2020「中技社科技獎學金」
2020 CTCI Foundation Science and Technology Scholarship
境外生研究獎學金
Research Scholarship for International Graduate Students

A Study on Deep Learning-based Approaches for Emergency Vehicle Detection

3rd year PhD student: Van-Thuan Tran, Advisor: Prof. Wei-Ho Tsai

Department of Electronic Engineering, National Taipei University of Technology

Abstract

This study investigates deep learning-based approaches for emergency vehicle detection (EVD). Recognizing that car drivers may sometimes be unaware of the approaching emergency vehicles (EVs), leading to delayed responses, we developed systems that can accurately detect the nearby EVs based on their siren sound and visual presence, and then alerts drivers to respond appropriately. We applied audio recognition together with object detection techniques to build EVD systems. The proposed systems will be helpful for preventing traffic accidents and providing the safety function for autopilot systems.

Research Focus

Audio-based EVD (A-EVD): Siren Sound Detection

Background

- ✓ Siren sounds: Special signals.
- ✓ Each country has its own Spec.
- ✓ Traffic soundscape: Siren, Car Horn, and Noise.

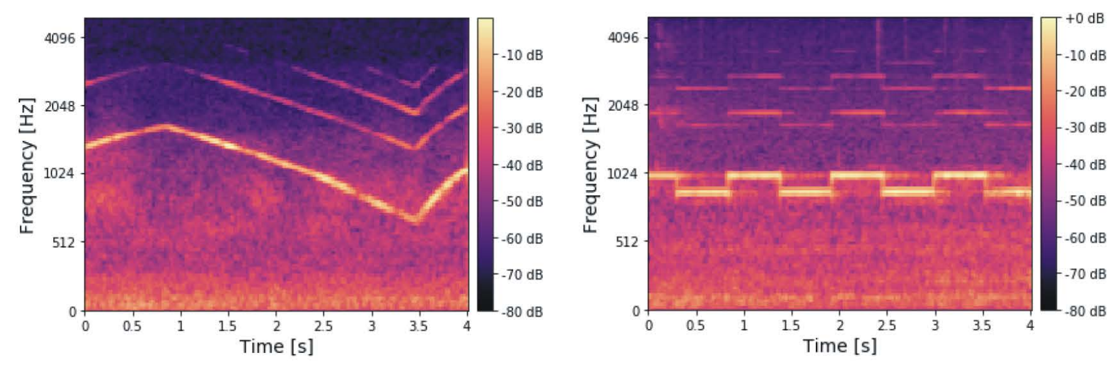

Fig 1. Spectrograms of two siren sound examples: Wail (left), and Two-tone siren (right).

Network Architectures

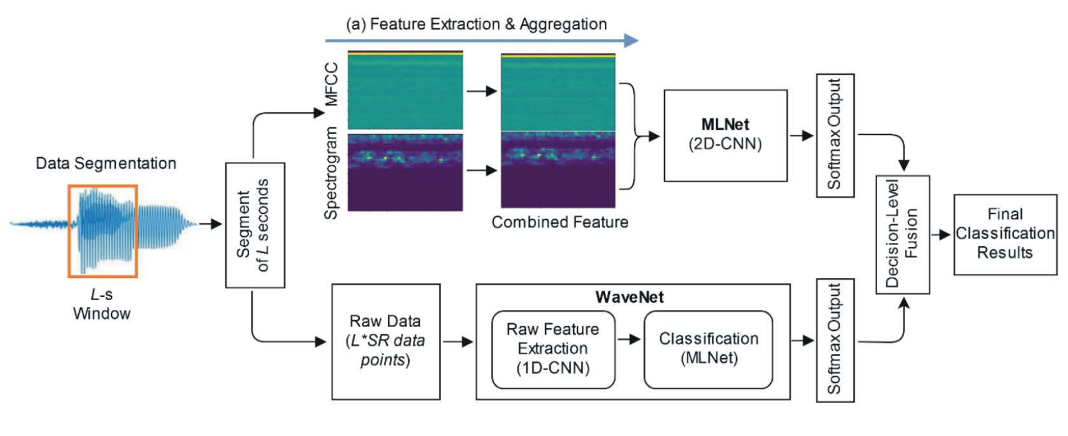

Fig 2. The framework of the CNN-based ensemble model (SirenNet) for A-EVD, in which *SR* is the sampling rate.

