Jack of many Faces: A Step Towards Facial Expression and Physiological State Analysis with a Single Network

Abdullah Tariq[1], **Martin Masek**[1] ,**R. Muhammad Atif Azad**[3], **Syed Zulqarnain Gilani**[1,2]

[1] *Center for Artificial Intelligence & Machine Learning, Edith Cowan University, Joondalup, Western Australia*
[2] *Department of Computer Science, University of Western Australia*
[3] *Department of Computer Science,Birmingham City University, UK*

Introduction

- **Human face** encodes a wealth of information, providing **critical cues** related to i**dentity, emotion, and cognitive states**.
- Both l**ocal and global** spatio temporal facial patterns are shared across a wide spectrum of affective and cognitive states.
- Developing a **unified facial representation model** that can generalize across diverse facial analysis tasks remains an open and underexplored challenge.

Challenges

- To **effectively disentangle facial features** from videos for making them transferable to down stream tasks**.**
- Using **adversarial learning** as a regualrisation instead of attacking.

Major Challenges: To Differentiate (Happy, Surprise), (Intoxicated, Drowsy, Fatigue)as they exhibit almost similar behavior

Our Contributions

- **First study** to develop a **single** facial feature analysis framework that **generalizes across multiple tasks**.
- A **novel ConAdv-ELCA module** designed to capture crucial l**ocal features, enhance feature disentanglement, and improve generalizability**. Also a unique way to align features.
- Extends non-local networks with dynamic channel wise interactions, enabling efficient modeling of inter channel context.

Datasets

- Used the **MiX Telematics dataset, NTHU-DDD, DFEW,AFEW and FERv39K.**

Figure 1. ***: Sample images showing spatial indicators of intoxication: (Top) Drooped eyelids and flushed cheeks/chin in intoxicated state; (Bottom) No such signs in sober state.***

Methodology

- **Face Detection & Alignmen**t Using 5 facial landmarks to align the face to extract facial features.
- **3D-Spatio-Temporal Network**: To help analyse video sequences.
  → ConAdv-ELCA A dual-path attention mechanism to disentangle emantically relevant facial features from early feature maps.
  → EGCA Enabling efficient modeling of inter channel context to enhance global feature representation in facial video analysis

