A Framework for Automatic Personality Recognition in Dyadic Interactions

Euodia Dodd, Siyang Song and Hatice Gunes
Department of Computer Science and Technology, University of Cambridge, UK

***Abstract*—Research has shown that the way in which an individual interacts with others contains vital cues for recognising their real personality traits. The ability to recognise and adapt to the personality of users is key to developing more intelligent social robots, especially in real-world scenarios. However, most methods for personality recognition focus on apparent personality recognition of individuals in isolated settings. In this work, we propose the first multi-modal framework for human behaviour primitives-based automatic real personality recognition in dyadic interactions. It leverages the use of the spectral representations of behavioural primitives to exploit the temporal nature of the data whilst retaining as much vital information pertaining to personality as possible. We experiment on a range of standard fusion methods to evaluate their effectiveness at combining information from multiple modalities and both interactants in a dyadic interaction. At the multi-subject level, our attention-based fusion approach using a multimodal transformer enabled with cross-subject attention was the most successful. The experimental results show that our approach improved on the previous state-of-the-art on the UDIVA dataset by up to 46%.**

***Index Terms*—Real personality recognition, Behavioural primitives, Dyadic interaction, Multimodal personality recognition, Social robotics**

I. Introduction

From our thought patterns to our behaviours, personality governs various aspects of human experiences and impacts many areas of human lives. Accurately recognising human personality would allow better understanding of different human behaviours and status, e.g. mental health [1]. This is particularly important to the development of intelligent social robots as previous works have shown that incorporating the users' and robot's personalities and interpersonal features affects engagement [2], [3], and perceived enjoyment [4]. A common way for evaluating personality is through the trait models which provide a taxonomy of personality traits. One of the most widely-used model is the "Big-Five" model [5] which groups traits into 5 factors; Openness to experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism. These traits measure aspects of the human personality that have been shown to remain relatively stable over time but differ across individuals. They also generalise across age and gender whilst remaining valid under different methods of testing [6].

Existing automatic personality computing can be divided into two categories [7], [8]: (i) predicting the self-reported personality of an individual (Automatic Personality Recognition (APR)), where self-reported personality traits are typically collected through the use of questionnaires where individuals describe how they see themselves; and (ii) predicting the apparent personality of an individual (Automatic Personality Perception Recognition (APP)), where apparent personality traits reflects how the individual is perceived by others. Though the two may be related, they contain different information about an individual. They are affected by certain biases such as reputation and self-presentation [9] as an individual is likely to answer the questions in a way that maintains the image they wish to portray to others or that they determine to be more socially desirable [10].

Since previous psychological studies [11]–[14] frequently show that personality traits can be reflected by human non-verbal behaviours, most existing personality computing approaches aim to directly recognise apparent personality traits from the target subject's audio [15]–[17], visual [18]–[20] or audio-visual behaviours [21], [22]. There is evidence suggesting that an individual's response to certain situations largely depends on their personalities [23]. For example during dyadic or small group interactions, the interaction between the personalities of the individuals involved has an important impact on the outcome of the interaction style [24]. Despite this, only a few studies [25]–[30] attempted to explore interaction behaviours for self-reported personality traits recognition, all of which are building on raw audio-visual clips. However, some of these approaches only investigated the target subject's behaviours without considering the conversational partner's behaviours for personality recognition, while others usually rely on complex feature-extraction or feature engineering techniques. Moreover, most of these works fail to capture emergent behaviours over varying time-scales (multi-scale temporal dynamics of human behaviours).

Considering that human behaviours especially face-related behaviours vary based on different demographic factors, recent studies [31], [32] show that human behaviour primitives such as facial action units (AUs) [33] can also provide objective, informative, confidential, and anonymous cues [34] for various human behaviour understanding tasks. In this paper, we systematically investigate: (i) the feasibility of applying various human behaviour primitives to automatic self-reported personality traits recognition; (ii) a set of standard fusion strategies for combining audio-visual behaviour primitives for self-reported personality traits recognition; and (iii) a set of standard fusion strategies for combining speaker and listener behavioural cues, and their benefits in recognising subjects' self-reported personality traits. The main contributions of this paper are summarised as follows:

- To the best of our knowledge, this is the first study that applies human behaviour primitives to automatic self-reported personality traits recognition, which achieved more than 46% performance improvements over the state-of-the-art method that directly predicts personality traits from raw audio-visual data.
- We provide the first study that investigated the effectiveness of different fusion strategies that combine audio and visual behaviours of the target subject for self-reported personality recognition under different dyadic interaction scenarios (competitive game-play, presentation, story-telling and collaborative problem-solving).
- We provide the first study that investigated the effectiveness of different fusion strategies that combine multi-modal behaviour primitives of the target subject and the conversational partner for self-reported personality recognition under different dyadic interaction scenarios.