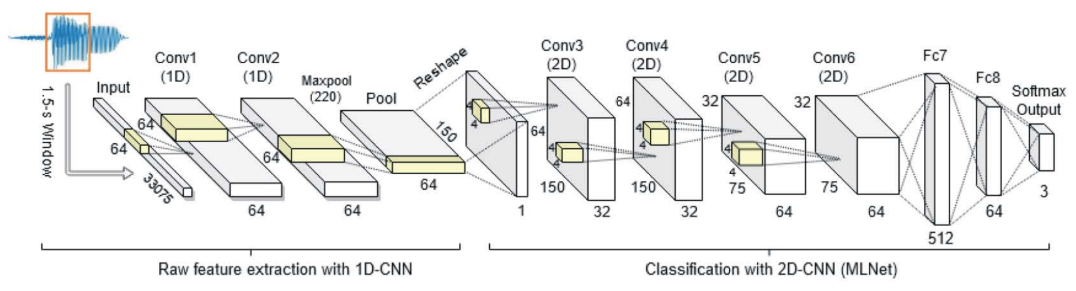

Fig 3. The end-to-end CNN model (WaveNet) for A-EVD.

Experiments & Results

Table I. Summary of experimental dataset in A-EVD

Data Class	Data Sources			
	Our Collection	UrbanSound8K	ESC-50	*Total (no. samples)*
Siren Sounds	**7,773**	929	40	8,742
Car Horns	**7,083**	429	40	7,552
Urban Noise	**1,087**	7,374	1,920	10,381
Total (no. sample)	**15,943**	**8,732**	**2000**	**26,675**
Total duration	17.7 hours	9.7 hours	2.8 hours	30.2 hours
Length of each clip	4 seconds	1-4 seconds	5 seconds	-
Sampling rate	44.1 kHz	8-192 kHz	44.1 kHz	-

Table II. Mean Accuracies of the init-MLNet and the init-WaveNet according to different input lengths.

Model (Feature)	Input length (s)					
	0.25	0.5	1	1.5	2	3
init-MLNet (log-mel)	78.58 (1.42)	80.54 (1.16)	85.24 (1.72)	91.18 (0.39)	92.15 (0.42)	92.56 (1.41)
init-MLNet (MFCC)	87.30 (0.54)	88.78 (0.63)	88.88 (0.77)	89.71 (0.67)	89.55 (0.98)	90.67 (0.61)
init-MLNet (MFCC+log-mel)	**91.21** (0.55)	**92.07** (0.53)	**92.80** (0.49)	**94.26** (0.51)	**94.17** (0.36)	**94.48** (0.96)
Init-WaveNet (Raw data)	91.76 (0.10)	92.19 (0.54)	92.89 (0.11)	93.99 (0.26)	94.70 (0.28)	94.79 (0.25)

Table III. Results of the SirenNet and comparison to single networks

Model	Accuracy (%)	Model loading time (s)	Inference time (s)
SirenNet (Raw data, MFCC+log-mel)	**98.24** (0.36)	0.805	0.027
WaveNet (Raw data)	96.51(0.31)	0.389	0.011
MLNet (MFCC+log-mel)	96.42 (0.45)	0.415	0.014

Table IV. Results of the proposed models on different input durations

Model	Input Length (s)			
	1.5	1	0.5	0.25
SirenNet (Raw data, MFCC+log-mel)	98.24 (0.36)	97.74 (0.79)	97.42 (0.39)	96.89 (0.57)
WaveNet (Raw data)	96.51 (0.31)	95.59 (0.32)	94.43 (0.51)	92.20 (0.65)
MLNet (MFCC+log-mel)	96.42 (0.45)	95.49 (0.16)	95.37 (0.46)	94.47 (0.59)

Table V. The comparison with other systems

Work	Feature(s)	Model/Method	Accuracy (%)
L. Marchegiani et al., 2017	Empirical Binary Masks (EBMs), Gammatonegrams (GTG), MFCC	k-NN	83.00 ± 1.30 (EBMs), 75.60 ± 1.14 (GTG), 62.20 ± 2.77 (MFCC)
J. Schroder et al., 2013	MFCC, Spectrogram	Part-based Models (PBMs), HMM	86.00 (PBMs), 80.00 (HMM+MFCC), 74.00 (HMM+Spectrogram)
J.J. Liaw et al., 2013	Longest Common Subsequence (LCS)	LCS Comparison	85.00
This work	Raw data	1D-CNN (WaveNet)	**96.51**
This work	MFCC+Spectrogram	2D-CNN (MLNet)	**96.42**
This work	Aggregated features: Raw data, MFCC, Spectrogram	CNN (SirenNet)	**98.24**

A-EVD Summary

The proposed SirenNet (Fig.2) is composed of two networks, including the WaveNet (Fig.3) which works with raw waveform, and the MLNet trained on handcrafted features. We conducted experiments on an extensive dataset of 30-hour recordings (Table I). The WaveNet and MLNet yielded accuracies of 96.51% and 96.42%, respectively. Equally important, the ensemble architecture (SirenNet) further boosted the classification accuracy to 98.24%, which is much higher than results of prior works. Also, the SirenNet requires a short inference time of 27ms, so it is well acceptable for real-time detection.

