Pose-aware Multi-feature Fusion Network for Driver Distraction Recognition

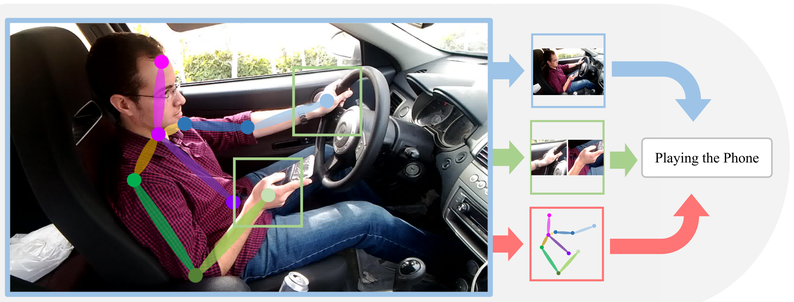
Mingyan Wu[1, 2], Xi Zhang[1, 2], Linlin Shen[1, 2*], Hang Yu[3]

[1]Computer Vision Institute, College of Computer Science and Software Engineering
[2]Guangdong Key Laboratory of Intelligent Information Processing
[3]Shenzhen Laboratory of IC Design for Internet of Things

Abstract

We propose a novel multi-feature fusion network based on pose estimation, for image based distracted driving detection. Since hand is the most important part of driver to infer the distracted actions, our proposed method firstly detects hands using the human body posture information. In addition to the features extracted from the whole image, our network also include the important information of hand and body posture. The global feature, hand and pose features are finally fused by concatenation of feature maps. The experimental results show that our method achieves state-of-the-art performance on SZ Bus Driver dataset and AUC dataset.

Motivation

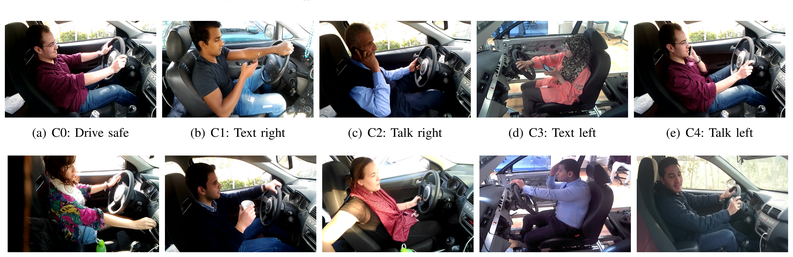


- The whole original image contains global information.
- The actions of hand are important cues in driver distraction recognition.
- The pose information is robust against the interference of backgrounds.

Datasets

(a) C0: Playing the phone (b) C1: Talking on the phone (c) C2: Smoking (d) C3: Safe driving

Fig: Examples of SZ Bus Driver dataset.



(a) C0: Drive safe (b) C1: Text right (c) C2: Talk right (d) C3: Text left (e) C4: Talk left

(f) C5: Adjust radio (g) C6: Drink (h) C7: Reaching behind (i) C8: Hair and makeup (j) C9: Talk to passanger

Fig: Examples of AUC Distracted Driver dataset.

Method

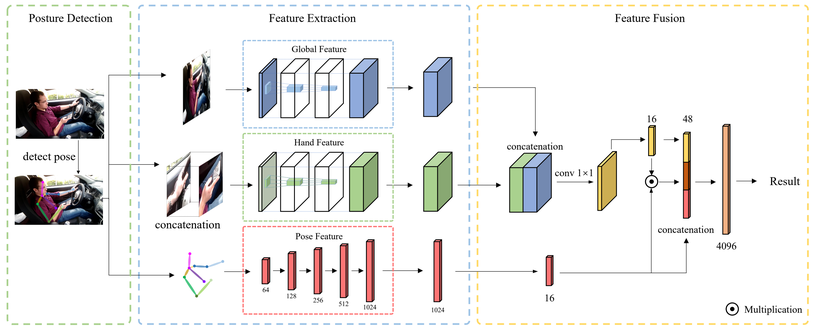


Fig: Pose-aware multi-feature fusion network.

