Emotion Recognition with Audio-visual input

Kavita Anant ka2744
Raksha Nandanahosur Ramesh rn2486

Advanced Big Data Analytics
Task A1: Face recognition, Feeling Recognition and Interaction

COLUMBIA UNIVERSITY

***Abstract*— Emotion Recognition is an important milestone in Affective computing today. Our project aims to combine visual and auditory input modalities to design a real-time automatic emotion recognition system. We propose to design pipelines to extract features from speech and facial expressions using pre-trained CNN architectures. Next, features will be fused through a feature-level fusion model and an attention mechanism proposed in [1] and validate the models on the AFEW dataset. Our final project will be an empathetic and context-aware conversational agent by feeding emotional cues from the audio-visual ER system to generate responses targeted for the recognized emotion.**

I. INTRODUCTION

Emotion is an intrinsic part of being human and deeply influences perception, communication and thinking. The field of affective computing aims to model human emotional expression to enhance the capabilities of computing technology.

Emotion recognition is an important challenge in Affective computing and has applications spanning many fields. For example, in healthcare, social robots can recognize a patient's emotional state and contribute towards improving mental health services and improve communication technologies for autistic children. Emotion recognition technology can even be used to conduct market research to understand product tester's emotions. Other use cases include affective video games and personalized education technologies to detect a child's engagement levels. [2]

Another important application of Affective computing, which we attempt to address in the final phase of our project is - emotion simulation in dialog agents. But the first step towards achieving more empathetic and personalized machines is accurate emotion recognition through multimodal inputs, which is the main focus of our study.

While emotions can be inferred through multiple input modalities, as human beings we use a combination of speech and facial expressions as universal indicators to understand and interpret a person's emotions. A lot of current commercial technologies [3] [4] [5] albeit comprehensive, involve only an independent analysis of either speech or facial expressions. Thus research in utilizing bimodal inputs to improve ER tasks is required to progress the Affective computing field forward.

II. RELATED WORK

With the advent of deep learning, CNN-RNN based end-to-end emotion recognition models have gained traction when compared to traditional pipelines for Facial emotion recognition (FER) and Speech Emotion Recognition (SER) that involve hand-crafted features. CNNs can produce automatic rich internal representations of videos and RNNs can model the time-varying patterns and capture the sequential evolution in videos accurately.

The EmotiW challenge [6] is a benchmark challenge held every year since 2015 to improve emotion recognition in audio-video datasets, and we observed most teams use CNN based models for the emotion recognition tasks. We reviewed [7] and [8] - two of the best models implemented in this challenge the past two years that obtained a ~60% benchmark accuracy on the AFEW dataset.

In [7] the authors present a hybrid net that uses 4 main feature extraction and classification techniques (i) VGG-LSTM model where features are extracted using VGGnet and emotion details are analysed by LSTM i.e, LSTM layer returns the temporal features based on continuing video frames on facial expressions. (ii) CNN (Inception V3) based features after fine-tuning. (iii) Euclidean distances computed on 3D facial landmarks fed to an SVM classifier (iv) acoustic features are extracted using OpenSMILE [9] pre-trained on SoundNet [10] which is a state-of-the-art network used to extract sound representations from unlabelled sound data collected in the wild. The final output is predicted by taking a weighted sum of each of the outputs of the hybrid net.

Similar to [7], the authors in [8] implement an ensemble of several models and show that using industry level face recognition networks and fine-tuning on the FER2013 data improves the emotion recognition accuracy. The authors also randomly shuffle the order of frames during training of LSTM as a method of augmentation which interestingly contributed to higher accuracies.

We hypothesized that emotion recognition accuracy can be improved further by refining the methodology by which features are combined from multi-modal inputs. In the above mentioned papers, the audio and video features are extracted and trained on two different classifiers. The bimodal outputs are then linearly fused or concatenated or integrated together with a weighted sum. But these models do not capture the inherent interactions present between the audio and video frames. In [1] the authors use a feature-level-fusion model [Fig 1] with bilinear pooling and an attention mechanism elaborated below. We implement our model based on this architecture.

