Vieru Tudor-Gabriel

937/2

Emotion Recognition in Facial and Vocal Patterns

AMS Classification : 68T10 Pattern Recognition, speech recognition

ACM Classification : I.2.0 Artificial Intelligence - General – Cognitive simulation

2. Related Work

The problem of identifying emotions in Facial and Vocal Patterns is well-known and many approaches have been proposed, having different implementations and obtaining small yet important results.

The input data, as it may seem from the title, is a Video file, with sound, representing a person saying sentences from which emotions shall be extracted. This has divided previous approaches into two categories : those which try to obtain said emotional features from both video and audio at the same time, and those which separated the initial input into two separate streams, video and audio, treating them in separate networks for better results.

2.1 Combined processing

As it can be observed in [2], the focus is on treating both the video and audio at the same time, having the results aggregated and posting a final prediction.

The proposed network architecture is a Hybrid CNN-RNN and C3D Network for the Video part, and an SVM with Learn Kernel, resulting in an effective yet rather heavy model. However, the performance is improved by using A Long Short Term Memory (LSTM) Network as the Recurrent Neural Network (RNN). The paper in question uses AFEW 6.0 Database as part of the EmotiW 2016 Challenge.

2.2 Separate processing

Since processing both the audio and the video file at the concurrently may require more computing power, some of the papers tackling the same problem have tried separating these tasks into two separate networks, train them individually and then interpret the results from both of the networks.

As [9] proposes, one approach to tackle audio recognition within two paths : one is to extract Audio Features such as Pitch, Log-Energy, Zero crossing Rate and TEO, then pass them through an Audio Feature-Level Fusion and the other one is to interpret the MFCC (Mel Frequency Cepstral Coefficients) and split them in 3 emotion groups, then pass the results in an aggregating Audio Decision-level Fusion, resulting in one single emotion.

3. Proposed Solution

This paper will focus on separating the input, separating a video passed as input into two channels, one video and one audio, process them separately through two different networks and each of them returning a set of probabilities at a given timespan. The timespan duration will be one second, based on the words spoken per second by the average person, which is 2 or 3.

Having this said, frames within a second shall provide enough data to indicate a single emotion. If there are discrepancies between those two, they shall be flagged to be further interpreted.

3.1 Audio processing

For the audio processing, the input audio stream will be converted into a spectrogram in the range 0 – 16384 Hz. The frequencies that can be heard by the human ear are in the range 0 – 20000Hz, however, on average, human voices have range between 125Hz and 8000Hz.

Due to the fact that differences between emotions are fine, the frequencies generated by an audio clip in the dataset (Fig. 1) have been mapped in the full audible spectrum, so these differences could be easily detected.

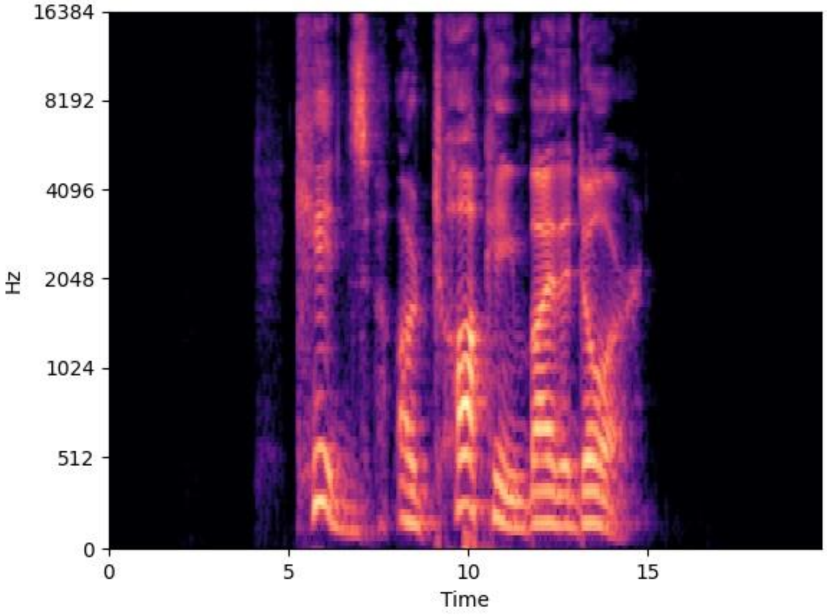

Fig 1. Spectrogram generated by a person displaying an angry emotion

Spectrograms as presented in the previous figure will be passed in a Network that will have the starting point on a pre-trained model, EfficientNetB4 [16], known to provide better results with a smaller model compared to other similar networks, especially when tackling audio classification.