Vision-based EVD (V-EVD): Object Detection

- ✓ YOLO Algorithm (EVD-YOLO)
- ✓ Binary Detector: EV/Non-EV

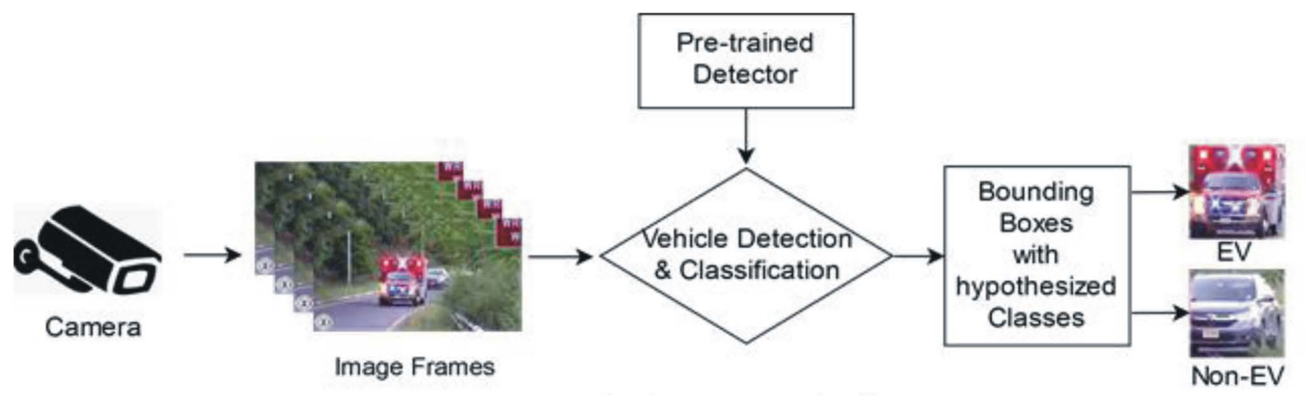
Pre-trained Detector
Camera
Image Frames
Vehicle Detection & Classification
Bounding Boxes with hypothesized Classes
EV
Non-EV

Fig 4. V-EVD Pipeline.