II. Methodology

In this section we present our framework for automatic personality recognition in dyadic interactions. The behaviour primitives are automatically extracted using OpenFace 2.0 [35]. The behaviour primitives in addition to the audio modality are then converted to their spectral representations to capture multi-scale behavioural cues whilst transforming the videos to a fixed-length representation with lower dimensionality. We consider four different fusion strategies to generate video-level predictions for the target participant. Finally, we propose an attention-based fusion approach for capturing both multimodal and multi-subject relationships.

A. Frame-level multi-modal human behaviour primitives extraction

Feature extraction for the visual modality was performed using the OpenFace 2.0 toolkit [35] which automatically detects the presence and intensity of 17 different Facial Action Units (AU), 6 gaze directions per eye, and 6 head pose movements. We then split these features into the AU, gaze and pose modalities. We then normalise the values. We choose to use behaviour primitives as they have been shown to be successful at capturing vital information relating to an individual's state of mind and have demonstrated success when applied to depression recognition and personality recognition [36] [37] [1]. Representations are much lower in dimensionality than raw video data with the primitives used having 35, 8 and 6 dimensions for AU, pose, gaze respectively. The reduced dimensionality resulted in a reduction in the computational demand.

For the audio modality, we extract the audio from the video using FFmpeg, an open-source suite of libraries for handling video and audio files. We extract the audio signal in stereo wave form at 44.1kHz per second.

B. Multi-scale behaviour representation generation

A key challenge presented by time-series data is the arbitrary length of each sample which makes it difficult to use with most standard ML models. To overcome this, the data from each modality is transformed into its spectral representation through the use of Fourier Transform as presented in the work by Song et al. [31]. Each time-series is converted to the frequency domain using the Discrete Fourier Transform where a fixed frequency resolution R is selected such that the frequency components will be a multiple of R allowing k common frequencies to be collected from each signal thus creating a fixed length representation. The amplitude map is computed as:

$$|F_c(w)|/N = \sqrt{Re_c^m(w)^2 + Im_c^m(w)^2/N} \tag{1}$$

while the phase map is computed as:

$$arg(F_c^m(w) = arctan\frac{Im_c^m(w)}{Re_c^m(w)} \tag{2}$$

where $F_c^m(w)$ represents the time-series signal m and w represents any real number. $Re_c^m(w)$ and $Im_c^m(w)$ represent the real and imaginary part of $F_c^m(w)$.

The amplitude heatmap and phase heatmap are concatenated to create the representation used as input. Further details can be found in [31]. By using the spectral representation of the videos, it becomes possible to create a fixed sized input from variable length videos whilst still retaining as much important information related to personality as possible. By doing so, we are also able to retain important temporal dynamics. The final spectral representations were in matrix form with 72, 18, 14 and 4 rows for AU, pose, gaze and audio modalities respectively. We fix $k = 80$ such that all representations have 80 columns.

Previous methods have addressed temporal relationships between frames by dividing videos into chunks of a predetermined time window. Determining the optimal length of the time-window is challenging as it may be dependent on the type of task, dataset and personality trait. Using the spectral representation mitigates these issues by creating multi-scale representations that encode the participant's behaviour throughout the video. A diagram of our method can be seen in Figure 1.

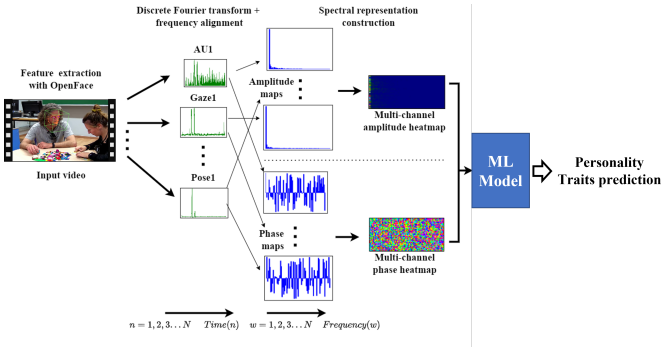

Fig. 1: Diagram of preprocessing method used to achieve spectral representations from raw input (adapted from the paper by Song et al. [31])

C. Multi-modal fusion framework for true personality recognition

We extract frame-level behaviour primitives and audio data which we then transform into their spectral representations. We then fuse these representations to generate video-level descriptions of a target individual. Our proposed framework is depicted below in Fig. 2. Specifically, we evaluate the following fusion methods (illustrated in Fig. 3):