On top of that, the preliminary results will be further processed on a custom network.

3.2 Video processing

For the video processing part, the proposed solution is set to follow a similar approach as in [2] or [3], due to the fact that as opposed to audio streams, videos cannot be compressed to a static form, such as a spectrogram, and they need continuous processing, as provided by a Recurrent Neural Network (RNN). On top of that, classification will be performed on the video frames, therefore, a Convolutional Neural Network (CNN) shall be used.

4. Dataset and data processing

The dataset used for the training of this model is The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [15], which is composed of 24 actors saying two sentences, in multiple takes and with one of the 8 primary emotions (Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust and Surprised), in different modalities : audio-video, video-only or audio-only.

For the audio processing the sound clips have been converted to spectrograms by first extracting the preliminary spectrogram and then mapping accordingly to the dB identified in the clip. After that, the resulting spectrograms are saved as image files, so they can be further transformed and processed into the network.

Bibliography

[1] Kathleen, M. “Toronto Emotional Speech Set (TESS) - University of Toronto Dataverse.” Scholars Portal Dataverse, 13 Feb. 2020, doi:10.5683/SP2/E8H2MF.

[2] Fan, Yin, et al. “Video-Based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks.” *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ACM, Oct. 2016. *Crossref*, doi:10.1145/2993148.2997632.

[3] Ebrahimi Kahou, Samira, et al. “Recurrent Neural Networks for Emotion Recognition in Video.” Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, Nov. 2015. Crossref, doi:10.1145/2818346.2830596.

[4] Kahou, Samira Ebrahimi, et al. “EmoNets: Multimodal Deep Learning Approaches for Emotion Recognition in Video.” Journal on Multimodal User Interfaces, no. 2, Springer Science and Business Media LLC, Aug. 2015, pp. 99–111. Crossref, doi:10.1007/s12193-015-0195-2.

[5] Davletcharova, Assel, et al. “Detection and Analysis of Emotion from Speech Signals.” Procedia Computer Science, Elsevier BV, 2015, pp. 91–96. Crossref, doi:10.1016/j.procs.2015.08.032.

[6] Mingli Song, et al. “Audio-Visual Based Emotion Recognition-a New Approach.” Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., IEEE. Crossref, doi:10.1109/cvpr.2004.1315276.

[7] Livingstone, Steven R., and Frank A. Russo. “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, edited by Joseph Najbauer, no. 5, Public Library of Science (PLoS), May 2018, p. e0196391. Crossref, doi:10.1371/journal.pone.0196391.

[8] Liu, Chuanhe, et al. “Group Level Audio-Video Emotion Recognition Using Hybrid Networks.” Proceedings of the 2020 International Conference on Multim„odal Interaction, ACM, Oct. 2020. Crossref, doi:10.1145/3382507.3417968.

[9] Ooi, Chien Shing, et al. “A New Approach of Audio Emotion Recognition.” Expert Systems with Applications, no. 13, Elsevier BV, Oct. 2014, pp. 5858–69. Crossref, doi:10.1016/j.eswa.2014.03.026.

[10] Haddad, Jad, et al. “3D-CNN for Facial Emotion Recognition in Videos.” Advances in Visual Computing, Springer International Publishing, 2020, pp. 298–309, http://dx.doi.org/10.1007/978-3-030-64559-5_23.

[11] Chew, Li Wern, et al. “Audio-Emotion Recognition System Using Parallel Classifiers and Audio Feature Analyzer.” 2011 Third International Conference on Computational Intelligence, Modelling & Simulation, IEEE, Sept. 2011. Crossref, doi:10.1109/cimsim.2011.44.

[12] Sinith, M. S., et al. “Emotion Recognition from Audio Signals Using Support Vector Machine.” 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), IEEE, Dec. 2015. Crossref, doi:10.1109/raics.2015.7488403.

[13] Jannat, Rahatul, et al. “Ubiquitous Emotion Recognition Using Audio and Video Data.” Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, ACM, Oct. 2018. Crossref, doi:10.1145/3267305.3267689.

[14] Hossain, M. Shamim, and Ghulam Muhammad. “Emotion Recognition Using Deep Learning Approach from Audio–Visual Emotional Big Data.” Information Fusion, Elsevier BV, Sept. 2019, pp. 69–78. Crossref, doi:10.1016/j.inffus.2018.09.008.

[15] Livingstone, S. R., & 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. In J. Najbauer (Ed.), PLOS ONE (Vol. 13, Issue 5, p. e0196391). Public Library of Science (PLoS). doi:10.1371/journal.pone.0196391

[16] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.