EVD-YOLO

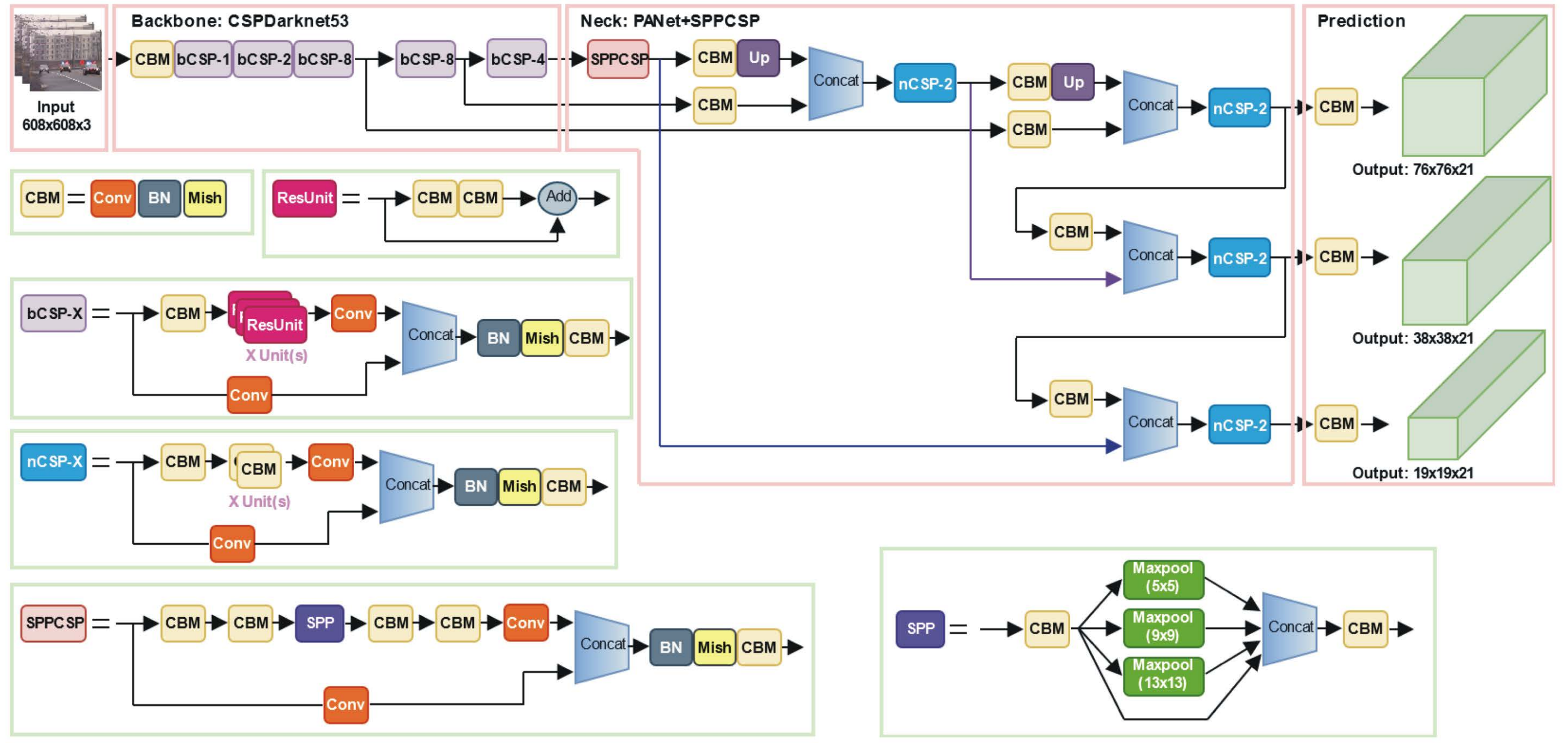

Fig 5. Architecture of the modified YOLOv4 for V-EVD (EVD-YOLO).

Experiments & Results

Table VI. Results of V-EVD systems.

Object Detector	Image Size	mAP	Time cost (ms)
EfficientDet-D0	608	82.5	25.8
EfficientDet-D1	608	87.5	29.4
EfficientDet-D2	608	90.3	35
EfficientDet-D3	608	92.2	44
EfficientDet-D4	608	91.2	67.5
YOLOv3	608	**92.9**	**18**
YOLOv4	608	**93.4**	**4.5**
YOLOv4 (Mish-neck)	608	**94.3**	**4.7**
YOLOv4 (SPPCSP)	608	**93.8**	**4.7**
YOLOv4 (CSPneck)	608	**94.0**	**4.9**
YOLOv4 (Mish-neck +SPPCSP+CSPneck) **→ EVD-YOLO**	608	**95.1**	**4.9**

