Audio–Visual Emotion Recognition

Subjects: Computer Science, Artificial Intelligence

Contributor: Anton Matveev , Yuri Matveev , Olga Frolova , Aleksandr Nikolaev , Elena Lyakso

Emotion recognition can be formulated as a problem where some source produces several streams of data (features) of various modalities (e.g., audio and video), each with its own distribution, and the goal is to estimate the distributions and map them onto the source.

audio–visual speech emotion recognition children

1. Introduction

Emotions play an important role in a person’s life from its very beginning to the end. Understanding emotions becomes indispensable for people’s daily activities, in organizing adaptive behavior and determining the functional state of the organism, in human–computer interaction (HCI), etc. In order to provide natural and user-adaptable interaction, HCI systems need to recognize a person’s emotions automatically. In the last ten to twenty years, improving speech emotion recognition has been seen as a key factor in improving the performance of HCI systems. While most research has focused on emotion recognition in adult speech [1][2], significantly less research has focused on emotion recognition in children’s speech [3][4]. That is because large corpora of children’s speech, especially audio–visual speech, are still not publicly available, and this forces researchers to focus on emotion recognition in adult speech. Nevertheless, children are potentially the largest class of users of most HCI applications, especially in education and entertainment (edutainment) [5]. Therefore, it is important to understand how emotions are expressed by children and whether they can be automatically recognized.

Creating automatic emotion recognition systems in a person’s speech is not trivial, especially considering the differences in acoustic features for different genders [6], age groups [7], languages [6][8], cultures [9], and developmental [10] features. For example, in [11], it is reported that the accuracies of speech emotion recognition are “93.3%, 89.4%, and 83.3% for male, female and child utterances respectively”. The lower accuracy of emotion recognition in children’s speech may be due to the fact that children interact with the computer differently than adults, as they are still in the process of learning social and conversational interaction linguistic rules. It is highlighted in [12] that the main aim of emotion recognition in conversation (ERC) systems is to correctly identify the emotions in the speakers’ utterances during the conversation. ERC helps to understand the emotions and intentions of users and to develop engaging, interactive, and empathetic HCI systems. The input data for a multimodal ERC is information from different modalities for each utterance, such as audio–visual speech and facial expressions, and the model leverages these data to generate accurate predictions of emotions for each utterance. In [13], it was found that in the case of audio–visual recognition of emotions in voice, speech (text), and facial expressions, the facial modality provides recognition of 55% of emotional content, the voice modality provides

38%, and the textual modality provides the remaining 7%. The last is the motivation to use audio–visual speech emotion recognition.

There are few studies on multimodal emotion recognition in children, and even fewer studies have been performed on automatic children's audio–visual emotion recognition. Due to the small size of the available datasets, the main approach was to use traditional machine learning (ML) techniques. The authors of [14] mentioned the following most popular ML-based classifiers: Support Vector Machine, Gaussian Mixture Model, Random Forest, K-Nearest Neighbors, and Artificial Neural Network, with the Support Vector Machine (SVM) classifier being employed in the majority of ML-based affective computing tasks. Recently, there has been a growing focus on automatic methods of emotion recognition in audio–visual speech. This is primarily driven by advancements in machine learning and Deep Learning [15], due to the presence of publicly available datasets of emotional audio–visual speech, and the availability of powerful computing resources [16].

2. Audio–Visual Emotion Recognition

Emotion recognition can be formulated as a problem where some source produces several streams of data (features) of various modalities (e.g., audio and video), each with its own distribution, and the goal is to estimate the distributions and map them onto the source. That, naturally, poses several questions that ought to be answered when building an emotion recognition system: which modalities are selected and represented, how the modalities are mapped on each other, and how the joint representations are mapped onto the sources of the distributions.

