PriMera Scientific Engineering
Volume 2 Issue 4 April 2023
ISSN: 2834-2550

Hazardous Behavior Recognition Based on Multi-Model Fusion

Type: Research Article
Received: March 13, 2023
Published: March 24, 2023

Citation:
Bingyi Zhang., et al. "Hazardous Behavior Recognition Based on Multi-Model Fusion". PriMera Scientific Engineering 2.4 (2023): 17-22.

Bingyi Zhang*, Bincheng Li and Yuhan Zhu

Dalian Maritime University, Dalian, China

***Corresponding Author:** Bingyi Zhang, Dalian Maritime University, Dalian, China.

Abstract

Distracted driving has become a serious traffic problem. This study proposes an image process- ing and multi-model fusion scheme to maximize the discrimination accuracy of distracted driving. First, the training dataset and the test dataset are processed to specific specifications by translation and clipping. Second, we set vgg16 as the benchmark model for evaluation, and train ResNet50, InceptionV3 and Xception model input images. Finally, considering that each model has its own advantages, we use frozen part of the network layer to fine-tune the model, remove the weights of each single model Fine-tune from the output before full connection, connect them in series, and then calculate each model weights through neural network training.

Introduction

There are many traffic accidents caused by distracted driving today. More than 3,700 people around the world die each day in road traffic collisions [1]. Many, if not all of these deaths could have been avoided. These sorts of crashes are the leading killer of kids and young people aged one to 25 in the US and five to 29 around the world. We need to create a safe systems approach so we can protect people and save lives. According to NHTSA [2], distracted driving behaviors mainly include:

Talking or Texting on one's phone, b) eating and drinking, c) talking to passengers, d) fiddling with the stereo, entertainment, or navigation system. We believe that detecting distracted driver gestures is the key to further preventive measures. Detection of driver distraction is also important for autonomous vehicles; We propose a hybrid model distracted driving attitude evaluation system, which mixes Xception, InceptionV3, and ResNet50, and adopts a more appropriate number of convolutional layers. Our main purpose is to improve the accuracy of the classifier.

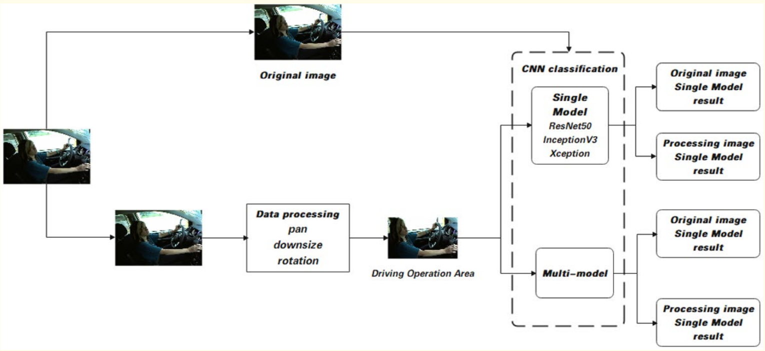

Original image
CNN classification
Single Model
ResNet50
InceptionV3
Xception
Original image Single Model result
Processing image Single Model result
Data processing
pan
downsize
rotation
Driving Operation Area
Multi-model
Original image Single Model result
Processing image Single Model result

Figure 1: Overall Idea Flow Chart.

Ease of Use

Maintaining the Integrity of the Specifications

At first, distracted driving behavior was an evaluation of multiple issues such as drivers making phone calls and not wearing seat belts. The Southeast University (SUE-DP) dataset [3] was proposed in 2011. It mainly includes holding the steering wheel, operating the gear lever, eating, Calling four categories of distracted driving behaviors, Berri proposed a polynomial kernel support vector machine (SVM) classification system for this dataset, and the classification success rate was 91.57%. Yan [4] proposed a random forest algorithm to extract PHOG features based on time series, with an accuracy of 96.56%.

With the refinement of the model and the development of machine learning, the impact of hand and face analysis on distraction detection has been discussed separately for different positions. The model uses the kaggle public data set to train the weighted model, and the classification accuracy reaches 95.98%.

Image enhancement

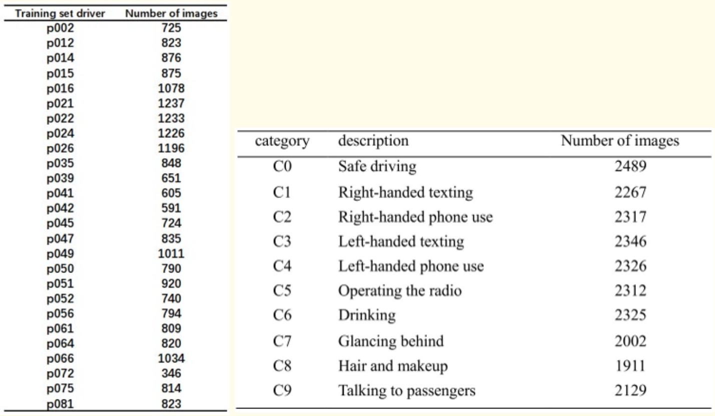
The closest work to our image enhancement part is Zhou, Bolei, et al [5], The author's purpose is to locate the core position of the image given the label, in order to exclude the interference of the external environment.

