ICCV 2021 Understanding Social Behavior in Dyadic and Small Group Interactions Challenge

Fact sheet: Automatic self-reported personality recognition Track

I. Team details

- Name of joint team leaders (and affiliation):
  Hanan Salam
  SMART Lab
  Department of Computer Science
  New York University Abu Dhabi, UAE
  Username on Codalab: **hanansalam**
  Email: hanan.salam@nyu.edu
  Oya Celiktutan
  SAIR Lab, Centre for Robotics Research
  Department of Engineering
  King's College London, UK
  Email: oya.celiktutan@kcl.ac.uk
- Name of other team members (and affiliation):
  Viswonathan Manoranjan
  SMART Lab
  Department of Computer Science
  New York University Abu Dhabi, UAE
  Email: vm2336@nyu.edu
  Iman Ismail
  SAIR Lab, Centre for Robotics Research
  Department of Engineering
  King's College London, UK
  Email: iman.a.ismail@kcl.ac.uk
  Himadri Mukherjee
  SMART Lab
  Department of Computer Science
  New York University Abu Dhabi, UAE
  Email: himadri.mukherjee@nyu.edu

II. Learning Personalised Models for Automatic Self-Reported Personality Recognition

Smart phones, voice assistants, and home robots are becoming more intelligent every day to support humans in their daily routines and tasks. Achieving the user acceptance and success of such technologies makes it necessary for them to be socially informed, responsive, and responsible. They need to understand human behaviour and socio-emotional states and adapt themselves to their user's profiles (e.g., personality) and preferences. Motivated by this, there has been a significant effort in recognising personality from multimodal data in the last decade [1], [2]. However, to the best of our knowledge, the methods so far have focused on one-fits-all approaches only and performed personality recognition without taking into consideration the user's profiles (e.g., gender and age). In this paper, we took a different approach, and we argued that one-fits-all approach does not work sufficiently for personality recognition as previous research showed that there are significant gender differences in personality traits. For example, women tend to report higher scores for extraversion, agreeableness and neuroticsm as compared to men [3]. Building upon these findings, we first clustered the participants into two profiles based on their gender, namely, female and male, and then used Neural Architecture Search (NAS) to automatically design a model for each profile to recognise personality. A separate network was designed and trained with visual and text features. The final prediction was obtained by aggregating the results of both video and text modalities. Figure 1 presents the overview of our proposed approach.

A. Pre-processing

Each participant's video and corresponding transcript file was divided into 1 minute data slices. To split the videos into 1 minute video clips, the number of frames was considered. As the videos were recorded at 25 frames per second, each 1500 frames corresponding to 1 minute video slices were extracted. The speech transcripts were divided into 1 minute transcripts based on the provided timestamps.

A single turn is presented as follows:
1
00:00:00,115 –¿ 00:00:01,865
PART.2: Preguntas
de ”sí” o ”no”, ¿eh? Acuérdate.

Here, “1” in the 1^{st} line corresponds to the turn number. This is followed by timestamps of the start and end of dialogue. The timestamp is in HH:MM:SS,SSS format. The MM value was used to split the transcripts per minute. If a dialogue spanned across a minute boundary, it was assigned to the previous minute (when started). Here, “PART.2” indicates it was spoken by person 2 in the respective video. This value was used to split the dialogues in a person specific manner.

B. Multimodal Features Extraction

We extract features from the video and text modalities. The following describes the features extraction process in detail.