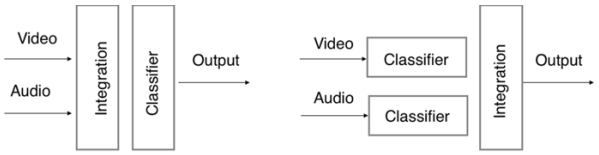

Fig 1: Feature-level fusion model *(left)* decision-level fusion model *(right) [11]*

A. *Salient Features of Reference Papers*

a) Feature fusion: A factorized bilinear pooling mechanism is implemented to deeply integrate the features from both audio and video input modalities.

b) Attention Mechanism: Since not all frames in a video encode emotion information, each frame is weighted based on its importance to the emotion classification task. The same mechanism is applied to the audio features as well i.e, every time-frequency region in the speech spectrogram is weighted based on its contribution toward determining the emotional state. [12]

III. OBJECTIVE AND GOALS

A. *Design a real-time emotion recognition (ER) system using facial expressions and speech as input modalities.*

- Implement a pipeline for Facial Emotion Recognition (FER) – face detection and tracking, face recognition, experiment with feature extraction methods – facial landmarks, CNN based features.
- Implement a pipeline for Speech Emotion Recognition (SER) – preprocess, pre-train and extract CNN based features.
- Train and implement optimized model with attention mechanism to recognize emotions – Anger, disgust, fear, happy, neutral, sad in real-time.
- Evaluate model with different audio-visual datasets.

B. *Integrate emotional cues from ER system into conversational agent to guide responses to be empathetic and context aware*

- Our hypothesis is that leveraging information from both speech and facial expressions (as opposed to using only a single modality) can significantly improve ER accuracies when tested in real-time.

IV. DATASET

Typically datasets for this problem space are based on expressions acted out by actors [13] or expressions induced after subjects listen to stories [14] or spontaneous emotions taken in the wild [6]. We have selected open-source audio-visual corpuses that use these 3 different affective state elicitation methods to implement our ER model. The AFEW dataset contains audio-visual clips taken in the wild i.e. from movies. This is a very challenging dataset to train on due to the number of speakers, background noise, different face orientations and illumination range. But for our use case, it is important to test the model on a dataset where emotions are spontaneous rather than acted out or induced as our goal is to create a real-time ER model. It is also important to note that there is a general lack of consensus about the emotion labels when annotators label the datasets due to the variations in expressions; in such cases good performance on one dataset, does not generalize well to other datasets. Therefore, keeping these challenges in mind, we will experiment with multiple datasets while training - eNTERFACE and or SAVEE dataset and use the AFEW as a test dataset.

Apart from audio-visual datasets, we also want to separately evaluate the models for facial emotion recognition using FER2013 and speech emotion recognition using IEMOCAP (Interactive emotional dyadic motion capture database) [15] as state-of-the art results are published using these corpuses. We also require these datasets for fine-tuning during pre-training the CNN based models.

TABLE I: AUDIO-VISUAL DATABASES

Dataset	***Description***
AFEW (Acted Facial Expressions in the Wild)	330 subjects, 1426 sequences
eNTERFACE'05 Audio-Visual Emotion Database	42 subjects, 1166 video sequences
SAVEE (Surrey Audio Visual Expressed Emotion)	4 subjects, 480 sequences

TABLE II: FACIAL EMOTION RECOGNITON DATABASE

Dataset	***Description***
FER2013	FER-2013 contains 35,887 images, with 4,953 labelled as Anger, 547 as Disgust, 5,121 as Fear, 8,989 as Happiness, 6,077 as Sadness, 4,002 as Surprise, and 6,198 as Neutral

TABLE II: SPEECH EMOTION RECOGNITON DATABASE

Dataset	***Description***
IEMOCAP	10 Actors: 5 male and 5 female, emotion elicitation is through improvisation and scripts. Also contains valence activation and dominance apart from the emotional categories

V. OVERVIEW OF METHODOLOGY

A. *Audio stream:*

When compared to traditional hand crafted features, CNN based speech features can obtain comparable classification accuracies in SER tasks. Therefore the raw speech spectrograms can be fed directly to the CNN. We will also extract features using OpenSMILE which support low-level acoustic descriptors like energy, loudness, pitch etc and statistical features, and pick the feature set that performs the best.

We will pre-train the model on ImageNet as its suggested in [12] that pretraining on a natural scene image-based dataset improves the classification accuracy.

The output of the FCN layers from the AlexNet architecture is fed to the attention block which helps the network focus on emotionally salient regions of the spectrogram.

The pipeline for audio stream is illustrated in Fig 3. We will validate this model on the IEMOCAP dataset to compare the state-of-the-art accuracies.

B. *Video Stream:*

To preprocess the video frames for the FER task, we will first implement face detection and face tracking. This step can be challenging on the AFEW dataset due to the nature of the dataset discussed previously. We will therefore experiment with the other FER databases.

The features are then extracted from the VGG-face network that is fine-tuned on FER2013 to make it emotion relevant and these are fed to the attention block.

C. *Audio-video Fusion:*

For our initial fusion model, we will linearly concatenate the features or use a decision level fusion model and test on the AFEW dataset.