Dataset

Since most of the datasets from Southeast University and related distracted driving datasets are not public, we use the competition dataset provided by kaggle. The data features are as follows:

1. There are a total of 102,150 pictures, of which 79,726 are in the test set, and the number of the test set is much larger than that of the training set;
2. The training set is different from the drivers collected in the test set. The training set is collected from 26 drivers, and the test set is selected from 55 drivers;
3. The picture has obvious temporal continuity;
4. The data is divided into 9 tags, and the number of images per tag is relatively average;
5. The camera placement is slightly different, and pre-processing is required to reduce the image range to prevent overfitting.

The approximate distribution of the images is shown in the following table:



Training set driver	Number of images
p002	725
p012	823
p014	876
p015	875
p016	1078
p021	1237
p022	1233
p024	1226
p026	1196
p035	848
p039	651
p041	605
p042	591
p045	724
p047	835
p049	1011
p050	790
p051	920
p052	740
p056	794
p061	809
p064	820
p066	1034
p072	346
p075	814
p081	823

category	description	Number of images
C0	Safe driving	2489
C1	Right-handed texting	2267
C2	Right-handed phone use	2317
C3	Left-handed texting	2346
C4	Left-handed phone use	2326
C5	Operating the radio	2312
C6	Drinking	2325
C7	Glancing behind	2002
C8	Hair and makeup	1911
C9	Talking to passengers	2129

(a) Pictures of Training Set. (b) Pictures of Different Categories.

Proposed Method

Our proposed solution consists of image augmentation and fusion models. We choose to use Fine- tune, that is, on the basis of transfer learning, training without locking part of the weights of Xception [6], InceptionV3 [7] and Resnet50 [8], the weights of ImageNet are partially locked at the level, and some levels are retrained. I choose to use multiple models, because the design ideas of different models are different, and the principles of extracting image features are also different. In theory, the use of multi-model fusion can improve the accuracy, and the fusion method I use is not to find an average, but to The weights of the single model Fine-tune remove the output before the full connection, connect them in series, and then solve the weights of each model through neural network training.

Image Processing

Image zooming and panning

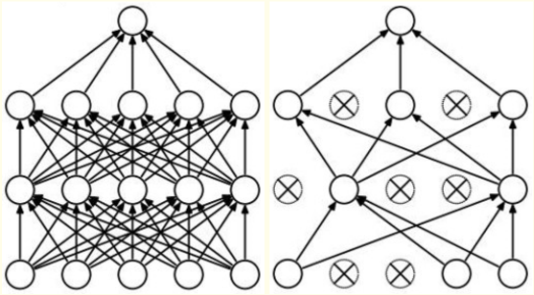
Because the collection of the dataset mainly uses the camera in the car to take the media files of the driver's driving process [9]. The collection process collects the driver's face and hands, and the resolution, installation location, camera angle or the driver's habitual driving posture of the cameras in different datasets brings redundant obstacles to the classification system.

Due to the difference in camera placement and the small amount of data, in order to prevent overfitting later, the data is enhanced before the training data. We translate, zoom, rotate, etc. for each image, increasing the diversity of the data.

Parametric Optimization

Use Dropout Layers

There are roughly two reasons why Dropout can prevent overfitting [10]: averaging, and reducing the complex co-adaptation relationship between neurons (reducing weights makes the network more robust to losing specific neuron connections). We use 0.5 probability dropout before the final Dense layer.



(c) Standard Neural Net. (d) After applying dropout.

Classification Methods

- ***Resnet50*** [8]
  Solve the problem that when the network depth increases to a certain extent, the stacking effect of the deeper network becomes worse. Resnet50 can increase accuracy without increasing complexity.
- ***InceptionV3*** [7]
  Based on the original Inception, the number of parameters is significantly reduced by sharing weights between adjacent blocks. This design can reduce the time complexity.
- ***Xception*** [6]
  Xception uses depth wise separable convolution to replace the multi-size convolution kernel feature response operation in the original InceptionV3, reducing parameters and increasing accuracy.

Our Classification Model

I choose to use multiple models, because the design ideas of different models are different, and the principles of extracting image features are also different. In theory, the use of multi-model fusion can improve the accuracy, and the fusion method I use is not to find an average, but to The weights of single-model Fine-tune remove the output before full connection, connect them in series, and then solve the weights of each model through neural network training.

