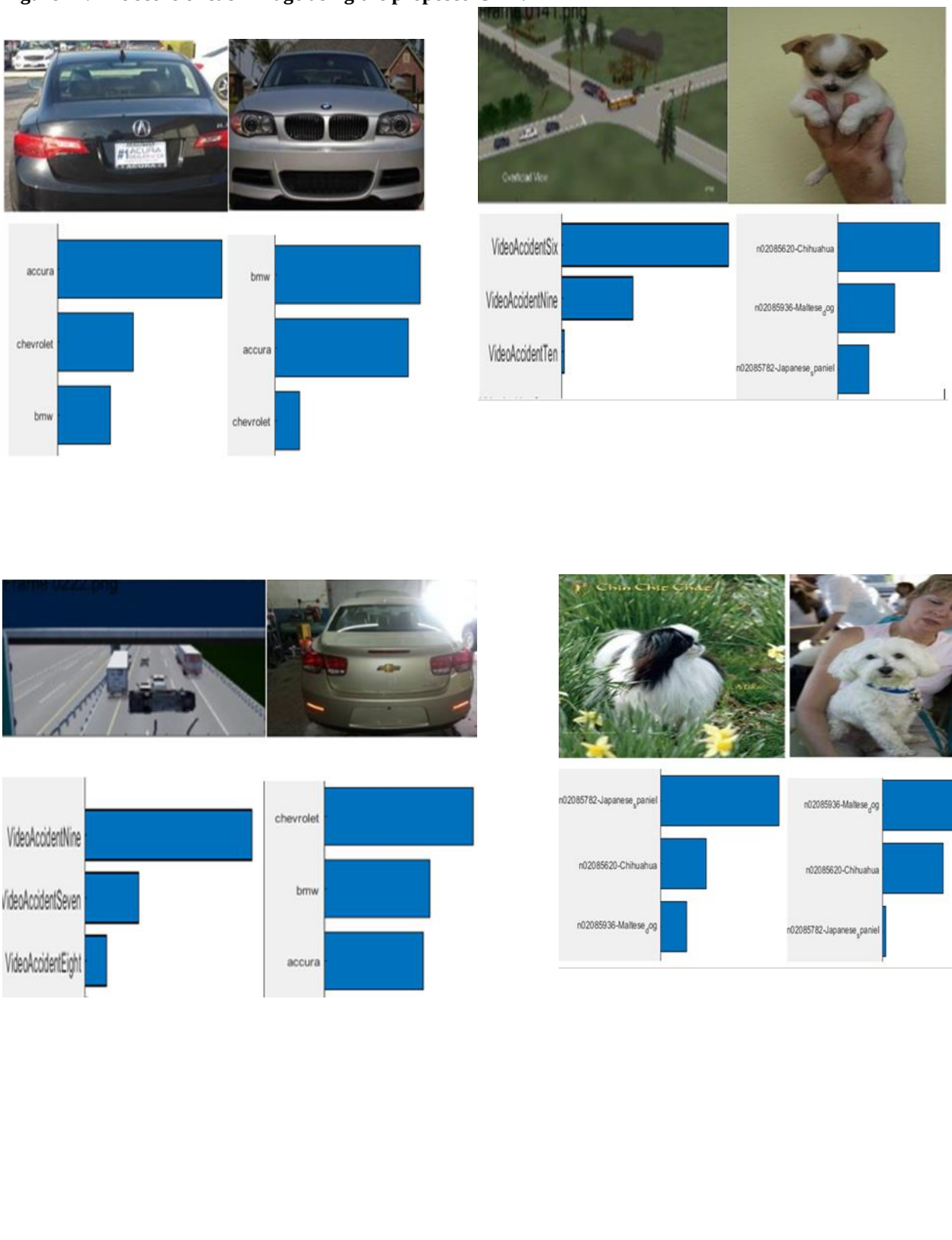
5. CONCLUSIONS AND FUTURE WORK

In this research, an optimized technique for accidents vehicles dataset images reduction has been used and a CNN is designed from scratch to train and do the classification process for a self-generation report. Optimizing the dataset in video manipulation increases the accuracy. Since reducing frames rate per each video focusing only

on accidental frames, leads to processing fewer features during the reduced video duration. This dataset along with the other two datasets were trained using the proposed CNN. The architectures and parameters for the GoogleNet and AlexNet have been modified to build the proposed net.

In future work, we hope to corporate with traffic police to obtain a high number of accident vehicle videos to create a dataset with 100 videos and generate detailed report about the accidents. Besides, using cloud computing to easily manipulate large datasets to be easily ingested and managed to train algorithms.

Figure 12. The score of each image using the proposed CNN.



Acknowledgements

I would like to thank the supervisors that help to enrich the research.

References

Anwer M., Shareef M, Ali A. (2019)' Smart Traffic Incident Reporting System in e-Government' ECIAIR conference, UK DOI: 10.34190/ECIAIR.19.061

Chang H.S., S. Sull, and S.U. Lee (1999) 'Efficient video indexing scheme for content- based retrieval'. IEEE Transactions on Circuits and Systems for Video Technol- ogy, 98:1269–1279.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich (2015) ' Going deeper with convolutions'. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1–9.

Clady, Xavier & Negri, Pablo & Milgram, Maurice & Poulenard, Raphael (2008)' Multi-class Vehicle Type Recognition System'. 5064. 228-239. 10.1007/978-3-540-69939-2_22.