- **Feature-level fusion**: Features are concatenated prior to training to create a combined input which is fed to the model. Feature-level fusion, also known as early fusion, has the advantage of being able to learn low-level relationships between modalities whilst only requiring one model [38]. It is the simplest of the fusion models to implement.
- **Decision-level fusion with averaging**: Models for each modality are first trained independently. The individual predictions are then averaged at the end to produce a single set of predictions. Decision-level fusion or late fusion allows for more flexibility as the different models can learn individually but ignores the lower-level correlations between modalities [38].
- **Decision-level fusion with a fully-connected layer and full back propagation**: A large ensemble model is created from individual models for each modality. Each individual model receives it's own set of input features and generates a set of outputs for a single modality. The outputs from each model are then concatenated and fed into a final, fully-connected layer. The whole model is trained end-to-end with back-propagation enabled for the full network. This allows the model to better learn the relationships between modalities at a higher-level [39].
- **Attention-based fusion**: Cross-modal transformers are used so each modality is able to receive information from the other modalities. This method includes both cross-modal attention and self-attention. It learns the low-level relationships between each pair of modalities and doesn't require them to be temporally aligned [40].

For the first three fusion methods, a ResNet-50 is used as the model. In the case of decision-level fusion, the ensemble model consists of multiple ResNet-50 models with an additional fully-connected layer. We then use the attention-based fusion method proposed by Tsai et al. [40] which uses a Multimodal Transformer(MulT) model composed of multiple cross-modal transformers. This method extends the transformer architecture by being able to learn representations from unaligned multi-modal streams. It is built from stacks of pairwise and bidirectional cross-modal attention blocks. Each of the transformers reinforces one modality with the low-level features of another modality using attention. This is modelled for each pair of modalities. MulT outperformed prior methods on a range of multimodal affect recognition datasets. The implementation is available publicly available on GitHub[1]. For multi-subject fusion, we increased the number of cross-modal transformers, cross-modal attention and self-attention blocks to 4 to include the audio modality. For each participant in the interaction, we pass the 4 input features into the modified MulT model and output the features from the final projection layer before the final fully connected layer. These output features are then used as the input to another modified MulT to introduce cross-subject attention which captures the relationships between the features of both interactants.

[1]https://github.com/yaohungt/Multimodal-Transformer

III. Experiments

We produce video-level predictions on the personality traits of an individual using 1) a range of audio-visual cues from the target individual and 2) a combination of audio-visual cues from the target individual and the cues from their conversational partner. Specifically, this paper conducted three sets of experiments:

1) We evaluate a set of fusion frameworks using the spectral representation of behaviour primitives from only the target participant;
2) We then improve upon the fusion frameworks and evaluate them on the spectral representation of behaviour primitives from both interlocutors;
3) Finally, we propose and evaluate a task-independent multimodal framework for personality recognition in dyadic interactions.

A. Dataset

We carry out our experiments on the multilingual UDIVA dataset [41]. The UDIVA dataset consists of 90.5 hours of dyadic interactions with 147 participants. There are in total 188 sessions divided into 4 different tasks: Animals, Ghost, Lego, and Talk. Each task was designed to elicit certain behaviours from participants. Participants appear between 1 and 5 videos with an average of 2.5 videos per participant. For example, the Lego section was designed to foster collaboration, whereas the Ghost section was designed to elicit competitive behaviour. The recordings were captured using 6 cameras, and audio was recorded through a microphone worn by each participant in addition to a microphone placed on the table. All recording devices are time-synchronised. Participants completed questionnaires to provide information including age, gender, ethnicity, occupation, maximum level of education, and country of origin which are all included in the dataset metadata. Scores for each of the Big-5 personality traits were assessed through standardized questionnaires and included in the metadata.

The dataset is also multilingual with Spanish being the predominant language followed by Catalan and English. The dataset is split into the training, validation and test splits with 116, 18 and 11 sessions respectively. Each split also contains 99, 20 and 15 participants respectively.

B. Training and model settings

For each experiment, we performed an automated bayesian search on the learning rate and the batch-size. We limited our automated search to these two hyperparameters as the

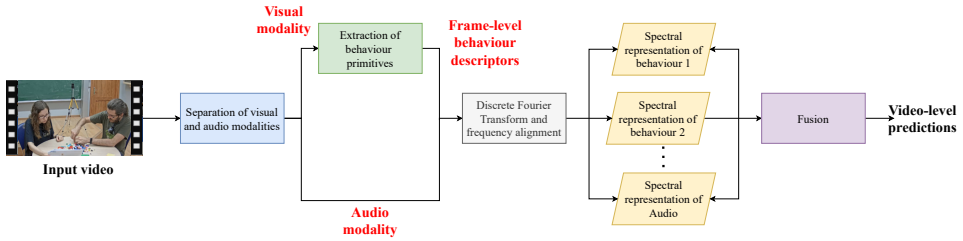

Fig. 2: The proposed framework.