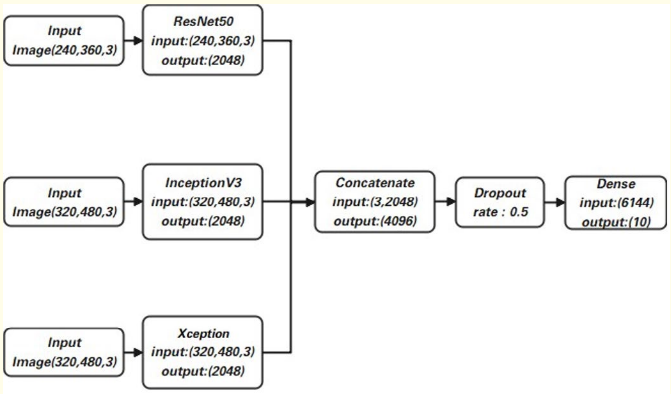

Input Image(240,360,3)
ResNet50 input:(240,360,3) output:(2048)
Input Image(320,480,3)
InceptionV3 input:(320,480,3) output:(2048)
Input Image(320,480,3)
Xception input:(320,480,3) output:(2048)
Concatenate input:(3,2048) output:(4096)
Dropout rate : 0.5
Dense input:(6144) output:(10)

Figure 2: Multi-model CAM Activation Image.

Experiments

Single Model Optimization

Benchmark Model

For final optimization and comparison, I use a unified fully connected layer. The fully connected layer I use is as follows.

When the input model of ResNet, the graph is scaled to (224, 224, 3), and the output after removing the fully connected layer is a vector of length 2048.

Various Experimental Models

- ***ResNet50***
  The image is reduced to (240, 320, 3), probably because many scenes have very small actions, and the reduction is too small to propose features. And after repeated attempts, the result of keeping the image ratio unchanged and entering the neural network is that it is best to first train with Adam for 6 rounds, and then using RMSprop to train for 6 rounds with a very small learning rate of 0.00001. After many experiments, it is found that the best effect of tune is to start the tune at the 152nd layer, that is, to lock the weights of the 0-151 layers. From the 152nd layer, the weights can be trained.
- ***InceptionV3***
  The image is reduced to (360, 480, 3), use Adam to train for 4 epochs, and then using RMSprop to train for 6 epochs with a minimal learning rate of 0.00001.From the 172nd layer, the weights can be trained.
- ***Xception***
  The image is reduced to (360, 480, 3) and using Adam to train for 4 epochs, and then using RMSprop to train 6 epochs with a very small learning rate of 0.00001. From layer 172, the weights can be trained.

Comparison between different models

In order to verify the superiority of the hybrid model and the effectiveness of the network, we conduct experiments on the kaggle public dataset. Mixing model from accuracy also has advantages. And the accuracy of most models is higher than 90%, and the model results have good recognition accuracy.

Model	***Accuracy (%)***
$BaseModelV tttt16$	64.49
$ResNet50$	85.96
$InceptionV3$	91.76
$Xecption$	90.36
$ResNet50 + InceptionV3 + Xecption$	93.45

Conclusion

Distracted driving has become an important culprit in traffic tragedies. In order to identify distracted driving behaviors and solve problems. First, image enhancement is used to extract driving behavior-related regions in the image. Second, we propose a new classification model (hybrid model), which achieves the classification accuracy of distracted driving behavior with large and small accuracy. This system is formed by the combination of three models. The weight of each model is calculated through the neural network, and the connection layers of the model are connected to each other, which strengthens the confidence of the model. Finally, through the convolutional network classification system, the original data set is classified, and the model accuracy is obtained. Experiments on the kaggle dataset show that the accuracy of the mixture model after image processing is far better than that of the other existing models. Therefore, this method has better classification efficiency.

References

1. World Health Organization. Global Status Report on Road Safety 2018: Summary; World Health Organization: Geneva, Switzerland (2018).
2. Timothy M Pickrell, Hongying (Ruby) Li and Shova KC. Traffic Safety Facts (2016).
3. Berri Rafael A., et al. "A pattern recognition system for detecting use of mobile phones while driving". 2014 International Conference on Computer Vision Theory and Applications (VISAPP). IEEE 2 (2014).
4. Yan C, Coenen F and Zhang B. "Driving posture recognition by joint application of motion history image and pyramid histogram of oriented gradients". International journal of vehicular technology (2014).
5. Abouelnaga Yehya, Hesham M Eraqi and Mohamed N Moustafa. "Real-time distracted driver posture classification". arXiv preprint arXiv:1706.09498 (2017)
6. Chollet F. "Xception: Deep Learning with Depthwise Separable Convolutions". 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2017).
7. Szegedy C., et al. "Rethinking the Inception Architecture for Computer Vision". 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2016): 2818-2826.
8. Ou C, Ouali C and Karray F. "Transfer learning based strategy for improving driver distraction recognition". International Conference Image Analysis and Recognition. Springer, Cham (2018): 443-452.
9. Wang J., et al. "A data augmentation approach to distracted driving detection". Future internet 13.1 (2021): 1.
10. Saito K., et al. "Adversarial dropout regularization". arXiv preprint arXiv:1711.01575 (2017).
11. Zhang Z. "Improved adam optimizer for deep neural networks". 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE (2018): 1-2.
12. Wichrowska O., et al. "Learned optimizers that scale and generalize". International Conference on Machine Learning. PMLR (2017): 3751-3760.