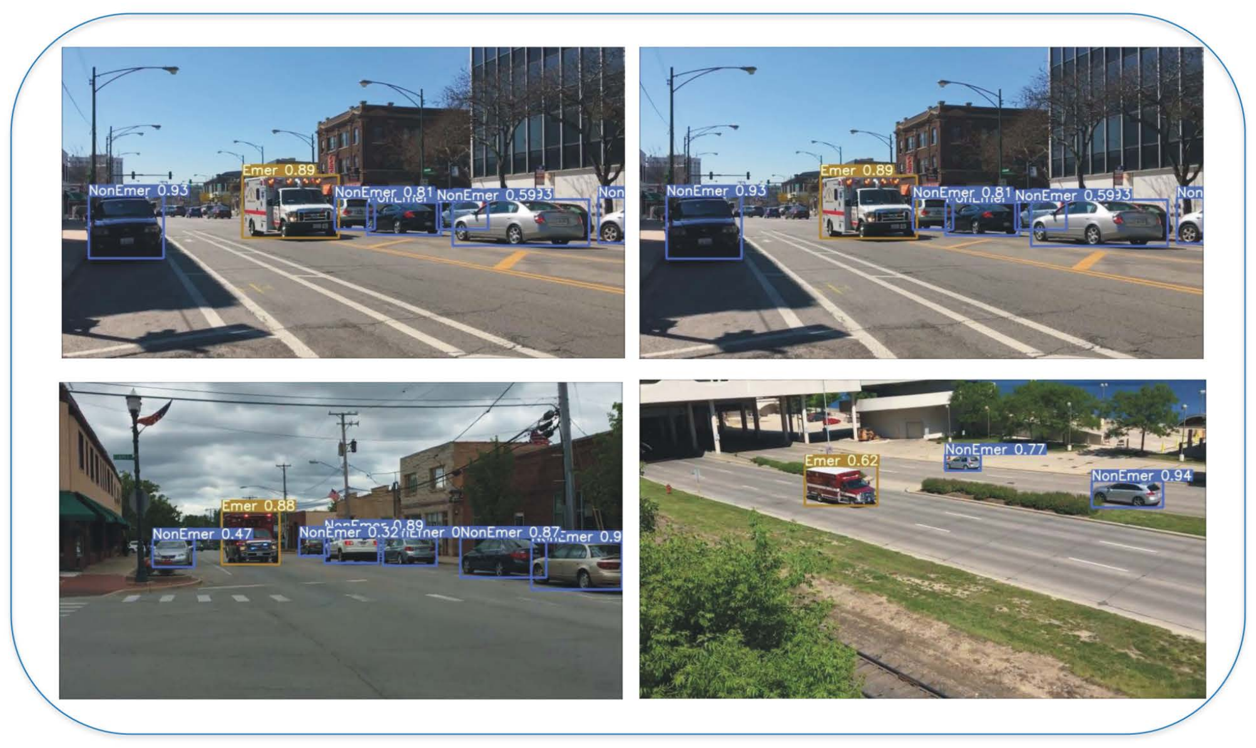

Fig 6. Examples of EVD-YOLO Detection.

V-EVD Summary

- ✓ Dataset: 20,000 images.
- ✓ YOLO showed promising results.
- ✓ EVD-YOLO further boosted the accuracy.
- ✓ Real-time processing.

Further Work: Proposed AV-EVD

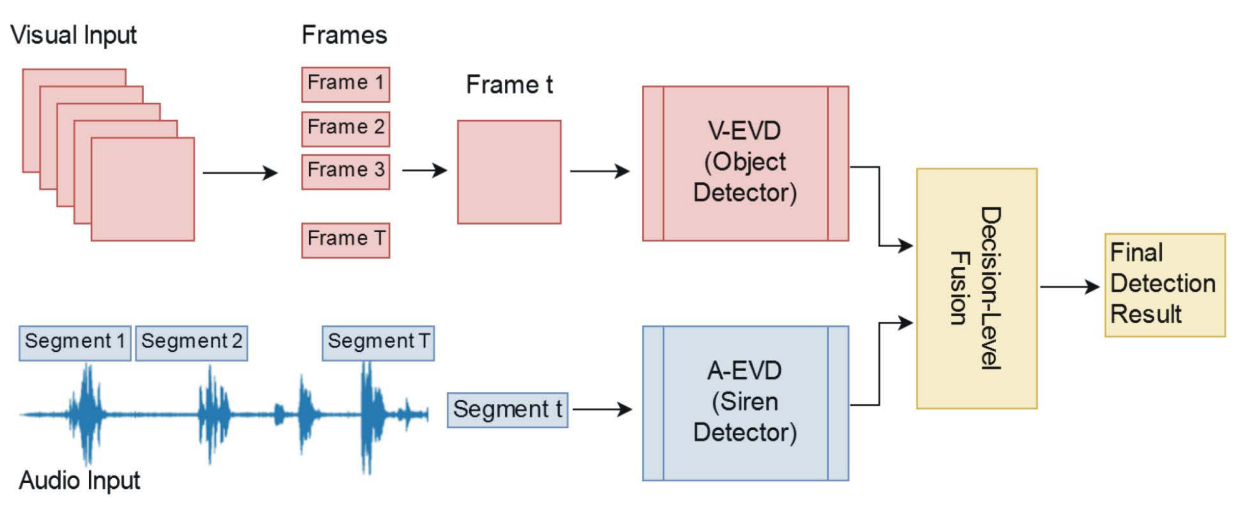

Fig 7. The proposed Audio-Vision-based EVD system.

Publications

1. **Van-Thuan Tran** and Wei-Ho Tsai, “Acoustic-Based Emergency Vehicle Detection Using Convolutional Neural Networks,” IEEE Access, vol. 8, pp. 75702-75713, 2020, doi: 10.1109/ACCESS.2020.2988986.
2. **Van-Thuan Tran** and Wei-Ho Tsai, “Detection of Ambulance and Fire Truck Siren Sounds Using Neural Networks,” 51st Research World International Conference, Hanoi, 2018/07/26.

財團法人中技社
CTCI FOUNDATION