STATISTICAL, SPECTRAL AND GRAPH REPRESENTATIONS FOR VIDEO-BASED FACIAL EXPRESSION RECOGNITION IN CHILDREN

Nida Itrat Abbasi[†] [*], *Siyang Song*[†], *and Hatice Gunes*

Department of Computer Science and Technology, University of Cambridge, United Kingdom.

ABSTRACT

Child facial expression recognition is a relatively less investigated area within affective computing. Children's facial expressions differ significantly from adults; thus, it is necessary to develop emotion recognition frameworks that are more objective, descriptive and specific to this target user group. In this paper we propose the first approach that (i) constructs video-level heterogeneous graph representation for facial expression recognition in children, and (ii) predicts children's facial expressions using the automatically detected Action Units (AUs). To this aim, we construct three separate length-independent representations, namely, statistical, spectral and graph at video-level for detailed multi-level facial behaviour decoding (AU activation status, AU temporal dynamics and spatio-temporal AU activation patterns, respectively). Our experimental results on the LIRIS Children Spontaneous Facial Expression Video Database demonstrate that combining these three feature representations provides the highest accuracy for expression recognition in children.

Index Terms— Affect Recognition, Child Facial Expressions, Heterogeneous Graph Representation, Deep Learning

1. INTRODUCTION

The human face is a rich and reliable source for understanding complex cognitive processes. Therefore, automatic analysis of facial expressions is essential not only in gaining insights into human-human interactions (e.g., depression [1] or pain detection [2] in clinical settings) but also into human-agent and human-robot interactions (e.g., inferring engagement and interest [3, 4]). In particular, accurately recognising facial expressions can provide a deeper insight into human nonverbal behaviours [5], especially for people that do not have advanced verbal communication skills, e.g., children.

A large number of automatic facial expression analysis solutions have been proposed in recent years [6, 7, 8]. However, several works have shown that models trained on adult expression datasets do not generalize well on child facial expression recognition tasks [4] as children and adults vary significantly in how they display expressions of emotions, leading to different manifestations in facial expressions [9]. For example, facial expressions of children are often exaggerated, incomplete and unique as compared with their adult counterparts [9, 10]. Moreover, child facial expressions are very dependent on demographic factors such as gender, age and ethnicity [11]. While a wide-range of machine learning-based facial expression classification approaches have been proposed for analysing and detecting adult facial expressions, very few studies [9, 11, 12] have specifically investigated facial expression analysis in children.

Predominantly, child facial expression analysis studies have used raw images or videos as input for expression recognition models [9, 11]. However, utilising behaviour primitives like facial action units (AUs) [13] have several advantages over using raw images / videos. Facial AUs provide (i) an objective and descriptive representation for measuring facial behavioural changes [14]; and (ii) an ethical advantage over using raw images and videos for vulnerable demographic groups including children and the elderly by discarding video data after AU detection in real-time, and hence addressing issues related to confidentiality and anonymity [15, 16].

In this work, we create an AU-based methodology for facial expression recognition in children. To this aim, we propose the construction of three length-independent representations: statistical, spectral and heterogeneous graph representation, to encode the video-level facial statistics, temporal dynamics as well as facial expression-related spatio-temporal activation patterns, from the AU time-series of the target video. To the best of our knowledge, we propose the first approach that constructs video-level heterogeneous graph representation for child facial expression recognition. Also, to the best of our knowledge, this is the first work that predicts children's facial expressions from their automatically detected AUs. Our experimental results utilising multilayer perceptron (MLP) and Bidirectional-Long Short-Term Memory (Bi-LSTM) neural networks show that using a combination of feature representations (statistical, spectral and graph) provide the highest accuracy for expression recognition. The pipeline of our approach is illustrated in Fig. 1.

2. METHODOLOGY

This section describes the details of the methodology including the extraction of AUs, the three types of representation,