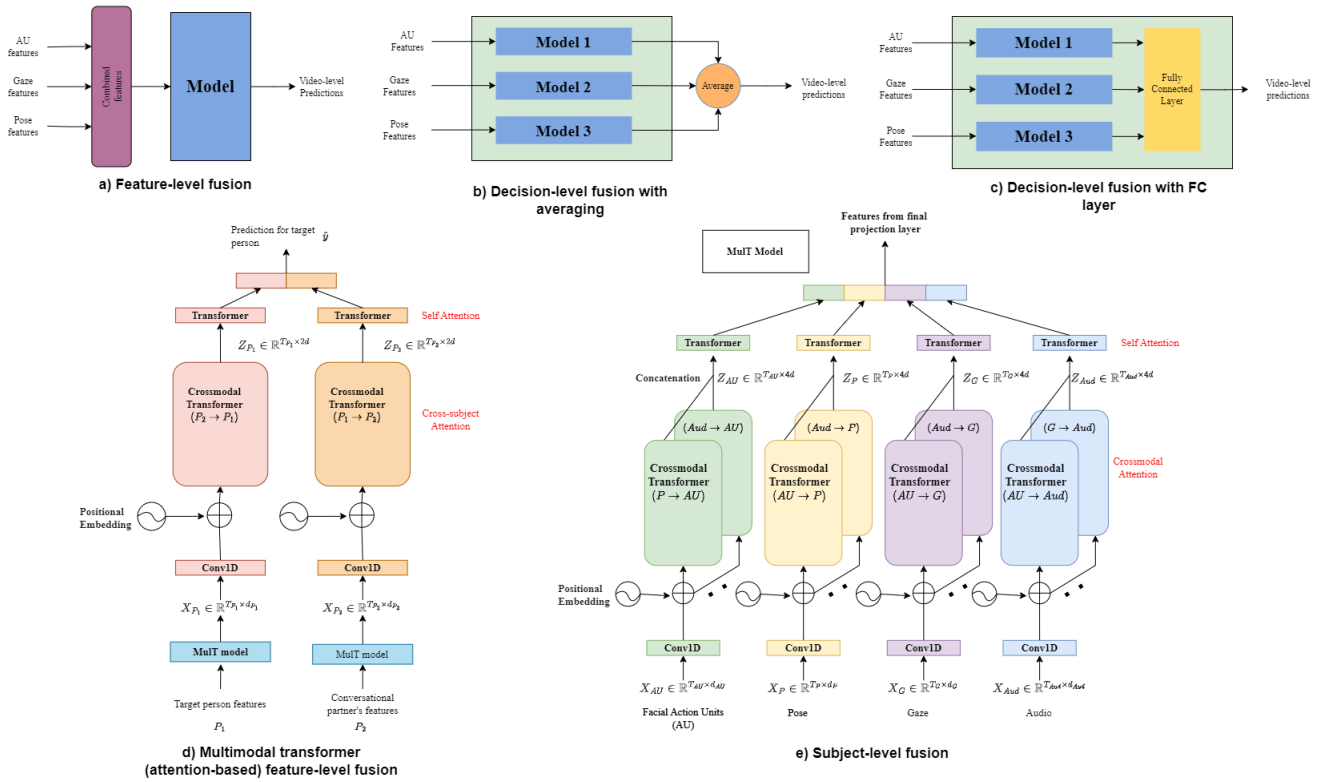

Fig. 3: Diagram depicting a) Feature-level fusion b) Decision-level fusion wit averaging c) Decision-level fusion with a fully-connected layer and full back propagation d) the MulT model augmented to take all four modalities as input for a single participant and e) the augmented MulTModel for two participants taking the output features from the final projection layer of the first MulT model as input

hyperparameter space was too large to search exhaustively. We bound the batch-size to a range of $[5, 30]$ to balance the constraints imposed by our GPU capacity with maintaining training stability. We also bound the learning rate to fall in the range $[1 \times 10^{-1}, 1 \times 10^{-6}]$ which we obtained from experimentation. We utilised early-stopping to preevent overfitting.

The full ResNet-50 architecture was not always appropriate for our experiments due to the low dimensionality of the spectral representations which sometimes led to overfitting. Through experimentation, we observed it was necessary to reduce the number of residual blocks to mitigate this. Furthermore, we found that reducing the width of the first residual block from 64 to 8 helped stabilise the training process. The model is trained end-to-end and optimised on the video-level MSE (MSE_{seq}) loss using Adam.

