Data Augmentation via Neural-Style-Transfer for Driver Distraction Recognition

Che-Tsung Lin[1*], Thomas Streubel [2,3], Marco Dozza [2], Giulio Bianchi Piccinini [2], Christopher Zach[1]

[1] Department of Electrical Engineering, Chalmers University of Technology, SE-41296, Gothenburg, Sweden

[2] Department of Mechanics and Maritime Sciences, Chalmers University of Technology, SE-41296, Gothenburg, Sweden

[3] Current affiliation: Volvo Car Corporation AB, Safety Center, SE-40531, Gothenburg, Sweden

*chetsung@chalmers.se;alexofntu@gmail.com

Outline

- **Motivation**
- **Introduction**
- **Do CNNs learn from texture or shape?**
- **Methodology**
- **Experimental Results**
- **Conclusion and Future work**

Motivation

- According to the National Highway Traffic Safety Administration, 3142 people were killed in motor vehicle crashes involving distracted drivers in 2019.
- Research on distraction has largely benefit from the analysis of Naturalistic Driving Data (NDD) in the last 15 years.
- One of the main concerns associated to the use of NDD is the time intensive and costly process of video reduction to extract variables from the videos.
- Convolutional Neural Networks (CNNs) are commonly thought to recognize objects by learning increasingly complex representations of object shapes. However, some recent studies suggested a more important role of image textures instead.
- Overfitting and over-confidence are two major issues that easily arise when training CNNs.

National Highway Traffic Safety Administration (2022). Distracted Driving. Retrieved March 29, 2022.

Modeling driver monitoring as image classification

- Pros
  - 10 classes.
  - Most images were collected in real-driving scenario.
  - Collected in different cars, by people from different countries.
- Cons:
  - Captured with only two viewing angles

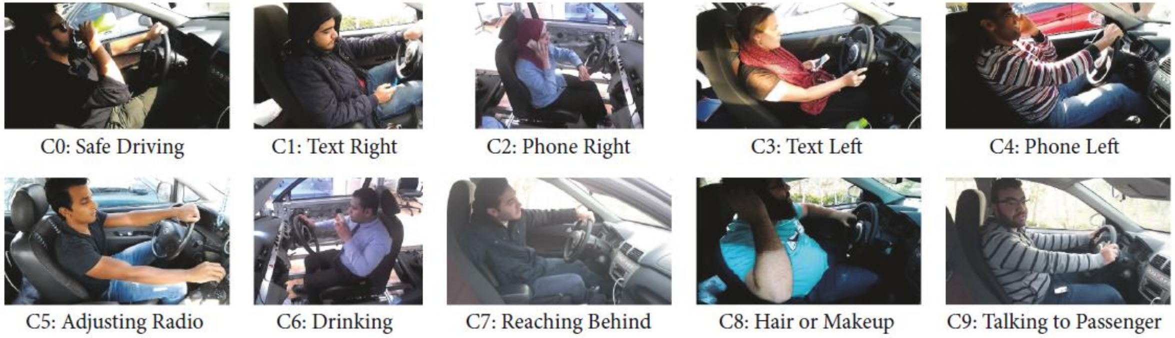
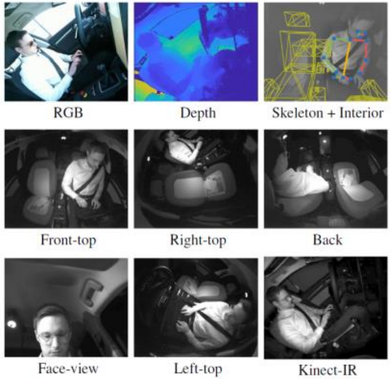


FIGURE 1: Ten classes of driver postures from our dataset.

Eraqi, et al. "Driver distraction identification with an ensemble of convolutional neural networks." *Journal of Advanced Transportation* , 2019.

Drive & Act dataset

- Different views lead to different results and fully-seen view could provide better results.
- The more the driver’s body captured by the camera; the higher the recognition accuracy.
- Model ensemble is beneficial.
- Pros:
  - The largest dataset of driver’s action recognition
  - 6 different views
  - 3 modalities(RGB, IR and depth)
  - 83 classes
- Cons:
  - The same vehicle (Audi A3) for every driver in a simulated environment.
  - The number of drivers is limited. (10 for training, 2 for validation, 3 for testing)



Camera	View	Validation	Test
NIR Cameras	front top	69.57	63.64
	right top	65.16	60.80
	back	54.70	54.34
	face view	49.73	42.98
	left top	68.72	62.83
	combined	72.70	67.17
Kinect Color	right top	69.50	62.95
Kinect Depth	right top	69.43	60.52
Kinect IR	right top	72.90	64.98
Combined	right top	73.80	68.51
All combined (score averaging)		**74.85**	**69.03**

Martin et al. "Drive & Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles." *ICCV* , 2019.