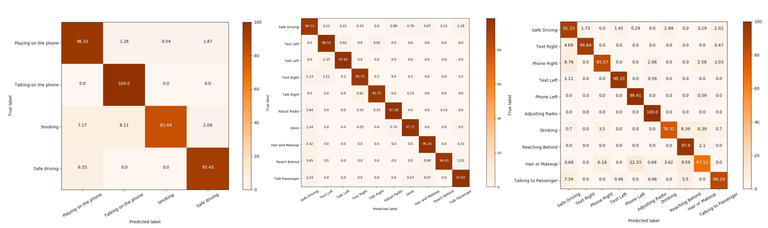
Experimental Results

Table: Result on SZ Bus Driver dataset

Feature	Backbone	C0	C1	C2	C3	Total
Global	VGG-16	74.04	83.01	41.78	85.80	73.35
	ResNet-50	87.11	47.89	48.85	83.48	75.70
	InceptionV3	83.19	82.14	22.48	90.53	77.28
Late Fusion	VGG-16	90.84	84.55	72.25	92.45	88.78
	ResNet-50	93.12	92.78	75.36	91.14	90.87
	InceptionV3	95.85	90.63	80.87	92.74	92.93
Early Fusion	VGG-16	94.13	95.38	82.35	66.16	91.09
	ResNet-50	95.43	**99.46**	64.13	94.76	92.58
	InceptionV3	**96.46**	97.66	**89.40**	**95.27**	**95.75**

Table: Result on AUC V1 and V2 dataset

Dataset	Method	Accuracy
AUC V1	GA-Weighted Ensemble (2017)	95.98
	DenseNet+Latent Pose (2018)	94.20
	VGG with Regularization (2018)	**96.31**
	I3D-two stream (2019)	77.10
	AlexNet+HOG features (2019)	93.19
	Our method	96.28
AUC V2	GA-Weighted Ensemble (2019)	90.07
	Our method	**90.38**



(a) SZ Bus Driver (b) AUC V1 (c) AUC V2

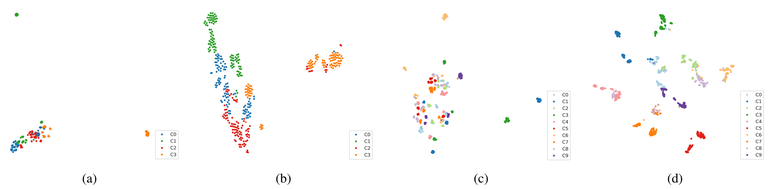
Fig: The confusion matrix of three datasets.

Ablation Study

Table: Ablation Study on three datasets

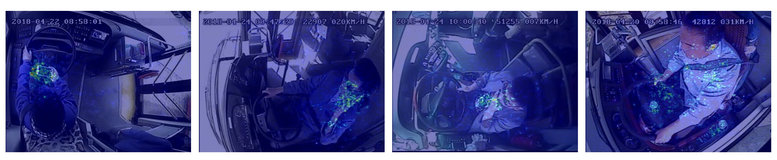
Dataset	Feature			Accuracy
	Global	Hand	Pose	
SZ Bus Driver	✓			77.28
		✓		85.58
			✓	88.29
		✓	✓	91.35
	✓	✓		88.68
	✓		✓	91.84
	✓	✓	✓	**95.75**
AUC V1	✓			95.22
		✓		90.86
			✓	91.36
		✓	✓	92.06
	✓	✓		95.65
	✓		✓	95.52
	✓	✓	✓	**96.28**
AUC V2	✓			85.12
		✓		67.86
			✓	74.88
		✓	✓	79.36
	✓	✓		87.15
	✓		✓	87.31
	✓	✓	✓	**90.38**

Visualization



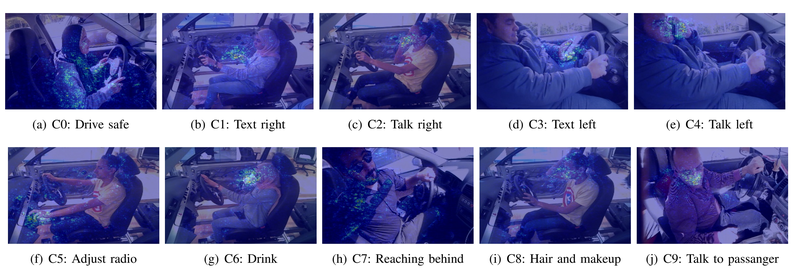
(a) (b) (c) (d)

Fig: The t-SNE visualization of global feature for SZ dataset (a) and AUC V2 dataset (c) and that of fused feature for SZ dataset (b) and AUC V2 dataset (d).



(a) C0: Playing the phone (b) C1: Talking on the phone (c) C2: Smoking (d) C3: Safe driving

Fig: The saliency maps on SZ Bus Driver dataset.



(a) C0: Drive safe (b) C1: Text right (c) C2: Talk right (d) C3: Text left (e) C4: Talk left

(f) C5: Adjust radio (g) C6: Drink (h) C7: Reaching behind (i) C8: Hair and makeup (j) C9: Talk to passanger

Fig: The saliency maps on AUC Distracted Driver dataset.