C. Metrics

We evaluate the predictions using the same metrics reported in USB challenge, namely the mean-squared error loss (MSE) though we also report the mean absolute error (MAE) to offer an alternative perspective. Our results reflect the MSE and MAE loss at both the participant level($_{part}$) and video-sequence level ($_{seq}$) in addition to the Pearson correlation

coefficient (PCC) all of which are defined below.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{4}$$

We calculate the average MSE over the 5 personality traits (AMSE) as:

$$AMSE = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{5}\sum_{i=1}^{5}(p_{i,j} - g_{i,j})^2 \tag{5}$$

The Pearson correlation coefficient (PCC) is a metric used to calculate the linear relationship between the variables with a score in the range$[-1, 1]$ where -1 means a perfect negative correlation and 1 represents a perfect positive correlation. A score of 0 means that the two variables are completely uncorrelated. This metric is useful to understand if there is any linear correlation between the model's predictions and the ground-truth though it ignores scale.The formula for the PCC between the predictions and ground truths $(\rho X, Y)$ across all samples in the test set is given by:

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma X \sigma Y} \tag{6}$$

where $p_{i,j}$ be the predicted personality trait score and $g_{i,j}$ be the ground truth for $(1 \leq i \leq 5)$ for a sample j. The models are trained to minimise MSE_{seq}, but are evaluated on both.

IV. Results and discussion

First, we report our results obtained from the spectral representations of the visual features from only the target participant with the results shown in the supplementary material. We then evaluate the same four fusion methods combining features from both interactants with the addition of the audio modality. The results can be seen in Tables I, III and II. We include the audio modality as previous works have shown this typically leads to an improvement in performance.

A. Multimodal fusion

When decision-level fusion was applied to features from the target participant alone, the Ghost task achieved the lowest MSE_{part} for Extroversion with Talk not far behind. The AMSE decreases significantly (64% in the case of the Animals task) when decision-level fusion is used. In fact, all fusion methods lead to an improvement in the AMSE in comparison to individual modalities. The results for the individual modalities can be found in the supplementary material. The individual modalities were not very effective indicators, but together are able to create a more complete picture of the participant's behaviour when combined at the decision-level therefore leading to a much lower error. This implies that the modalities interact in a non-trivial way that can be better learned through back-propagation. However the extent of the improvement is unprecedented in the literature as we have achieved state-of-the-art on this dataset which we attribute to the effectiveness of behaviour primitives at capturing behavioural cues.

B. Cross-subject fusion

Cross-subject fusion saw little variation in the performance across the four tasks. Out of the four fusion methods, our attention-based fusion approach achieved the lowest validation MSE_{seq} in 3 of the tasks. Not only does this method have the advantage of being able to represent complex relationships between modalities, it is also able to do this with multiple interactants. This advantage translated into a lower MSE_{part}. The worst performing task was Lego, though by a rather small margin. This was likely due to the increased MSE_{part} for Openness. Animals was the most informative for Openness which was also found in previous works [28] [41]. Ghost was most informative for Agreeableness and Neuroticism. The Ghost task is the most likely of the four to elicit the observable characteristics of Neuroticism as it was designed to encourage competitive behaviour from the interactants. Talk was the best performing-task for Conscientiousness and Extroversion. In the Talk task, each participant is required to speak continuously for 5 minutes about a subject of their choice. This correlates with behaviours typically associated with Extroversion such as longer speaking turns. Furthermore, the more formal context of a 5 minute speech or presentation style interaction is likely to have contributed to Extroversion being more easily discernable amongst the interactants as the speech of extroverts and introverts has been shown to differ in a more formal context [42].

Overall, the use of features from both interlocuters led to notable improvements in performance for 4 of the 5 traits. The error for Openness actually increased for 3 of the 4 tasks after we performed cross-subject fusion. The Ghost task was the only one to see an improvement of approximately 6% from single-person to multi-person fusion for Openness. It is unclear why this occurs, but it can be inferred that Openness does not benefit any further from cross-subject fusion than it does for multimodal fusion in the framework we propose. The Talk task also achieved a higher AMSE with cross-subject fusion. This could be due to the interaction being dominated by each interactant in turn with less casual interaction meaning that including the features of the other interlocutor may not have been of much benefit. This is supported by Openness and Neuroticsm, the traits more concerned with an indidvidual's internal state, suffering the most from cross-subject fusion. Conscientiousness and Extroversion were the only traits to see any benefits in this task with around a 4% improvement.

The AMSE was improved by 3.9%, 5.7% and 11.2% for the Animals, Ghost and Lego tasks respectively. This adds to the improvements already gained from performing multimodal fusion.

C. Task-independent cross-subject fusion

We extend the framework to be trained across all four tasks in the dataset. This allows us to directly compare our results against similar works that train task-independent models. We repeat the experiments using features from both interlocutors, but use the the videos from all the tasks to train a single model. Our results are shown in Table II.