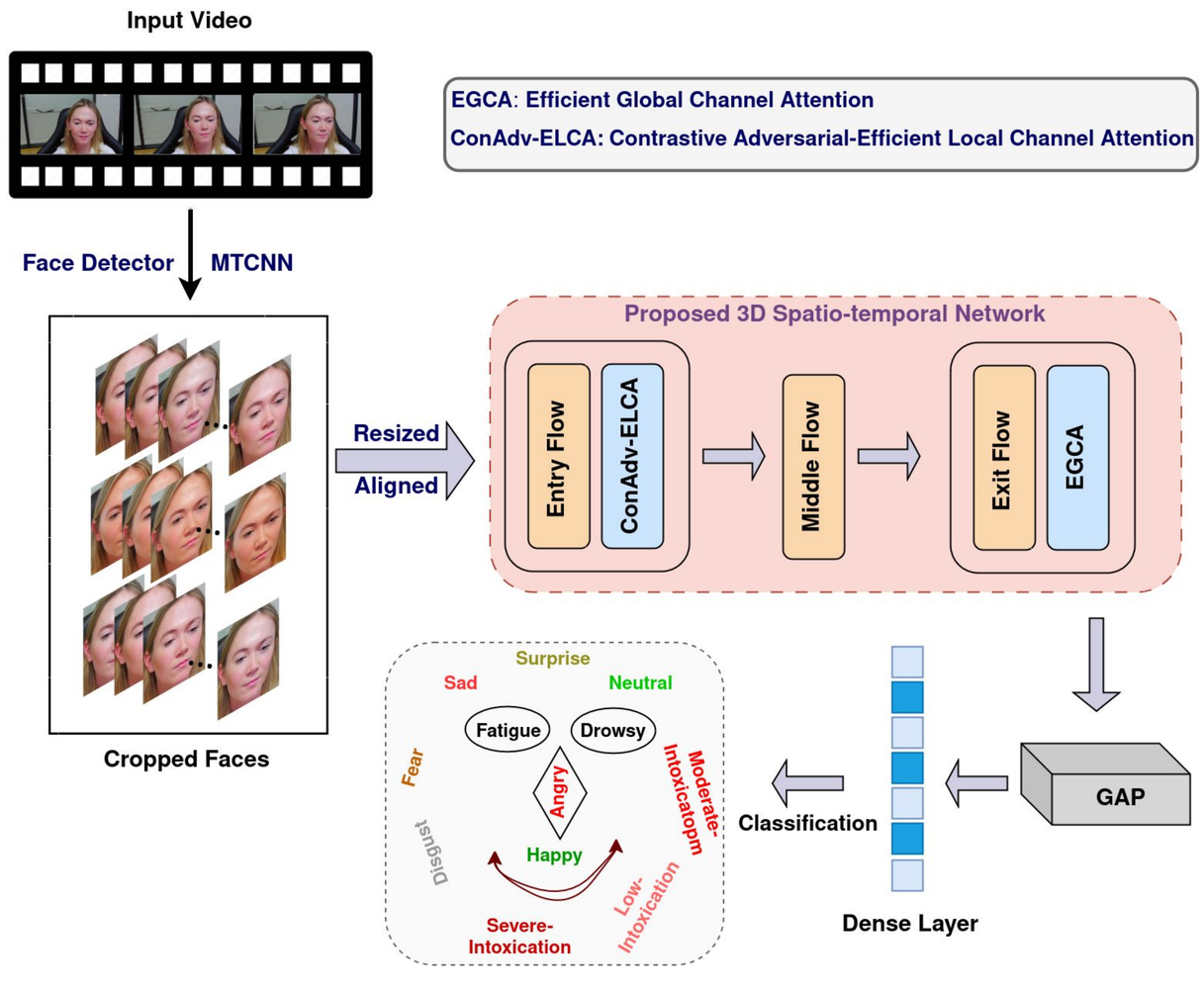

Figure 2. ***Overview of the Proposed Architecture***

Experimental Results

Table 1. ***Comparison with Previous Work on BAC Estimation***

Ref.	Data	Class type	AUROC	Accuracy	F1
[23]	IR Videos/Eye tracking system	Binary 0.05	0.88	0.76	0.75
		Binary 0.08	-	-	-
		Multi-class	-	-	-
[28]	IR Videos/Eye tracking system	Binary 0.05	-	-	-
		Binary 0.08	0.80	0.82	-
		Multi-class	-	-	-
[21]	RGB Videos/Conventional camera	Binary 0.05	-	-	-
		Binary 0.08	-	-	-
		Multi-class	-	0.75	0.75
Ours	RGB Videos/Conventional camera	Binary 0.05	**0.91**	**0.87**	**0.88**
		Binary 0.08	**0.93**	**0.90**	**0.89**
		Multi-class	**0.90**	**0.87**	**0.86**

Binary 0.05*: Sober [BAC < 0.05] or Drunk [BAC ≥ 0.05] g/dL*
Binary 0.08*: Sober [BAC < 0.08] or Drunk [BAC ≥ 0.08] g/dL*
Multi-class*: Sober [BAC = 0.00], Moderate [0.05 ≤ BAC ≤ 0.07], or Severe [BAC ≥ 0.08] g/dL*