It has been shown that regardless of the model and representations, multimodal approaches virtually always outperform unimodal ones [17], i.e., adding another modality can only benefit the performance. While this may seem obvious, the notion actually relies on the fact that, in the worst-case scenario, a model is able to learn an identity mapping for the driving modality and disregard the other one. However, as has been shown in practice, it is rarely the case that additional modalities carry no valuable information. As for the selection of modalities, the most common ones in the literature are images (or sequences of images, i.e., video), audio, and text.

Representation is one of the key concepts in machine learning [18]. While the task of machine learning imposes a number of limitations on the representations of data, such as smoothness, temporal and spatial coherence, over the years, a bevy of various representations have been used to solve various machine learning problems, and while some are more common than the other, there is no clear rule for choosing the best representation. Traditional machine learning algorithms rely on the representation of the input being a feature and learn a classifier on top of that [19]. Meanwhile, the most agile modern models attempt to learn not only the representations but also the architecture and the hyperparameters of the model [20]. Both extremes, however, have several issues. The traditional approach lacks the capability to discover deep, latent features and is mostly unable to achieve high efficiency associated with learning hierarchical and spatial-temporal relationships within feature sets, and since there is no space to learn cross-modal relationships, multimodal models either rely on some sort of decision-level fusion or expert heuristics for joint representations. The end-to-end approach, on the other hand, has a high computational cost and requires a precise, structured approach to training [21]. With those limitations, most of the modern models take reasonably preprocessed input data, then attempt to learn their efficient representations, including joint representations, and finally learn to classify those representations.

There are several ways to present audio data to a model. The most common include [22]:

- Waveform/raw audio, seldom used outside of end-to-end models, is simply raw data, meaning the model has to learn efficient representations from scratch;

- Acoustic features such as energy, pitch, loudness, zero-crossing rate, often utilized in traditional models, while allowing for simple and compact models, are mostly independent by design and prevent a model from learning additional latent features;

- A spectrogram or a mel-spectrogram, which shares some similar issues with raw audio, has found its way into many models due to extensive research into convolutional neural networks, since, being presented as an image, it enables learning efficient representations as shown in various practical applications;

- Mel-Frequency Cepstral Coefficients, which represent the short-term power spectrum of a sound—very commonly used as they provide a compact but informative representation.

In [23], a relatively recent example of representation learning was proposed—a large-scale self-supervised pre-trained WavLM model for speech processing. This model, which is a transformer encoder, efficiently encodes audio features for classification and is trained on a large dataset. The frozen encoder can then be utilized as a feature extractor for general purpose speech processing.

For image processing, the traditional approaches are extremely computationally expensive. For example, when a raw image is processed through a fully connected neural network, the network has to treat each pixel as an individual input and learn to extract relevant features from all locations within the image. In contrast, a convolutional neural network (CNN) [24] can learn to recognize patterns in an image regardless of where they are located, using shared weights across the entire image and reducing the number of parameters required. By design, CNNs learn hierarchical representations of the raw input data and, due to the shown efficiency of this approach, this is the most common approach for the representation of visual data. However, while a static image is a common input for a variety of computed vision problems, there is also a large field of problems concerned with sequences of images, i.e., video. Since, for most of the practical tasks, there are strong relationships between consecutive frames of the input video. It is natural that efficient representations of those relationships are key for achieving high performance. For example, optical flow is a technique used in computer vision that enables one to recognize and track movement patterns in video footage [25]. Another option to employ an implementation of a recurrent neural network (RNN), for example a long short-term memory (LSTM) network or a convolutional RNN, in which case a network is able to collect global context and produce representations enhanced with those shared latent features [26]. Another relatively recent approach is to implement a 3D CNN [27], where the temporal dimension is added to both the input tensor and the filters. While the idea of considering a sequence of images as just another dimension of the input tensor is relatively natural, the significant increase in the number of weights presents the need for a large amount of training video data and incurs a high computational cost. However, as the CNN architectures for image processing became highly optimized and somewhat larger video datasets have become available, this approach became legitimately viable.