[†] contributed equally, [*]corresponding author email: nia22@cam.ac.uk

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

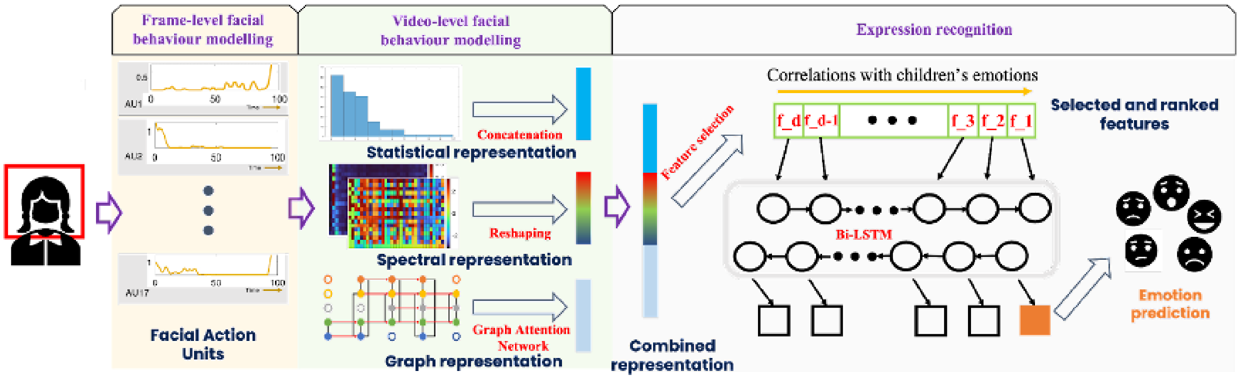

Fig. 1. The proposed pipeline for child facial expression recognition.

children-specific feature selection and the deep learning-based expression recognition.

2.1. Frame-level representation

In a given video, facial AU detection was obtained using the OpenFace 2.0 toolkit [17]. OpenFace provides AU intensity and occurrence information for each frame of the input video in terms of 17 facial AUs (AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26 and AU45). This provided us a 34-channel time series (34 × video duration) that was utilised for further processing of the video level dynamics.

2.2. Video-level representation

Since the goal is to predict the category of facial expression of emotion in children from their facial videos, we propose to encode all frame-level facial descriptors in a video of an arbitrary length into three types of length-independent representations, summarising three types of facial patterns (static AU displays, multi-scale temporal evolution of AUs, and local spatio-temporal AU activation patterns) for video-level expression recognition (Fig. 1).

Statistical representation: We first calculate 24 statistics to summarise each AU time-series, i.e., 12 statistics to represent its intensity information and 12 statistics to represent its occurrence information [18]. For an AU intensity/occurrence time-series, the first four statistics are the mean, standard deviation, median and maximum of the original time-series. Then, we compute the first order derivatives utilising these four values as well as the the second order derivatives of the original time-series. Finally, we concatenate the statistics of all AUs as a 1D video-level facial statistical representation which results in a feature vector with dimensionality = 24.

Spectral representation: A main drawback of the statistical representation is that it ignores the temporal dynamics of AUs (i.e., the evolution of an AU over time) , which is a key element for facial behaviour analysis, and thus crucial for recognizing and distinguishing facial expressions. To address this, we introduce the spectral representation [18, 19] to summarize multi-scale video-level facial dynamics. For an AU, this approach firstly converts its time-series to a spectral signal using Fourier Transform, where each component in the spectral signal corresponds to the intensity of a unique video-level frequency (i.e., a unique scale of facial dynamics). For time-series of variable-lengths, the components of their K common frequencies are selected. As a result, each time-series can be represented by a K-D vector that summarise K-scales of video-level facial dynamics. Again, we concatenate the spectral representations of all AUs as a 1D video-level facial spectral representation with dimensionality $= K$.

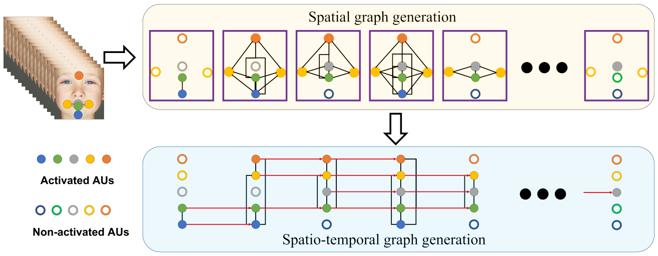

Fig. 2. Illustration of the proposed graph representation. It starts with producing a spatial graph for each frame, i.e., all activated AUs are connected by an undirected edge. Then, directed edges that represent the temporal evolution of every AU are employed to connect adjacent frames, i.e., generating a spatio-temporal graph for the video-level facial behaviours.

Graph representation: While the two representations described above can summarise AU video-level static and dynamic status, they can not encode spatio-temporal details at each time-stamp or thin-slice level. Considering that AUs represent different facial muscle movements, they can be treated as different components of the behaviour of a whole face. To this end, we propose a novel AU graph representation to describe the video-level facial behaviours. As illustrated in Fig.