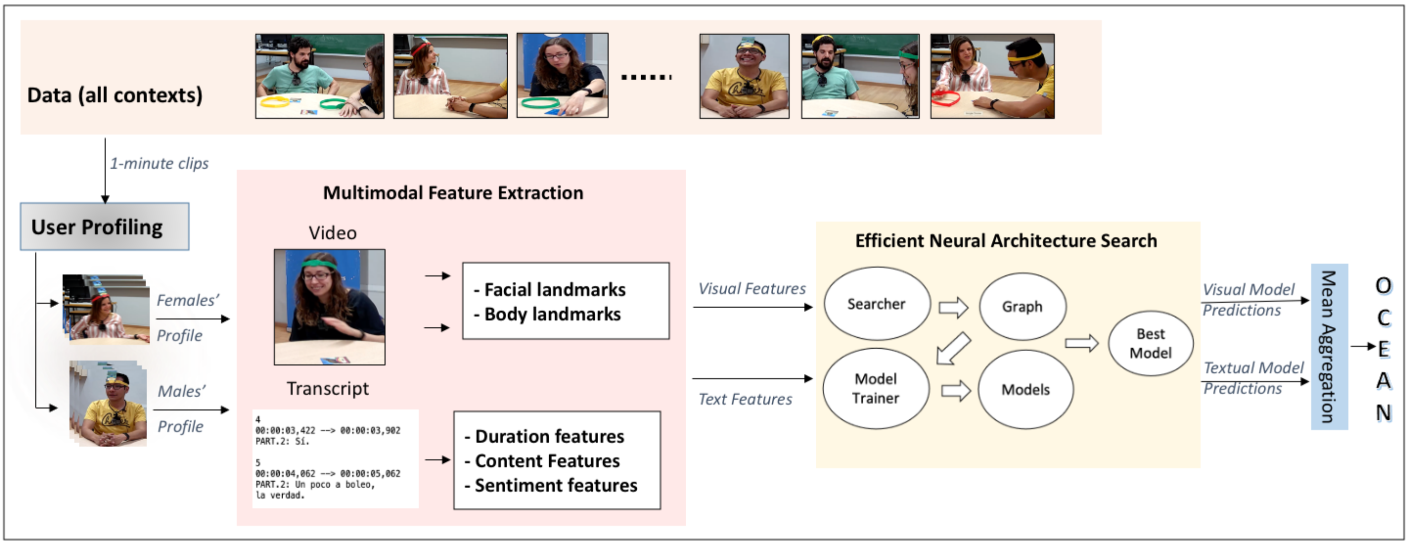

Fig. 1. Overview of the proposed approach: We first divided the videos into 1 minute short clips and clustered the participants into two profiles based on their gender, namely, female and male. We then extracted a set of visual and text features, which were given as input to the Neural Architecture Search (NAS) framework to automatically design a model for each profile and for each modality to recognise personality. For each user profile, the final prediction was obtained by aggregating the results of both video and text modalities.

1) Video-based features: We consider both facial and body pose landmarks for personality prediction. These video-based features were extracted from the annotations provided by the challenge organizers. The features taken into consideration are:

a) Facial landmarks: 68 facial landmarks were provided for each video frame along 3 dimensions. The data was first flattened to obtain a facial landmark array of dimension $(1, 204)$. We then calculated the mean and standard deviation of each facial landmark point over all the frames in a 1 minute video clip, resulting in a feature vector of $(1, 408)$.

b) Body landmarks: 24 three dimensional body landmarks were provided for each video frame. The data was first flattened to obtain an array of dimension $(1, 72)$. We then calculated the mean and standard deviation of each body pose landmark points over all the frames in each 1 minute video clip, resulting in a feature vector of $(1, 144)$.

Both the face and body landmark statistics were concatenated, resulting in a feature vector of dimension 552 for each 1 minute video clip.

2) Text-based features: The transcripts of the interactions were analyzed based on each talk turn content and duration. The extracted features include talk turn duration, content and sentiment.

a) Talk turn duration features: The duration of interaction for a person in a single minute was analyzed to generate a 5 dimensional feature set consisting of the following:

- Minimum turn duration: The minimum time (at the turn level) for which a person talked.
- Maximum turn duration: The maximum time (at the turn level) for which a person talked.
- Average turn duration: The average time across all turns for a particular person in a single minute.
- Standard deviation of turn duration: The standard deviation of time taken in each turn for a single person over a 1 minute segment. This gave an idea of the variation in the time spent on different interactions.
- Total duration of turns: The total amount time a person spoke in a single minute.

b) Talk turn content features: The number of turns and the content of every dialogue was analyzed and 5 features were generated, which consist of the following:

- Turn percentage: The percentage of turns for a particular person out of the total number of turns in a single minute.
- Average words per turn: The average number of words spoken by a person in a turn across a 1 minute window.
- Longest turn: The largest number of words among all the turns over a minute for a particular person.
- Total number of words: The total number of words uttered over all the turns in a minute for a particular person.
- Standard deviation of words per turn: The standard deviation of the number of words per turn was computed to quantify the variance of the amount of vocal interaction by a particular person over a minute.

c) Talk turn sentiment features: Each of the 1 minute transcripts was analyzed to generate 10 sentiment-based features. Since the majority of the conversations was in Spanish (71.8%) [4], a Spanish sentiment recognition library was used [5]. Moreover, to the best of our knowledge, there is no sentiment recognition system for Catalan language. The generated sentiment values ranged in between 0 to 1 where 0 corresponds to fully negative and 1 corresponds to fully positive sentiment. The following sentiment-based features were computed:

- Most negative turn: The sentiment of texts across the turns over a minute for each person was computed and the smallest value was used.

- Most positive turn: The sentiment of texts across the turns over a minute for each person was computed and the largest value was used.
- Average sentiment: The average sentiment over all the turns in one minute was computed per person.
- Sentiment variation: The standard deviation of sentiment values across the turns for a person over a minute was computed. This gave an idea of the variation of sentiment over successive expressions.
- Sentiment range: The sentiment range was divided into 5 equi-spaced classes corresponding to highly negative, negative, almost neutral, positive, and highly positive. The number of turns across a minute over these classes was computed and then normalized with the total number of turns by the person in that particular minute. This resulted in 5 features.
- Overall sentiment: The sentiment value was computed over the 1-minute segment per person, without dividing into turns.

C. Personalized Neural Architecture Search Strategy

In order to train personalized personality prediction models, first we created different profiles by grouping the individuals in the dataset into females and males. An adaptive neural architecture was designed automatically with Neural Architecture Search (NAS) [6] and trained for each profile.

Neural architecture search (NAS) has been proposed to automatically adjust deep neural networks, without the need for manually designing the architecture. Existing search algorithms include NASNet [7], PNAS [8], and Auto-Keras [6]. We used the NAS framework proposed by [6] due to its efficiency. The approach employs an efficient training during search via network morphism, which keeps the functionality of a neural network while changing its neural architecture through morphism operations. The framework uses Bayesian optimization to guide the network morphism to enable efficient neural architecture search. To this end, a neural network kernel based on edit distance was designed and an algorithm was proposed to optimise the acquisition function in the tree-structured space. The algorithm was also implemented in an open-source AutoML system called Auto-Keras [6].

1) Implementation details: We used Auto-Keras [6] to perform the neural architecture search. For both visual and text modalities, we used two dense layers with 32 units as the default architecture. The original training data is divided into training and validation sets using an $85-15\%$ split strategy for text features and an $85-25\%$ split strategy for visual features. The best models were searched by employing network morphism operations such as inserting new layers, expanding existing layers, or adding skip connections. The loss function was mean squared error and each network was trained with ADAM optimiser. The number of epochs was set to 1000. The number of trials was set to 100. Finally, we used an early stopping with patience equal to 30.

D. Decision Fusion

We applied decision fusion to predict the personality of an individual. The scores obtained per minute were averaged over all the sessions. Thereafter, the resulting values were aggregated across the different modalities using the average predictions of both modalities.

E. Challenge Results

The trained best models were evaluated on the unseen test samples provided by the challenge organisers. In Table I, we provided our obtained results as shown in the leaderboard of the challenge[1].

F. Final Remarks

The proposed system has two main advantages. Firstly, it combines multiple modalities to predict the personality of an individual. Secondly, the system is scalable and can adapt itself to changing trends in the data as the neural architecture search-based approach enables generation of a deep learning model depending on the user profile.

We observed that visual and textual features performed almost equally well. On the test set, textual features (i.e., talk turn-based features) yielded slightly smaller error on average as compared to visual features (i.e., facial and body landmark features) and provided an improvement by 0.006. The best error value of 0.769153 was obtained by combining both the visual and textual features.