	Animals			Ghost			Lego			Talk		
	MSE	**MAE**	**PCC**	**MSE**	**MAE**	**PCC**	**MSE**	**MAE**	**PCC**	**MSE**	**MAE**	**PCC**
Feature-level fusion	0.4395	0.5283	0.7837	**0.4265**	**0.5223**	**0.7936**	0.4428	0.5435	0.7826	0.4409	0.5402	0.7818
Decision-level fusion	0.7130	0.6679	0.6616	0.4977	0.5696	0.7507	0.8489	0.7389	0.6858	0.6522	0.6354	0.7556
Attention-based fusion	**0.4362**	**0.5325**	**0.7879**	0.4323	0.5333	0.7893	**0.4375**	**0.5383**	**0.7887**	**0.4396**	**0.5352**	**0.7851**
Decision-level (simple)	1.124	0.8503	0.2355	1.132	0.8505	0.2288	1.107	0.8433	0.2524	1.129	0.8522	0.2182
Test	0.4456	0.5578	0.7504	0.4443	0.5661	0.7432	0.4587	0.5618	0.7501	0.4496	0.5594	0.7518

TABLE I: The validation MSE_{seq}, MAE_{seq}, and PCC achieved using the different fusion methods for each task. The results for the best performing trait for each task are highlighted in bold.

	O	C	E	A	N	Avg
Baseline [41]	0.744	0.794	0.886	0.653	1.012	0.818
Dyadformer [28]	-	-	-	-	-	0.722
SMART-SAIR [25]	0.711	0.723	0.867	0.548	0.997	0.769
Gender-wise Bimodal NAS [25]	0.684	0.588	0.830	0.550	0.796	0.690
Task-independent framework (ours)	**0.5978**	**0.3683**	**0.3262**	**0.4016**	**0.5020**	**0.4392**

TABLE II: The AMSE achieved by previous works on the UDIVA dataset in comparison to the task-independent variant of the proposed framework

Attention-based fusion was yet again the most successful of the four fusion methods, closely followed by feature-level fusion. Feature-level fusion performed better than attention-based fusion on the Ghost task. We theorise that concatenating the features from both interactants allowed the model to still learn the relationships of the features whilst allowing the model to learn how much weight to apply to features from the non-target interactant.

V. Conclusion

In this work, we propose a novel multimodal framework for automatic personality recognition in dyadic interactions. We demonstrated the quantitative and qualitative benefits of using automatically extracted behaviour primitives over deep-learned features in addition to using their spectral representations to capture multi-scale temporal relationships between frames. We then investigated a set of strategies for multimodal fusion and multi-subject fusion. The single-person task-specific variant of our proposed framework out-performs the state-of-the-art on the UDIVA dataset by up to 44%. The addition of cross-subject fusion increases this to almost 46%. These results persisted when we extended the framework to be task-independent as it out-performed the state-of-the-art by 36%. Our most successful fusion approach was a multimodal transformer architecture enabled with both cross-modal and cross-subject attention though feature-level fusion achieved a comparable performance. The low-level interactions between the features from both interactants captured by feature-level fusion appeared to be almost as effective for the prediction of personality traits. This opens up the possibility of our framework being model-agnostic as feature-level fusion is a pre-training step which has no dependency on the model architecture.

A. Limitations and further work

There are several limitations of our proposed framework. We only experiment with a ResNet-50 and a MulT model as the core models, it is unclear how this may have contributed to our results or if our findings will remain consistent with a different model architecture. Investigating with different models would be helpful to better understand any dependencies our framework has on the model architecture and if it can be made truly model-agnostic. We also conduct a rather limited bayesian search on the hyperparameter space due to computational constraints. A more extensive search could produce even better results than the ones we obtained. This work could be extended to include metadata as a complimentary modality similar to the approaches in the related works [28] [25] [41] as this has been shown to improve performance in almost all cases. This work can further be extended to other datasets on dyadic or group interactions, particularly datasets of a different context such as the NoXi dataset [43] and the AMIGOS dataset [44]. Finally, an evaluation of this framework in the context of human-robot interaction would serve to understand the improvements gained by the addition of improved contextual information and its contribution towards more intelligent interaction in multi-person scenarios.