2, we first construct a spatial graph representing the static AU status of each frame, where each AU is represented as a vertex. In particular, for those activated AUs, their vertice features are denoted as 1 while vertice features of un-activated AUs are 0. In the spatial graph, we only consider these activated AUs, and set them to be mutually connected (undirected edge). After that, we propose to build a spatio-temporal graph by combining spatial graphs of all frames in that video. For an activated AU (AU m) in the n_{th} frame, if the AU m in the $n + 1_{th}$ frame is also activated, they will be connected by a directed edge. This way, the video-level AU spatio-temporal activation details can be encoded by a single graph representation with a unique topology, where the spatial graphs provide the *frame-level AU activation status* and the temporal graphs represent the *AU activation duration and dynamics*. Importantly, the produced spatio-temporal graphs retain all local spatio-temporal details of the AU activation, allowing discriminative local facial patterns to be used for facial expression classification. To further extract the expression-related patterns, we feed the produced video-level heterogeneous representation to the well-trained Graph Attention Network (GAT) [20], and use the output of the first fully connected layer in the GAT as the 1D facial graph representation for each video. In summary, the statistical and spectral representations can summarise video-level AU activation status and their multi-scale temporal dynamics for a target video. The graph representation can also provide facial expression-related local spatio-temporal AU activation patterns. As a result, we first compare and then combine these three representations (which encode multi-level facial behaviour patterns) as the final video-level representation for facial expression classification in children.

2.3. Expression recognition

To utilize the produced multi-level representation for child facial expression classification, we propose a novel Bi-LSTM-based approach. Our Bi-LSTM has one hidden layer with 100 units, and a output layer with softmax activation to provide one-hot expression prediction. Before feeding the combined representation to Bi-LSTM, we compute the correlation between each feature of the representation and children's facial expressions, and then select the top-d features that are most correlated with the children's facial expressions. As a result, the final representation only consists of d features and they are ranked based on their correlations to children's facial expressions in the representation. Then, we treat each feature as an individual step to feed to the Bi-LSTM model. This way, the Bi-LSTM would adjust its forget gate based on the importance of the input features. In comparison to standard classifiers such as Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) that equally consider all features at the input level, the proposed strategy treat these features differently based on their correlations with children's facial expressions. Specifically, the Bi-LSTM model would first learn features with lower correlations and learn the most informative feature at the last step, allowing the model to emphasize most important features and forget less informative information.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

Dataset: In this work, we used the LIRIS Children Spontaneous Facial Expression Video Database that contains facial expression videos of 12 children with varying gender, ethnicity, and age-group (208 videos, total running time of 17 mins 35 secs) [11, 21]. Other existing child facial expression datasets like CAFE [22], Darthmouth [23] and NIMH [24] only contain static images of posed expressions. LIRIS DB instead consists of annotated videos recorded spontaneously using emotional inducers corresponding to six emotion categories (sadness, happiness, disgust, surprise, fear and anger) along with some combined emotion categories (happy-surprise etc). Therefore we utilise the LIRIS DB for our experiments.

Training details: For facial expression classification, we have excluded video clips belonging to the anger class (not sufficient number of clips) and the combined categories, and the clips that have a very short duration (less than 3% of files) from further analysis. We have conducted 12-fold leave-one-child-out cross-validation (at each fold, videos of 11 children were used for training and videos of the remaining child were used for testing). To train the Bi-LSTM and MLP models, cross-entropy was used as the loss function and Adam was used as the optimizer with the learning rate of 0.001 and 0.005, respectively.

3.2. Results

In this work, we trained 12 models for eight video-level representation-classifier combinations following the leave-one-child-out cross-validation strategy. The accuracy obtained by the different models is provided in the Table 1.

Statistical		Spectral		Graph		Combined	
MLP	Bi-LSTM	MLP	Bi-LSTM	MLP	Bi-LSTM	MLP	Bi-LSTM
48.9%	54.4%	62%	57.1%	47.3%	51.1%	66.3%	67.4%

Table 1. Child facial expression classification accuracy obtained with different representations proposed.