EuroFOT and Drive C2X

- Pros:
  - Provides significant more images with different views
  - 8 classes of distractions and 5 phone usages and four hands-on-wheel classes.
  - More drivers (127)
- Cons:
  - Monochrome images.
  - Limited resolution (352x288)

Class and images	**Training**	**Validation**
No activities	169213	200432
Interaction with passenger	1244	504
Talking or singing	18106	732
Reaching for an object	19232	8046
Interaction with center stack	6836	3439
Eating/Drinking	4975	2784
Hands-face interaction	29847	19598
Reading	1920	1058

https://www.eurofot-ip.eu/
https://cordis.europa.eu/project/id/270410

CNN learns from texture more!

- This paper has shown that ImageNet-trained CNNs are strongly biased towards recognizing textures rather than shapes.

(a) Texture image		(b) Content image		(c) Texture-shape cue conflict	
81.4%	**Indian elephant**	71.1%	**tabby cat**	63.9%	**Indian elephant**
10.3%	indri	17.3%	grey fox	26.4%	indri
8.2%	black swan	3.3%	Siamese cat	9.6%	black swan

Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.

Geirhos et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." *ICLR* , 2020.

CNN’s performance degrades when lacking of texture information

- Pretraining with stylized-Imagenet is beneficial in terms of image classification and object detection since a shape-based representation is more beneficial than a texture-based representation.
- However, due to copyright issue, stylized-Imagenet can’t be directly provided.
- The neural-style transfer model needs 1 min (500 iterations) for each mage.
- For 1.2 million images, we need 833 days.

Stylized-ImageNet

name	training	fine-tuning	top-1 IN accuracy (%)	top-5 IN accuracy (%)	Pascal VOC mAP50 (%)
vanilla ResNet	IN	-	76.13	92.86	70.7
	SIN	-	60.18	82.62	70.6
	SIN+IN	-	74.59	92.14	74.0
Shape-ResNet	SIN+IN	IN	**76.72**	**93.28**	**75.1**

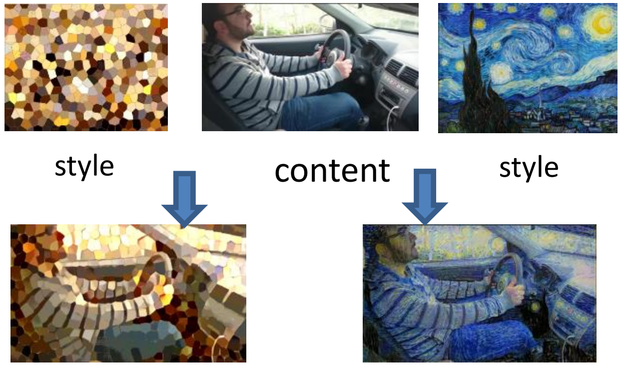
Imagenet classification and VOC detection results

Huang et al. "Arbitrary style transfer in real-time with adaptive instance normalization." *ICCV* , 2017.

Neural-style transfer as a data augmentation strategy for DDD

- The number of Images in Distracted Driver Dataset: 14478 images.
- By reducing iterations to 200 iterations, we need 3.35 days.

FIGURE 1: Ten classes of driver postures from our dataset.



Label smoothing in image classification

- Label Smoothing (Rafael et al., 2019) is a regularization technique that introduces noise for the labels. This accounts for the fact that datasets may have some mistakes in them, so maximizing the likelihood of log p(y|x) might result in over-fitting.
- Assume for a small constant ϵ, the target value of the training label y is 1−ϵ and ϵ/k−1 for the target class and others, respectively. i.e., the original target value of each class is

$$P_i = \begin{cases} 1, i = y, \\ 0, i \neq y. \end{cases} \quad (1)$$

After label-smoothing, they become

$$P_i = \begin{cases} 1 - \epsilon, i = y, \\ \epsilon/(k-1), i \neq y. \end{cases} \quad (2)$$

Therefore, for cross-entropy loss

$$Loss = -\sum_{i=1}^{k} p_i log\, q_i, \quad (3)$$

Experimental Results

- Under the same backbone-ResNet-50, Label Smoothing is always helpful if cross-entropy is applied.
- The backbone trained by Stylized-ImageNet, learns to capture shapes instead texture so that the model is more robust against noise.
- The backbone also generalizes to EuroFOT well.

Table 3 *ResNet-50 results on Distracted Driver Dataset*

Pretraining	Finetuning	Label smoothi ng	Accuracy
ImageNet	DDD	No	81.695% (Eraqi et al., 2019)
ImageNet	DDD	Yes	86.95%
ImageNet + Stylized-ImageNet	DDD	Yes	88.05%
ImageNet + Stylized-ImageNet	DDD+ Stylized-DDD	Yes	89.01%
ImageNet + Stylized-ImageNet	EuroFOT	Yes	84.72%

Conclusion and Future work

- We expect to apply the same model on phone-usage and hands-on-wheel classification.
- Most driving monitoring application is actually a multi-label image classification application, i.e., the driver might actually be talking (a distraction behavior) and using smartphone with his left hand (a phone usage class) simultaneously.
- Large-scale ensemble is actually an option when this application is applied on a labelling system.
- Driver distraction classification can actually be trained in semi-supervised setting via mean teacher model.

CHALMERS
UNIVERSITY OF TECHNOLOGY

Thank you for listening!