Deep Sequential Image Features for Acoustic Scene Classification

Zhao Ren[1,2], Vedhas Pandit[1,2], Kun Qian[1,2,3], Zijiang Yang[1,2], Zixing Zhang[2], Björn Schuller[1,2,4]

[1]Chair of Embedded Intelligence for Health Care & Wellbeing, Universität Augsburg, Germany
[2]Chair of Complex & Intelligent Systems, Universität Passau, Germany
[3]MISP Group, MMK, Technische Universität München, Germany
[4]GLAM – Group on Language, Audio & Music, Imperial College London, UK

Introduction

- **Motivations:**
  - **2D spectrograms** are applied successfully in acoustic scene classifidcation
  - **Wavelet transform** incorporates multiple scales and localisations
- **Major Contributions:**
  - Use **scalograms** to extract powerful representations
  - Combine pre-trained CNNs with GRNNs by **transfer learning**

Deep Sequential Images

- The short-time Fourier transform (STFT) for a signal $x(t)$ is defined by,

$$X(\tau,\omega)=\int_{-\infty}^{\infty}x(t)\omega(t-\tau)e^{-j\omega t}, \quad (1)$$

where t: time, $\omega(t)$: window function, τ: time index.

- The *bump* wavelet transform is defined by,

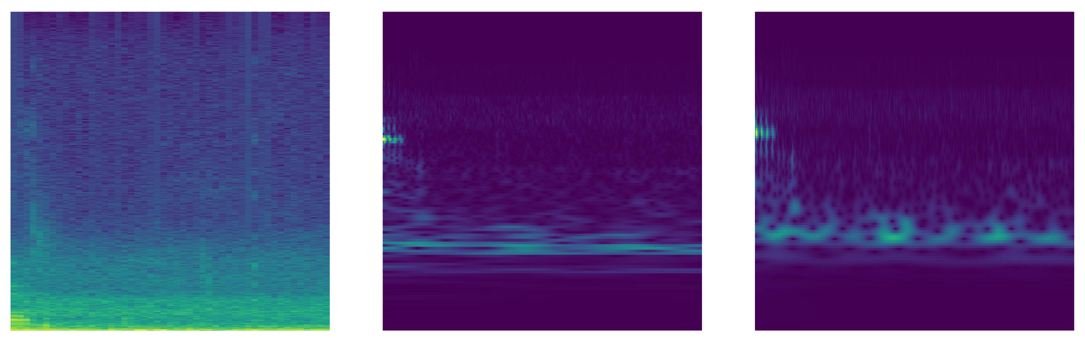
$$\Psi(s\omega)=e^{\left(1-\frac{1}{1-(s\omega-\mu)^2/\sigma^2}\right)}1_{[(\mu-\sigma)/s,(\mu+\sigma)/s]}, \quad (2)$$

where s: scale, ω: window, μ and σ: two constant parameters.

- The *morse* wavelet generation is defined by,

$$\Psi_{P,\gamma}(\omega)=u(\omega)\alpha_{P,\gamma}\omega^{\frac{P^2}{\gamma}}e^{-\omega^{\gamma}}, \quad (3)$$

where $u(\omega)$: unit step, ω: window, $\alpha_{P,\gamma}$: a normalising constant, P: time-bandwidth product, γ: symmetry.



(a) STFT (b) *bump* (c) *morse*

Figure: Images of the first audio sequence of "*a001_0_10.wav*" with a label *residential area*.

