Emotion Recognition Using Fusion of Audio and Video Features

Grigor Keropyan
supervisor: Vazgen Mikayelyan

Department of Mathematics
ASDS program
Yerevan State University

May, 2021

Outline

Why emotion recognition is important?

1. Interpersonal Relationships
2. Human Computer Interaction

Why emotion recognition is important?

1. Interpersonal Relationships
2. Human Computer Interaction

Applications

1. Call centers (Zoom, Hangout, Skype, etc.)
2. Business meetings
3. Tutor Agents

Applications

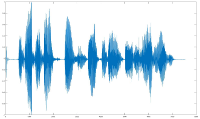
1. Call centers (Zoom, Hangout, Skype, etc.)
2. Business meetings
3. Tutor Agents

Applications

1. Call centers (Zoom, Hangout, Skype, etc.)
2. Business meetings
3. Tutor Agents

Emotion Expression Modalities

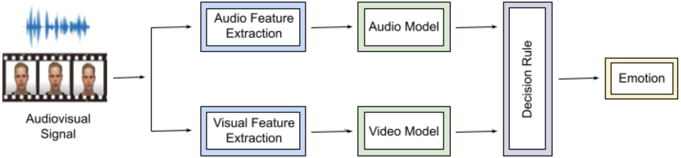
PURE TEXT



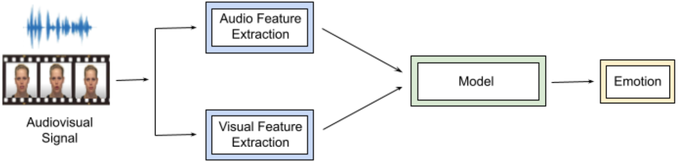
Datasets

1. RAVDESS (Livingstone and Russo, 2018)
2. SAVEE (Philip and Haq, 2014)
3. IEMOCAP (Busso et al., 2008)
4. SEMANIE (Mckeown et al., 2010)
5. AFEW (Dhall et al., 2012)
6. eNTREFACE'05 (Martin et al., 2006)
7. ...

Methods

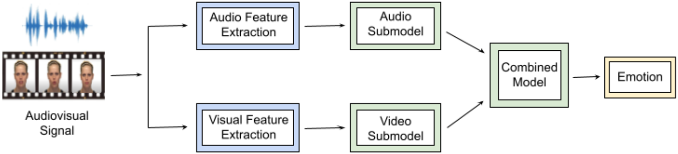


Audiovisual Signal
Audio Feature Extraction
Visual Feature Extraction
Audio Model
Video Model
Decision Rule
Emotion



Audiovisual Signal
Audio Feature Extraction
Visual Feature Extraction
Model
Emotion

Methods


Audiovisual Signal
Audio Feature Extraction
Visual Feature Extraction
Audio Submodel
Video Submodel
Combined Model
Emotion

Major contributions

1. Use hybrid method for modality fusion on the raw data (e.g., audio and pictures from video) to be able to use existing whole content
2. Train and test sets separation based on the speakers (this is important as models tend to overfit to speakers and so the generalization error will be high in this cases)
3. Use mixture of different datasets with augmentation of real world noise in order to provide robustness

Major contributions

1. Use hybrid method for modality fusion on the raw data (e.g., audio and pictures from video) to be able to use existing whole content
2. Train and test sets separation based on the speakers (this is important as models tend to overfit to speakers and so the generalization error will be high in this cases)
3. Use mixture of different datasets with augmentation of real world noise in order to provide robustness

Major contributions

1. Use hybrid method for modality fusion on the raw data (e.g., audio and pictures from video) to be able to use existing whole content
2. Train and test sets separation based on the speakers (this is important as models tend to overfit to speakers and so the generalization error will be high in this cases)
3. Use mixture of different datasets with augmentation of real world noise in order to provide robustness

Emotions and Datasets

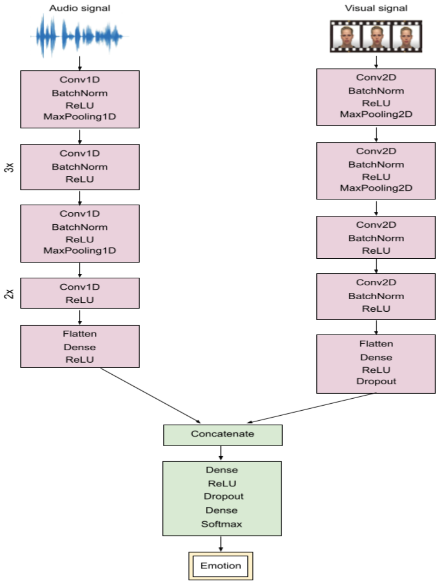
Eemotions

1. Happy
2. Angry
3. Sad
4. Neutral

Datasets

1. RAVDESS (Livingstone and Russo, 2018)
2. IEMOCAP (Busso et al., 2008)
3. eNTREFACE'05 (Martin et al., 2006)

Proposed Architecture

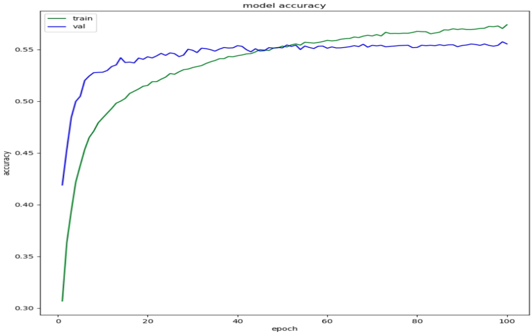


Results Random Split

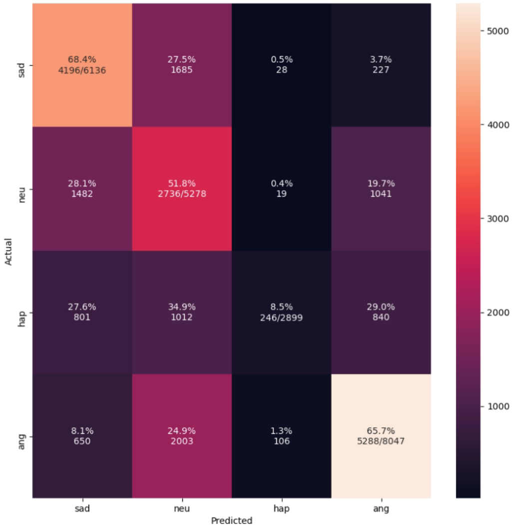
models	Accuracy random.s.	Accuracy speaker s.
Baseline	54	-
Lightgbm Audio	**87.7**	**57.6**

Results Speaker Split

models	Accuracy
Baseline	54
Audio model	54
Video model	**57.7**



Confusion Matrix



Demo

Videos: Link
Source code: Link

Thank you!

References

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Dhall, A., Goecke, R., Lucey, S., and Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41.

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The enterface'05 audio-visual emotion database. *Data Engineering Workshops, 2006. Proceedings*, pages 8 – 8.

Mckeown, G., Valstar, M., Cowie, R., and Pantic, M. (2010). The semaine corpus of emotionally coloured character interactions. *2010 IEEE International Conference on Multimedia and Expo, ICME 2010*, pages 1079–1084.

Philip, J. and Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Precision, Recall, F1score

-	precision	recall	f1-score	support
sad	0.60	0.68	0.63	6136
neu	0.36	0.37	0.37	5278
hap	0.40	0.22	0.28	2899
ang	0.68	0.71	0.69	8047
accuracy			0.56	22360

Features

Audio

1. ZRC
2. RMS
3. Mel Spectrogram
4. MFCC
5. Chromagram

Video

1. 20 faces of (50, 50) size