The key concept for multimodal classification is the fusion of modalities. Though earlier models relied on unimodal classification and consecutive ensemble learning for decision-level fusion such as averaging, voting, and weighted sum, it was quickly discovered that both the redundancy of features between modalities and latent cross-modal relationships can be utilized to achieve higher performance [28][29]. Another traditional approach is to implement an early fusion. While some of the works propose the fusion of modalities at the input data level [30], the most common approach is to combine modalities upon feature extraction, relying on some sort of heuristics [28][29]. In modern research, fusion is applied somewhere between the feature extraction and the decision level with the goal of learning efficient joint representations to both eliminate the redundancy in order to reduce the computational cost, and to align modalities to take advantage of cross-modal relationships.

There are several strategies for this kind of intermediate fusion, but the most common technique is to implement fusion via an attention mechanism [16]. This is a method to focus on the most relevant information from each modality, to determine which parts of each modality’s input should be given greater focus when making a prediction, and selecting the most important features from each modality and combining them in a meaningful way. In a more general sense, the attention technique can be understood from the distinction between soft and hard attention. To emulate human perception and reduce computations, ideally, a model should be able to ignore the clutter in the input data and attend only to the meaningful parts [31] sequentially and aggregate information over time—this approach would implement so-called hard attention. However, to achieve that, it would require the model to make choices where to look at and they are difficult to represent as differentiable functions which would be required for the most conventional techniques for training. Requiring a model to be differentiable means that the model is simply able to associate more importance with certain parts of the input data—this approach is called soft attention.

Another informative way to designate attention techniques is to focus on the dimensions across which they are applied. Though some terminology may be used interchangeably in the literature, the more common ones include:

- Channel attention—as channels of feature maps are often considered feature detectors, it attempts to select more relevant features for the task [32];

- Spatial attention—in the cases with multidimensional input data such as images, it attends to inter-spatial relationship of features [33];

- Temporal attention—though the temporal dimension can sometimes be considered simply as another dimension of input data, in practice it might be beneficial to view it separately and apply different logic to it, depending on the task [33];

- Cross-attention—mostly utilized in the cases with multiple modalities to learn relationships between modalities; since different modalities often have different dimensions, the modalities cannot be viewed as just another dimension of the input tensor, thus requiring a different approach from simply increasing the dimension of the attention maps; can be used to combine information from different modalities, in which case it is said to implement the fusion of modalities [34].

The authors of [35] suggested that applying attention along the input dimensions separately achieves lower computational and parameter overhead compared to computing attention maps with the same dimensions as the input. The authors of [36] proposed the “Squeeze-and-Excitation” block, an architectural unit that explicitly models interdependencies between channels and recalibrates feature maps channel-wise. The authors of [37] presented a self-attention mechanism for CNN to capture long-range interactions between features, which, in modern research, is mostly applied to sequence modeling and generative modeling tasks, they show that they can improve the performance of a model by increasing the number of feature maps by concatenating the feature maps with multihead attention maps. The authors of [38] implemented cross-attention for multimodal emotion recognition from audio and text modalities where the features from the audio encoder attend to the features from the text encoder and vice versa to highlight the most relevant features for emotion recognition. Though the features from those two modalities are eventually concatenated before passing them to the classifier, the attention block does not explicitly implement a fusion of modalities and is rather an example of late fusion. The authors of [39] proposed a universal split-attention block for the fusion of modalities where the attention block explicitly fuses features from different modalities and can be both placed at an arbitrary stage of a network and repeated multiple times across the network.

After the feature maps are generated by a network, the final step is to classify the sample into one of the target categories. The most common approach is to map the feature maps onto scalar values (flatten the feature maps) and present the output as a scalar vector so that it can be presented to a fully connected network which is trained to classify the input into one of the target categories, usually by a SoftMax layer with the number of neurons equal to the number of target classes [19]. Even though this approach is utilized in most of the modern models, flattening of the feature maps effectively discards the spatial and temporal relationships. To investigate some of those issues, the authors of [40] suggested generating so-called “class activation maps”, where the class activation map points to the segments of the input image which the network considers discriminative to detect the target class. Since the outcome of this procedure can encapsulate the spatial and temporal relationships between the input and the feature maps, this information can also be employed for classification.