Comparison between video-level representations: As seen from Table 1 and Fig. 3, among the three video-level representations, on average spectral representation provides the best results. However, the combination of all three representations provides the highest accuracy (66 % in MLP and 67.4 % in Bi-LSTM). This implies that these representations contain complementary information via detailed multi-level facial behaviour decoding (AU activation status, temporal dynamics

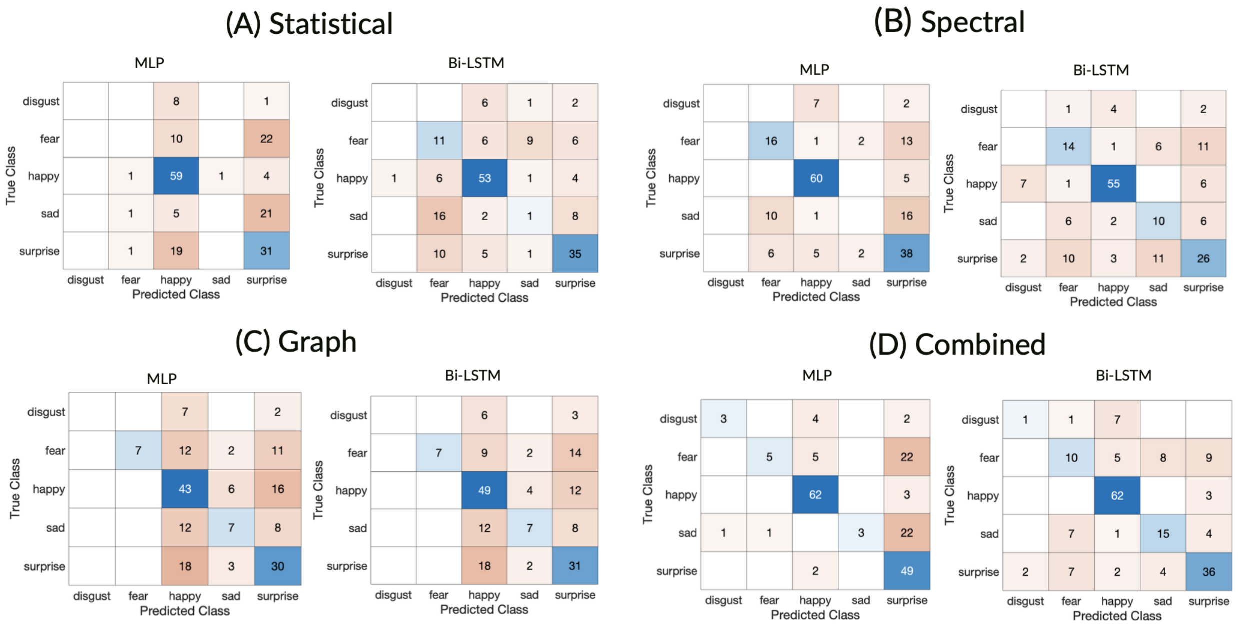

Fig. 3. Confusion matrices for the different facial expression recognition frameworks proposed in this work.

and local spatio-temporal activation patterns). The expression class *sad* is better recognized using the graph representation. Both statistical and spectral representations are informative for recognizing the expression class *happy*. Spectral representation provides the best accuracy for the expression class *fear*. This means that clues that enable the classification of *happy* are contained in both static AU status and dynamic AU information, while for *fear* AU temporal dynamics appear to be more informative.

Comparison between MLP and the proposed Bi-LSTM model: Bi-LSTM outperforms the standard MLP in all representations except in the spectral representation. Since Bi-LSTM makes use of forward and backward information of the feature input at every time-step, it is possible that the temporal dynamics of AUs (as reflected through spectral representation) may not be as informative for the Bi-LSTM as compared to the standard MLP. Moreover, in our experiments, we have used the subject-independent cross-validation protocol which is more challenging than using 80%-10% of frames for training and validation, respectively, as was done by Khan et al [11]. Another predominant advantage of our proposed methodology is that we have facial AUs as feature inputs instead of the raw images extracted from the video clips, making our analysis pipeline more generalisable and objective across target groups with varying demographics.

Comparison between expression classes: We found that the proposed AUs-based representation provide relatively accurate predictions for happy expression class as also observed in [11, 12] and surprise expression class which has also been reported in [11]. It is also seen that the negative expression of disgust has been barely recognised (only in combined representation) and is often classified as happy, while fear has been frequently classified as surprise, showing that these two expressions may have some similar AU patterns in children.