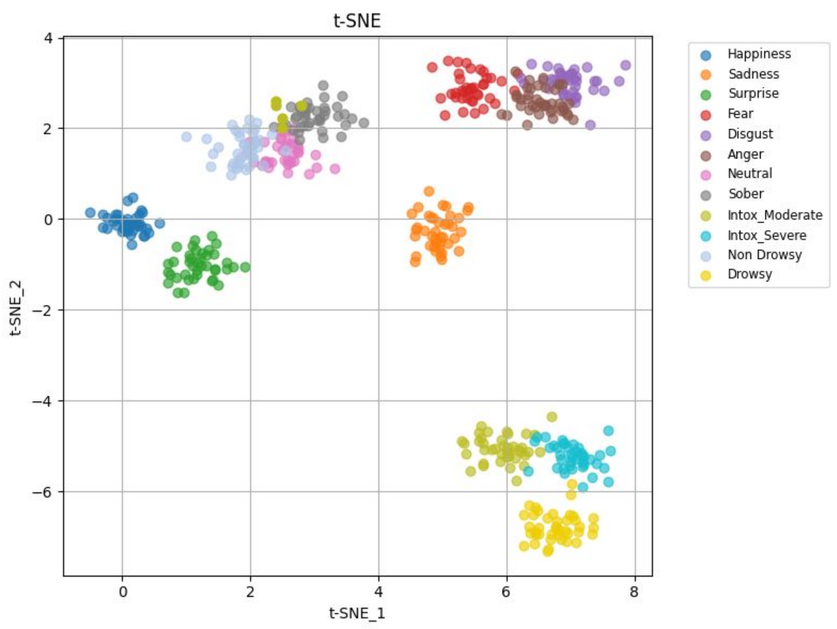

Figure 3. ***Visual analysis of data points indicates that drowsiness and blood alcohol concentration (BAC) classes are distinctly separated***

Table 2. ***Performance Comparison on Drowsiness***

Methods	Accuracy (%)
CNN + ConvCGRNN [53]	84.81
Hierarchical Temporal DBN-HMM [60]	84.82
CNN + TSC-LSTM-Refinement [15]	84.85
ResNeXt3D-101 + LSTM [67]	88.60
2s-STGCN [1]	92.70
IsoSSL-MoCo [40]	93.71
VBFLFLA [63]	93.10
Ours	**95.04**

Table 3. ***Performance Comparison on Emotion Recognition***

Methodology	DFEW		FERV39k		AFEW	
	UAR (%)	WAR (%)	UAR (%)	WAR (%)	UAR (%)	WAR (%)
Uni-modal						
I3D-RGB [4]	-	-	-	-	41.86	45.41
Resnet18+GRU [8, 17]	51.68	64.02	-	-	43.75	46.72
EmotiW-2019 Baseline [9]	-	-	-	-	-	38.81
Resnet18+LSTM [14, 17]	51.32	63.85	-	-	43.96	48.82
3D ResNet18 [16]	-	-	-	-	42.14	45.67
AEN [27]	56.66	69.37	38.18	47.88	50.88	54.60
NR-DFERNet [29]	55.77	68.01	33.99	45.97	48.37	53.54
GCA+IAL [30]	55.71	69.24	-	-	-	-
IAL [30]	-	-	35.82	48.54	-	-
STT [36]	54.58	66.65	37.76	48.11	49.11	54.23
C3D [51]	-	-	-	-	43.75	46.72
R(2+1)D [52]	-	-	-	-	42.89	46.19
M3DFEL [55]	-	-	35.94	47.67	-	-
DPCNet [59]	55.13	66.32	-	-	47.86	51.67
Freq-HD [59]	-	-	33.07	45.26	-	-
Former-DFER [68]	53.69	65.70	37.20	46.85	47.42	50.92
Ours	**60.89**	**71.98**	**44.08**	**53.78**	**54.02**	**56.77**

Conclusion

- Proposed study establishes state-of-the-art BAC estimation using facial video and differentiate between related cognitive states..
- **Proves potential for non-invasive physiological state monitoring systems.**

Future: Deploy on **edge devices**, enhance **robustness in the wild**

EDITH COWAN UNIVERSITY

Let's Connect

Abdullah Tariq a.abdullah@ecu.edu.au

CENTRE FOR ARTIFICIAL INTELLIGENCE & MACHINE LEARNING CAIML