References

1. Schuller, B.W. Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends. Commun. ACM 2018, 61, 90–99.

2. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access 2019, 7, 117327–117345.
3. Lyakso, E.; Ruban, N.; Frolova, O.; Gorodnyi, V.; Matveev, Y. Approbation of a method for studying the reflection of emotional state in children’s speech and pilot psychophysiological experimental data. Int. J. Adv. Trends Comput. Sci. Eng. 2020, 9, 649–656.
4. Onwujekwe, D. Using Deep Leaning-Based Framework for Child Speech Emotion Recognition. Ph.D. Thesis, Virginia Commonwealth University, Richmond, VA, USA, 2021. Available online: https://scholarscompass.vcu.edu/cgi/viewcontent.cgi?article=7859&context=etd (accessed on 20 March 2023).
5. Guran, A.-M.; Cojocar, G.-S.; Diosan, L.-S. The Next Generation of Edutainment Applications for Young Children—A Proposal. Mathematics 2022, 10, 645.
6. Costantini, G.; Parada-Cabaleiro, E.; Casali, D.; Cesarini, V. The Emotion Probe: On the Universality of Cross-Linguistic and Cross-Gender Speech Emotion Recognition via Machine Learning. Sensors 2022, 22, 2461.
7. Palo, H.K.; Mohanty, M.N.; Chandra, M. Speech Emotion Analysis of Different Age Groups Using Clustering Techniques. Int. J. Inf. Retr. Res. 2018, 8, 69–85.
8. Tamulevičius, G.; Korvel, G.; Yayak, A.B.; Treigys, P.; Bernatavičienė, J.; Kostek, B. A Study of Cross-Linguistic Speech Emotion Recognition Based on 2D Feature Spaces. Electronics 2020, 9, 1725.
9. Lyakso, E.; Ruban, N.; Frolova, O.; Mekala, M.A. The children’s emotional speech recognition by adults: Cross-cultural study on Russian and Tamil language. PLoS ONE 2023, 18, e0272837.
10. Matveev, Y.; Matveev, A.; Frolova, O.; Lyakso, E. Automatic Recognition of the Psychoneurological State of Children: Autism Spectrum Disorders, Down Syndrome, Typical Development. Lect. Notes Comput. Sci. 2021, 12997, 417–425.
11. Duville, M.M.; Alonso-Valerdi, L.M.; Ibarra-Zarate, D.I. Mexican Emotional Speech Database Based on Semantic, Frequency, Familiarity, Concreteness, and Cultural Shaping of Affective Prosody. Data 2021, 6, 130.
12. Zou, S.H.; Huang, X.; Shen, X.D.; Liu, H. Improving multimodal fusion with Main Modal Transformer for emotion recognition in conversation. Knowl.-Based Syst. 2022, 258, 109978.
13. Mehrabian, A.; Ferris, S.R. Inference of attitudes from nonverbal communication in two channels. J. Consult. Psychol. 1967, 31, 248–252.
14. Afzal, S.; Khan, H.A.; Khan, I.U.; Piran, J.; Lee, J.W. A Comprehensive Survey on Affective Computing; Challenges, Trends, Applications, and Future Directions. arXiv 2023, arXiv:2305.07665v1.