		O	C	E	A	N	Avg
Animals	**Mean value baseline**	0.731	0.871	0.988	0.672	1.206	0.894
	Palermo et al. [41]	0.737	0.756	0.887	0.58	1.023	0.797
	Dyadformer [28]	0.674	1.239	1.448	**0.134**	0.947	0.888
	Single-person fusion (Ours)	**0.370**	0.500	0.381	0.579	**0.427**	0.451
	Cross-subject fusion (Ours)	0.593	**0.373**	**0.333**	0.409	0.461	**0.434**
Ghost	**Mean value baseline**	0.733	0.887	0.991	0.674	1.220	0.901
	Palermo et al. [41]	0.741	0.893	0.844	0.667	1.139	0.857
	Dyadformer [28]	0.771	0.691	0.754	0.616	1.029	0.772
	Single-person fusion (Ours)	0.650	0.450	**0.331**	0.456	0.417	0.461
	Cross-subject fusion (Ours)	**0.614**	**0.415**	0.352	**0.387**	**0.405**	**0.434**
Lego	**Mean value baseline**	0.738	0.871	0.99	0.676	1.204	0.896
	Palermo et al. [41]	0.727	0.763	0.826	0.611	1.037	0.793
	Dyadformer [28]	0.741	0.635	0.736	0.747	0.908	0.753
	Single-person fusion (Ours)	**0.411**	0.598	0.531	0.447	0.544	0.506
	Cross-subject fusion (Ours)	0.601	**0.376**	**0.344**	**0.413**	**0.513**	**0.450**
Talk	**Mean value baseline**	0.731	0.872	0.991	0.673	1.211	0.896
	Palermo et al. [41]	0.773	0.79	0.869	0.67	0.985	0.817
	Dyadformer [28]	0.574	0.504	0.419	0.683	1.135	0.663
	Single-person fusion (Ours)	**0.423**	0.382	0.343	**0.404**	**0.303**	**0.371**
	Cross-subject fusion (Ours)	0.595	**0.365**	**0.328**	0.405	0.503	0.439

TABLE III: Results for both our single-person method and our cross-subject method per trait and task. The "Avg" column represents the average performance over all the traits (AMSE). We compare our results with the two best performing works on the dataset that reported task-specific results. Best result per task and trait are in bold.

Acknowledgments: H. Gunes is supported by the EPSRC/UKRI under grant ref. EP/R030782/1 (ARoEQ). For open access purposes, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising. The dataset(s) used for this publication can be accessed upon request following the terms and conditions of the dataset owners.

VI. Ethical Impact Statement

This research was conducted on a dataset of recordings with human participants containing both video and audio data in addition to metadata containing personal information. An ethics review on the dataset was conducted and approved by the IRB. All participants consented for their data to be used for research purposes, however our proposed method aims to provide anonymous but cues to mitigate ethical concerns.

It is to be noted that there was a demographic imbalance in the dataset with the majority of participants being of a particular racial group. This could lead to issues of bias in our method as it is unclear if our results will generalise to individuals of other racial groups. There may also be biases encoded in the toolkit we use to extract behavioral primitives which may be perpetuated in our results. Furthermore, the labels were generated using self-reported questionnaires which may be biased due to self-presentation.

Despite the intentions of this work to contribute towards research with positive societal impact, there are potentially negative applications of this work. The methods described could be applied towards surveillance, screening and algorithmic decision-making processes with some unintended consequences including exacerbating social biases. Individuals may also feel uncomfortable with aspects of their internal state being perceived and acted on in a way that appears manipulative or invasive. To mitigate these risks, we use a dataset in this work which has been limited by the authors to open-source research applications and can not be used commercially.

References

[1] S. Jaiswal, S. Song, and M. Valstar, "Automatic prediction of depression and anxiety from behaviour and personality attributes," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 1–7.

[2] H. Salam, O. Çeliktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, vol. 5, pp. 705–721, 2017.

[3] S. Andrist, B. Mutlu, and A. Tapus, "Look like me: Matching robot personality via gaze to increase motivation," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 3603–3612. [Online]. Available: https://doi.org/10.1145/2702123.2702592

[4] O. Celiktutan and H. Gunes, "Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2015, pp. 815–820.

[5] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, 1992.

[6] R. R. McCrae, "Why i advocate the five-factor model: Joint factor analyses of the neo-pi with other instruments," 1989.

[7] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[8] R. Liao, S. Song, and H. Gunes, "An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition," *arXiv preprint arXiv:2210.09138*, 2022.

[9] S. T. McAbee and B. S. Connelly, "A multi-rater framework for studying personality: The trait-reputation-identity model." *Psychological Review*, vol. 123, no. 5, p. 569, 2016.

[10] I. Krumpal, "Determinants of social desirability bias in sensitive surveys: a literature review," *Quality & quantity*, vol. 47, no. 4, pp. 2025–2047, 2013.

[11] P. D. Blanck, R. Rosenthal, M. Vannicelli, and T. D. Lee, "Therapists' tone of voice: Descriptive, psychometric, interactional, and competence analyses," *Journal of Social and Clinical Psychology*, vol. 4, no. 2, pp. 154–178, 1986. [Online]. Available: https://doi.org/10.1521/jscp.1986.4.2.154

[12] D. Rutter, I. E. Morley, and J. C. Graham, "Visual interaction in a group of introverts and extraverts," *European Journal of Social Psychology*, vol. 2, no. 4, pp. 371–384, 1972.