Feature Extraction and Classification

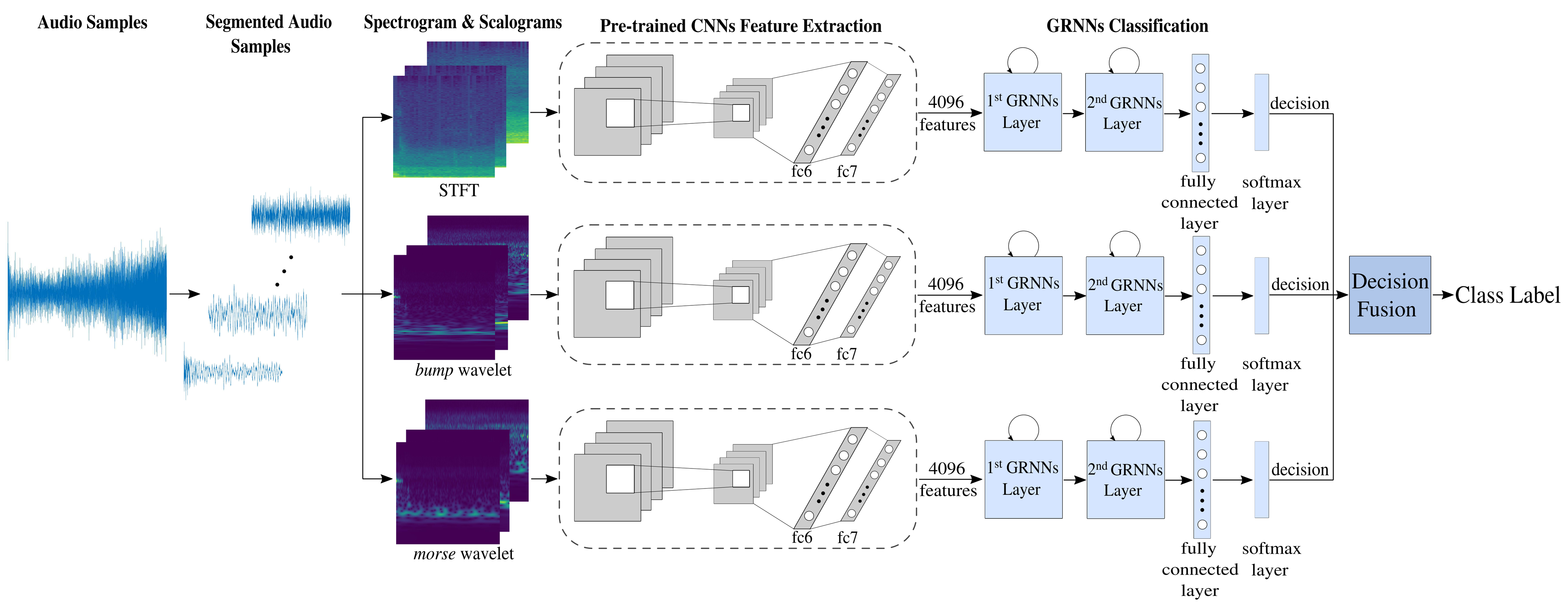


Figure: Framework of our proposed system.

Table: Configurations of the VGG16 CNNs.

Input: 224×224 RGB image
2×conv size: 3; ch: 64; Maxpooling
2×conv size: 3; ch: 128; Maxpooling
3×conv size: 3; ch: 256; Maxpooling
3×conv size: 3; ch: 512; Maxpooling
3×conv size: 3; ch: 512; Maxpooling
fc6 layer with 4096 neurons
fc7 layer with 4096 neurons
fc layer with 1000 neurons
Output: softmax layer for 1000 classes

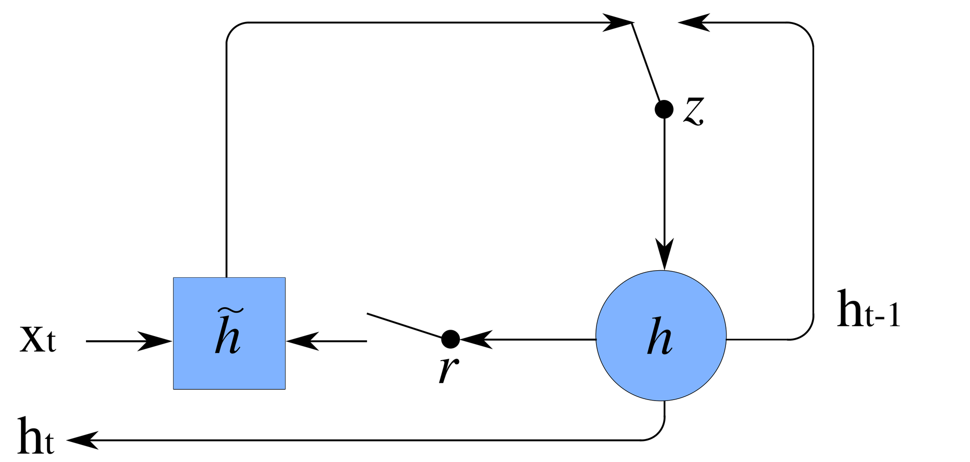


Figure: Illustration of a Gated Recurrent Unit (GRU).

- **Feature Extraction:**
  - VGG16 Convolutional Neural Networks (CNNs)
- **Classification:**
  - Gated Recurrent Neural Networks
  - Decision Fusion by Margin Sampling Value (MSV)

Database

- DCASE 2017 database:
  - 312 segments of 10 *s* from 52 minutes of audio recordings
  - 15 classes: *beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station, office, park, residential area, train,* and *tram*
  - An unlabelled evaluation set and four folds, each of which contains a training set and a development set

Experimental Results

Table: Performance comparison of different epochs of GRNNs(120-60), *learning rate*=0.0002.

accuracy [%]	Fold1	Fold2	Fold3	Fold4	Mean
(a) STFT					
epoch 30	77.9	72.5	73.1	79.3	75.7
epoch 50	79.2	74.7	74.3	77.7	**76.5**
epoch 70	77.1	75.8	72.9	77.4	75.8
(b) *bump* wavelet					
epoch 30	74.5	75.4	73.9	77.2	**75.2**
epoch 50	73.6	72.9	73.6	73.2	73.3
epoch 70	69.7	73.4	72.6	72.1	72.0
(c) *morse* wavelet					
epoch 30	74.5	75.4	73.9	77.2	**75.2**
epoch 50	73.6	72.9	73.6	73.2	73.3
epoch 70	69.7	73.4	72.6	72.1	72.0

Table: Performance comparison of different combinations of the three feature sets by decision fusion on GRNNs(120-60), *learning rate*=0.0002.

accuracy [%]	Fold1	Fold2	Fold3	Fold4	Mean
STFT+bump	82.6	79.5	77.5	80.9	80.1
STFT+morse	81.1	80.0	76.5	81.5	79.8
bump+morse	76.7	77.5	76.0	77.5	76.9
STFT+bump+morse	82.6	80.7	78.7	81.5	**80.9**

Conclusions

- Classify deep STFT and wavelet features on GRNNs
- Wavelet features are helpful to increase the accuracy
- Future work:
  - Investigate which CNNs infer the best representations
  - Experiment with data augmentations of the training data

Acknowledgements

This work was partially supported by the European Union's Seventh Framework under grant agreements No. 338164 (ERC StG iHEARu), and the China Scholarship Council (CSC).