Dey N., Ashour A.S (2018) 'Applied Examples and Applications of Localization and Tracking Problem of Multiple Speech Sources'. In: Direction of Arrival Estimation and Localization of Multi-Speech Sources. Springer Briefs in Electrical and Computer Engineering. Springer, Cham.

Fang J., Y. Zhou, Y. Yu, and S. Du(2017) 'Fine-grained vehicle model recognition using a coarse-to-fine convolutional neural network architecture' IEEE Trans. Intell. Transp. Syst., vol. 18, no. 7, pp. 1782–1792, Jul.

Hasanpour, S., Rouhani M., Fayyaz, M., Sabokrou, M., (2018). 'Let's keep it simple,using simple architectures to out perform deeper and more complex architectures'. arXiv:1608.06037v7.

Huang YL (2018) 'Video Signal Processing'. In: Dolecek G. eds Advances in Multirate Systems. Springer, Cham.

Jakob Verbeek, Nikos Vlassis, Ben Krose (2003) 'Efficient greedy learning of Gaussian mixture models '.Neural Computation, Massachusetts Institute of Technology Press MIT Press, 2003, 15 2, pp.469-485. 10.1162/089976603762553004.

Karpathy A., G. Toderici, S. Shetty (2014) 'Large-scale Video Classification with Convolutional Neural Networks.'

Kolekar M., (2018)' Intelligent video surveillance system: An algorithm Approach'. CRC Press, Tayor & Francis Group.

Krizhevsky, Alex , Sutskever, Ilya , Hinton, Geoffrey (2012) 'ImageNet Classification with Deep Convolutional Neural Networks'. Neural Information Processing Systems. 25. 10.1145/3065386.

Liu H., Tang H.,Xiao W.,Guo Z., Tian L., Gao Y. (2016) 'Sequential Bag-of-Words model for human action classification'. CAAI Transactions on Intelligence Technology, Volume 1, Issue 2, Pages 125-136

Minaee S., Abdolrashidi A., and Y. Wang (2015) 'Iris recognition using scattering transform and textural features', in Signal Processing and Signal Processing Education Workshop SP/SPE, 2015 IEEE, 2015, pp. 37-42.

Mustaffa A., K. HOKAO (2013) 'Database development of road traffic accident case study Johor Bahru, Malaysia' Journal of Society for Transportation and Traffic Studies JSTS Vol.3 No.1.

Navneet Dalal, Bill Triggs (2005) 'Histograms of Oriented Gradients for Human Detection. International Conference on Computer Vision & Pattern Recognition' (CVPR '05), San Diego, United States. pp.886—893.

Niebles J. C., C.-W. Chen, and L. Fei-Fei (2010) 'Modeling temporal structure of decomposable motion segments for activity classification'. In ECCV, pages 392–405. Springer.

Padalkar, Milind (2010) 'Histogram Based Efficient Video Shot Detection Algorithms'. 10.13140/RG.2.1.1590.3847.

Pickering M.J. and S. Ru¨ger (2003) 'Evaluation of key-frame based retrieval techniques for video'. Computer Vision and Image Understanding, 922-3:217–235.

Ranganatha S., Gowramma, Y. (2018) 'Image Training and LBPH Based Algorithm for Face Tracking in Different Background Video Sequence'. International Journal of Computer Sciences and Engineering. 6. 349-354. 10.26438/ijcse/v6i9.349354.

Salih. D., Ali .A (2019)' Appearance-based indoor place recognition for localization of the visually impaired person' ZJPAS (2019) , 31(4);70-81. DOI: http://dx.doi.org/10.21271/ZJPAS.31.4.8

Siddharth Das., (2017) 'CNN Architectures: LeNet, AlexNet, VGG, GoogLeNet, ResNet and more'. online at [https://medium.com/@sidereal/cnns-architectures-lenet-alexnet-vgg-googlenet-resnet-and-more-666091488df5]

Sochor J., A. Herout, and J. Havel (2016) 'Box Cars: 3D boxes as CNN input for improved fine-grained vehicle recognition' in Proc. Comput. Vis. Pattern Recognit., Jun., pp. 3006–3015.

Sreenu, G., Saleem Durai, M.A (2019) 'Intelligent video surveillance: a review through deep learning techniques for crowd analysis'. J Big Data 6, 48 doi:10.1186/s40537-019-0212-5.

Sun X., Gu J., Huang R.,Zou R.,Giron B. (2019) 'Surface Defects Recognition of Wheel Hub Based on Improved Faster R-CNN'. Electronics 2019, 8, 481; doi:10.3390/electronics8050481.

Sze K.W., K.M. Lam, and G. Qiu. (2005) 'A new key frame representation for video segment retrieval'. IEEE Transactions on Circuits and Systems for Video Technology, 159:1148–1155.

Xinchen Wang X., Zhang W.,Wu X., Xiao L., Qian Y.,Fang Z. (2017) 'Real-time vehicle type classification with deep convolutional neural networks'. J Real-Time Image Proc DOI 10.1007/s11554-017-0712-5, Springer.

Yang L., P. Luo, C. C. Loy, and X. Tang (2015) 'A large-scale car dataset for fine-grained categorization and verification'. in Proc. Comput. Vis. Pattern Recognit., Jun., pp. 3973–3981.

Zhang H.J., J. Wu, D. Zhong, and S.W. Smoliar (1997) 'An integrated system for content-based video retrieval and browsing'. Pattern Recognition, 304:643–658.