[13] D. C. Funder and C. D. Sneed, "Behavioral manifestations of personality: An ecological approach to judgmental accuracy." *Journal of personality and social psychology*, vol. 64, no. 3, p. 479, 1993.

[14] T. DeGroot and J. Gooty, "Can nonverbal cues be used to make meaningful personality attributions in employment interviews?" *Journal of business and psychology*, vol. 24, pp. 179–192, 2009.

[15] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[16] F. Valente, S. Kim, and P. Motlicek, "Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus," in *Proc. Interspeech 2012*, 2012, pp. 1183–1186.

[17] N. Madzlan, J. Han, F. Bonin, and N. Campbell, "Towards automatic recognition of attitudes: Prosodic analysis of video blogs," *Speech Prosody, Dublin, Ireland*, pp. 91–94, 2014.

[18] S. Song, S. Jaiswal, E. Sanchez, G. Tzimiropoulos, L. Shen, and M. Valstar, "Self-supervised learning of person-specific facial dynamics for automatic personality recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[19] F. Gürpınar, H. Kaya, and A. A. Salah, "Combining deep facial and ambient features for first impression estimation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 372–385.

[20] C. Beyan, A. Zunino, M. Shahid, and V. Murino, "Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 1084–1099, 2019.

[21] F. Alam and G. Riccardi, "Predicting personality traits using multimodal information," in *Proceedings of the 2014 ACM multi media on workshop on computational personality recognition*, 2014, pp. 15–18.

[22] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 337–348.

[23] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*. Cambridge University Press, 1995.

[24] R. Cuperman and W. Ickes, "Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables"," *Journal of personality and social psychology*, vol. 97, pp. 667–84, 10 2009.

[25] H. Salam, O. Celiktutan, V. Manoranjan, I. Ismail, and H. Mukherjee, "Iccv 2021 understanding social behavior in dyadic and small group interactions challenge," 2021.

[26] Z. Shao, S. Song, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, *Personality Recognition by Modelling Person-Specific Cognitive Processes Using Graph Representation*. New York, NY, USA: Association for Computing Machinery, 2021, p. 357–366. [Online]. Available: https://doi.org/10.1145/3474085.3475460

[27] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning person-specific cognition from facial reactions for automatic personality recognition," *IEEE Transactions on Affective Computing*, 2022.

[28] D. Curto, A. Clapés, J. Selva, S. Smeureanu, J. C. S. J. Junior, D. Gallardo-Pujol, G. Guilera, D. Leiva, T. B. Moeslund, S. Escalera, and C. Palmero, "Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions," 2021.

[29] S. Okada, L. S. Nguyen, O. Aran, and D. Gatica-Perez, "Modeling dyadic and group impressions with intermodal and interperson features," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1s, jan 2019. [Online]. Available: https://doi.org/10.1145/3265754

[30] W. Mou, H. Gunes, and I. Patras, "Alone versus in-a-group: A comparative analysis of facial affect recognition," in *Proceedings of*

the 24th ACM International Conference on Multimedia, ser. MM '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 521–525. [Online]. Available: https://doi.org/10.1145/2964284.2967276

[31] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[32] N. I. Abbasi, S. Song, and H. Gunes, "Statistical, spectral and graph representations for video-based facial expression recognition in children," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1725–1729.

[33] P. Ekman, "Facial expression and emotion." *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[34] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[35] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[36] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[37] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 829–844, 2022.

[38] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *CoRR*, vol. abs/1705.09406, 2017. [Online]. Available: http://arxiv.org/abs/1705.09406

[39] O. Kampman, E. J. Barezi, D. Bertero, and P. Fung, "Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction," *CoRR*, vol. abs/1805.00705, 2018. [Online]. Available: http://arxiv.org/abs/1805.00705

[40] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Florence, Italy: Association for Computational Linguistics, 7 2019.

[41] C. Palmero, J. Selva, S. Smeureanu, J. Junior, A. Clapés, A. Moseguí Saladié, Z. Zhang, D. Gallardo-Pujol, G. Guilera, D. Leiva, and S. Escalera, "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset," 01 2021, pp. 1–12.

[42] J.-M. Dewaele and A. Furnham, "Personality and speech production: A pilot study of second language learners," *Personality and Individual Differences*, vol. 28, pp. 355–365, 02 2000.

[43] A. Cafaro, J. Wagner, T. Baur, S. Dermouche, M. Torres Torres, C. Pelachaud, E. André, and M. Valstar, "The noxi database: Multimodal recordings of mediated novice-expert interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ser. ICMI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 350–359. [Online]. Available: https://doi.org/10.1145/3136755.3136780

[44] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2021.