Further, we will implement the improved model as proposed in [1] where the weighed features from the pipelines for audio and video stream are deeply integrated using a factorized bilinear pooling model. This is illustrated in Fig 2.

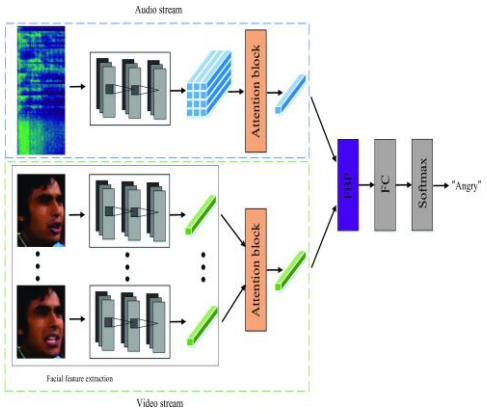

Fig 2: Pipeline for emotion recognition using audio and video streams and their fusion as proposed in *[*cite]*

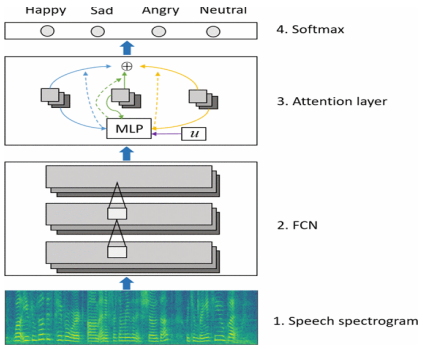

Fig 3: Pipeline for speech emotion recognition system as proposed in [12].

VI. MILESTONE PLANNING

1) Milestone 1:

- Implement initial pipeline to extract facial features – face detection, facial landmark detection. Test face detection and tracking in real-time.
- Preprocess speech – extract speech spectrograms to feed to CNN.

2) Milestone 2:

- Train models for FER and SER individually - fine-tune models on pretrained networks.
- Implement initial fusion model from both speech and facial expressions.

3) Milestone 3:

- Optimize models – hyperparameter tuning, improve time-lags, implement attention guidance mechanism for both FER and SER.
- Test on AFEW and other corpuses using CNN based features and facial landmarks-based features.

4) Final Presentation:

- Merge Project with Team A2[1]: Integrate above ER system with conversational agent from (if time permits) and test system in real-time.

REFERENCES

[1] Y. Zhang, Z.-R. Wang, and J. Du, "Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition," arXiv preprint arXiv:1901.04889, 2019.

[2] Park, Hae Won & Grover, Ishaan & Spaulding, Samuel & Gomez, Louis & Breazeal, Cynthia. (2019). A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education. Proceedings of the AAAI Conference on Artificial Intelligence. 33. 687-694. 10.1609/aaai.v33i01.3301687.

[3] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. 2016. AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16).

[1] *The conversational agent and context awareness is being implemented by Team A2. If time permits, we will attempt to fuse our projects for the final project presentation. The main focus of our project however is only emotion recognition with audio-visual input.*

Association for Computing Machinery, New York, NY, USA, 3723–3726. DOI:https://doi.org/10.1145/2851581.2890247

[4] https://www.ibm.com/watson/services/tone-analyzer/

[5] https://replika.ai/

[6] A. Dhall, R. Goecke, S. Lucey and T. Gedeon, "Collecting Large, Richly Annotated Facial-Expression Databases from Movies," in IEEE MultiMedia, vol. 19, no. 3, pp. 34-41, July-Sept. 2012.

[7] Chuanhe Liu, Tianhao Tang, Kui Lv, and Minghao Wang. 2018. Multi-Feature Based Emotion Recognition for Video Clips. In Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI '18). Association for Computing Machinery, New York, NY, USA, 630–634. DOI:https://doi.org/10.1145/3242969.3264989

[8] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," arXiv preprint arXiv:1711.04598, 2017

[9] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor", In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. doi:10.1145/2502081.2502224

[10] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." Advances in Neural Information Processing Systems. 2016.

[11] Avots, E., Sapiński, T., Bachmann, M. *et al.* Audiovisual emotion recognition in wild. *Machine Vision and Applications* **30,** 975–985 (2019). https://doi.org/10.1007/s00138-018-0960-9

[12] Y. Zhang, J. Du, Z. Wang, and J. Zhang, "Attention based fully convolutional network for speech emotion recognition," *arXiv preprint arXiv:1806.01506*, 2018.

[13] Jackson, Philip & ul haq, Sana. (2011). Surrey Audio-Visual Expressed Emotion (SAVEE) database.

[14] O. Martin, I. Kotsia, B. Macq and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA, 2006, pp. 8-8.

[15] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.