4. CONCLUSIONS

In this paper, we proposed the usage of statistical, spectral and graph representations to encode multi-level facial behaviours for video-based child facial expression recognition, where features are processed and combined by Bi-LSTM based on their correlations with child facial expressions. This work is the first to create video-level heterogeneous graph representation for facial expression recognition in children - a relatively less explored target user group. Our experimental results show that the combination of all three representations using the Bi-LSTM model provides the highest accuracy for child facial expression recognition. Models developed in this work can provide a valuable stepping stone for creating affect recognition frameworks for child-agent interaction research. In future, we aim to use more advanced deep-learning frameworks like gated graph convolutional networks and also compare other state-of-the-art end-to-end network architectures for improving the accuracy of the models proposed in this work.

Acknowledgements: N. I. Abbasi is supported by the W.D. Armstrong Trust Studentship and the Cambridge Trusts. S.Song is partially supported by the European Union's Horizon 2020 Research and Innovation programme under grant agreement No. 826232. H. Gunes is supported by the EPSRC under grant ref. EP/R030782/1.

5. REFERENCES

[1] E. R. Vieira et al., "Depression in older adults: screening and referral," *Journal of Geriatric Physical Therapy*, vol. 37, no. 1, pp. 24–30, 2014.

[2] J. Versloot et al., "Assessment of pain by the child, dentist, and independent observers," *Pediatric dentistry*, vol. 26, no. 5, pp. 445–449, 2004.

[3] N. Churamani et al., "Affect-driven modelling of robot personality for collaborative human-robot interactions," *arXiv preprint arXiv:2010.07221*, 2020.

[4] A. Howard et al., "Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems," in *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*. IEEE, 2017, pp. 1–7.

[5] Z. Liu et al., "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 668–676, 2017.

[6] E. Sariyanidi et al., "Learning bases of activity for facial expression recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1965–1978, 2017.

[7] F. Giroux et al., "Guidelines for collecting automatic facial expression detection data synchronized with a dynamic stimulus in remote moderated user tests," in *International Conference on Human-Computer Interaction*. Springer, 2021, pp. 243–254.

[8] D. K. Jain et al., "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognition Letters*, vol. 139, pp. 157–165, 2020.

[9] M. A. Witherow et al., "Transfer learning approach to multiclass classification of child facial expressions," in *Applications of Machine Learning*. International Society for Optics and Photonics, 2019, vol. 11139, p. 1113911.

[10] P. M. Cole, "Children's spontaneous control of facial expression," *Child development*, pp. 1309–1321, 1986.

[11] R. A. Khan et al., "A novel database of children's spontaneous facial expressions (liris-cse)," *Image and Vision Computing*, vol. 83, pp. 61–69, 2019.

[12] A. Dapogny et al., "On automatically assessing children's facial expressions quality: A study, database, and protocol," *Frontiers in Computer Science*, vol. 1, pp. 5, 2019.

[13] P. Ekman, "Methods for measuring facial action," *Handbook of methods in nonverbal behavior research*, pp. 45–90, 1982.

[14] M. Pantic and L.J.M. Rothkrantz, "Automatic analysis of facial expressions: the state of the art," *IEEE Trans. on PAMI*, vol. 22, no. 12, pp. 1424–1445, 2000.

[15] N. F. Tolksdorf et al., "Ethical considerations of applying robots in kindergarten settings: Towards an approach from a macroperspective," *International Journal of Social Robotics*, vol. 13, no. 2, pp. 129–140, 2021.

[16] S. Robson, "Producing and using video data in the early years: Ethical questions and practical consequences in research with young children," *Children & Society*, vol. 25, no. 3, pp. 179–189, 2011.

[17] T. Baltrusaitis et al., "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018, pp. 59–66.

[18] S. Song et al., "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 158–165.

[19] S. Song et al., "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, 2020.

[20] P. Veličković et al., "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[21] R. Ahmed Khan et al., "Automatic affect analysis: from children to adults," in *International Symposium on Visual Computing*. Springer, 2015, pp. 304–313.

[22] V. LoBue and C. Thrasher, "The child affective facial expression (cafe) set: Validity and reliability from untrained adults," *Frontiers in psychology*, vol. 5, pp. 1532, 2015.

[23] K. A. Dalrymple et al., "The dartmouth database of children's faces: Acquisition and validation of a new face stimulus set," *PloS one*, vol. 8, no. 11, pp. e79131, 2013.

[24] H. L. Egger et al., "The nimh child emotional faces picture set (nimh-chefs): a new set of children's facial emotion stimuli," *International journal of methods in psychiatric research*, vol. 20, no. 3, pp. 145–156, 2011.