15. Dresvyanskiy, D.; Ryumina, E.; Kaya, H.; Markitantov, M.; Karpov, A.; Minker, W. End-to-End Modeling and Transfer Learning for Audiovisual Emotion Recognition in-the-Wild. Multimodal Technol. Interact. 2022, 6, 11.
16. Wang, Y.; Song, W.; Tao, W.; Liotta, A.; Yang, D.; Li, X.; Gao, S.; Sun, Y.; Ge, W.; Zhang, W.; et al. A systematic review on affective computing: Emotion models, databases, and recent advances. Inf. Fusion 2022, 83–84, 19–52.
17. Chiara, Z.; Calabrese, B.; Cannataro, M. Emotion Mining: From Unimodal to Multimodal Approaches. Lect. Notes Comput. Sci. 2021, 12339, 143–158.
18. Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 8, 1798–1828.
19. Burkov, A. The Hundred-Page Machine Learning Book; Andriy Burkov: Quebec City, QC, Canada, 2019; 141p.
20. Egele, R.; Chang, T.; Sun, Y.; Vishwanath, V.; Balaprakash, P. Parallel Multi-Objective Hyperparameter Optimization with Uniform Normalization and Bounded Objectives. arXiv 2023, arXiv:2309.14936.
21. Glasmachers, T. Limits of End-to-End Learning. In Proceedings of the Asian Conference on Machine Learning (ACML), Seoul, Republic of Korea, 15–17 November 2017; pp. 17–32. Available online: https://proceedings.mlr.press/v77/glasmachers17a/glasmachers17a.pdf (accessed on 28 October 2023).
22. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. Sensors 2021, 21, 1249.
23. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. IEEE J. Sel. Top. Signal Process. 2022, 16, 1505–1518.
24. Alexeev, A.; Matveev, Y.; Matveev, A.; Pavlenko, D. Residual Learning for FC Kernels of Convolutional Network. Lect. Notes Comput. Sci. 2019, 11728, 361–372.
25. Fischer, P.; Dosovitskiy, A.; Ilg, E.; Häusser, P.; Hazırbaş, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning Optical Flow with Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile; 2015; pp. 2758–2766.
26. Patil, P.; Pawar, V.; Pawar, Y.; Pisal, S. Video Content Classification using Deep Learning. arXiv 2021, arXiv:2111.13813.
27. Hara, K.; Kataoka, H.; Satoh, Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 6546–6555.

28. Wu, C.; Lin, J.; Wei, W. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. APSIPA Trans. Signal Inf. Process. 2014, 3, E12.
29. Karani, R.; Desai, S. Review on Multimodal Fusion Techniques for Human Emotion Recognition. Int. J. Adv. Comput. Sci. Appl. 2022, 13, 287–296.
30. Ordóñez, F.J.; Roggen, D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. Sensors 2016, 16, 115.
31. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent Models of Visual Attention. In Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), Montreal, QC, Canada, 8–13 December 2014; Volume 2, pp. 2204–2212. Available online: https://proceedings.neurips.cc/paper_files/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf (accessed on 28 October 2023).
32. Hafiz, A.M.; Parah, S.A.; Bhat, R.U.A. Attention mechanisms and deep learning for machine vision: A survey of the state of the art. arXiv 2021, arXiv:2106.07550.
33. Bertasius, G.; Wang, H.; Torresani, L. Is Space-Time Attention All You Need for Video Understanding? arXiv 2021, arXiv:2102.05095.
34. Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; Wu, F. Multi-Modality Cross Attention Network for Image and Sentence Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10938–10947.
35. Woo, S.; Park, J.; Lee, J.-L.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the 15th European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Part VII; pp. 3–19.
36. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
37. Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention Augmented Convolutional Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3285–3294.
38. Krishna, D.N.; Patil, A. Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. In Proceedings of the 21th Annual Conference of the International Speech Communication Association (INTERSPEECH), Shanghai, China, 25–29 October 2020; pp. 4243–4247.

39. Lang, S.; Hu, C.; Li, G.; Cao, D. MSAF: Multimodal Split Attention Fusion. arXiv 2021, arXiv:2012.07175.
40. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2921–2929.

Retrieved from https://encyclopedia.pub/entry/history/show/117091