

500933321

EE8209

PROFESSOR: **Ghassem Tofighi**

proposal

Syeda Maham Rehan

**PROJECT PROPOSAL**

TITLE:

Multimodal scene reorganization for audio Categorization

INTRODUCTION:

Modality is basically the way in which something happens or is experienced and a research problem is said to be multimodal when it includes such multiple modalities. To make Artificial Intelligence progressive in understanding the world around us, it needs to be able to interpret such multimodal signals together. Multimodal machine learning aims to build models that can process and relate information from multiple modalities. It is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential.

CHALLENGES:

The research field of Multimodal Machine Learning brings some unique and different challenges for computational researchers given the heterogeneity of the data. Learning from multimodal sources gives the opportunity of capturing correspondences between modalities and to have an in depth understanding of natural phenomena.

* Representation

How to represent and summarize multimodal sound data in a way that exploits the complementarity and redundancy of multiple modalities

* Translation

How to translate data from one modality to another

* Alignment

To identify the direct relations between elements

* Fusion

To combine or join the information of two or more modalities to perform a prediction

* Co-learning

To transfer knowledge between modalities, their representation, and their predictive models

WHY THIS PROJECT?

Multimodal scene recognition in audio is of vital importance in deep audio and video understanding. User-generated videos are usually very short and the content is consistent. Video contains frame sequences and audio. Some scenes are incredibly difficult from images. However, from audio point of view, the task is much important and is the first step towards the scene reorganization. Therefore, general scene recognition with multimodal information is a promising direction.

APPROACH

First, in extracting the most informative features from each modality and second, in the most successful way of fusion both modalities. Ultimately, this leads to an improvement in the accuracy of speech recognition.

REFERENCE:

[1] J.-J. Aucouturier, B. Defreville, and F. Pachet. 2007. The Bag-of-frame Approach to Audio Pattern Recognition: A Sufficient Model for Urban Soundscapes but Not for Polyphonic Music. Journal of the Acoustical Society of America (2007), 881–91.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning Sound Representations fromUnlabeled Video. In Advances in Neural Information Processing Systems. 892–900.

[3] Soo Hyun Bae, Inkyu Choi, and NamSoo Kim. 2016. Acoustic Scene Classification Using Parallel Combination of LSTM and CNN. Technical Report.

[4] V. Bisot, S. Essid, and G. Richard. 2015. HOG and Subband Power Distribution Image Features for Acoustic Scene Classification. In European Signal Processing Conference. 719–723.