As a future work, we plan to incorporate audio features such as Mel Frequency Cepstral Coefficents and short-time average energy features. One limitation is that the models were separately trained on 1 minute segments for each of the modalities and after their predictions were aggregated. Instead, we plan to apply more sophisticated techniques and explore feature level fusion strategies to train a single model for predicting an individual's personality.

III. Additional method details

- **Mark with an X the modalities you have exploited.** (X) Visual, ( ) Acoustic, (X) Transcripts, (X) Metadata, (X) Landmark annotations, ( ) Eye-gaze vectors.

- **In case you used metadata, mark with an X the types of metadata you have exploited.** ( ) Age, (X) Gender, ( ) Country of origin, ( ) Max. level of education, ( ) Pre-session mood, ( ) Post-session mood, ( ) Pre-session fatigue, ( ) Post-session fatigue, ( ) Relationship among interactants, ( ) Task type, ( ) Task order, ( ) Task difficulty, ( ) Language, ( ) Other.
  If "other", or if you have used just a subset of info for a given type of metadata (e.g., just a subset of mood values), please detail:

[1] https://competitions.codalab.org/competitions/31326

TABLE I
RESULTS FROM LEADERBOARD (TEST PHASE) OBTAINED BY THE PROPOSED APPROACH.

Rank position	O	C	E	A	N	MSE
1	0.711249 (1)	0.723010 (3)	0.866556 (1)	0.548260 (1)	0.996690 (1)	0.769153 (1)

- **Mark with an X the tasks you used for training.** (X) Talk, (X) Lego, (X) Animals, (X) Ghost.
- **Mark with an X the tasks you used for evaluation.** (X) Talk, (X) Lego, (X) Animals, (X) Ghost.
- **Did you use the provided validation set as part of your training set?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any fusion strategy of modalities?** (X) Yes, ( ) No
  If yes, please detail:
  Decision fusion was applied as explained above. The average of predictions from both modalities were computed resulting in a single prediction for each person.
- **Did you use ensemble models?** (X) Yes, ( ) No
  If yes, please detail:
  The models were trained separately for the video and transcript-based features. The obtained results were averaged to predict the final personality score of an individual.
- **Did you follow a multi-task approach or trained each trait individually?** ( ) Multi-task, (X) Trained each trait individually.
- **Did you use information from the other interlocutor (e.g., their visual info) to predict the personality of the target interlocutor?** ( ) Yes, (X) No.
  If yes, please detail:
- **Did you use pre-trained models?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use external data?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any regularization strategies/terms?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use handcrafted features?** (X) Yes, ( ) No
  If yes, please detail
  The transcripts were analyzed to generate different handcrafted features based on duration, content, and sentiment of talk turns. The details of the features are presented in Section II-B.2.
- **Did you use any pose estimation method?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any face / hand / body detection, alignment or segmentation strategy?** ( ) Yes, (X) No
  If yes, please detail:
- **At what level of granularity did your method perform personality inference?** ( ) Frame-level, (X) Audio/video chunk-level (i.e., short audio/video snippet), ( ) Task-level, ( ) Session-level, ( ) Other.
  If "other", please detail. If you selected "chunk-level", please comment on the chunk length and why you selected it:
  We selected one minute as our previous work showed that one minute provides adequate information to infer personality, and averaging features over longer video clips leads to information loss [9].
- **Did you use any aggregation method to compute a single personality prediction per participant?** (X) Yes, ( ) No
  If yes, please detail:
  The personality prediction per participant involved mean aggregation of the obtained scores per minute across all the sessions. Then, these scores were averaged over the different modalities.
- **Did you use any spatio-temporal feature extraction strategy?** ( ) Yes, (X) No
  If yes, please detail:
- **Did you perform any data augmentation?**
  ( ) Yes, (X) No
  If yes, please detail:
- **Did you use any bias mitigation technique (e.g., rebalancing training data)?**
  ( ) Yes, (X) No
  If yes, please detail:

IV. CODE REPOSITORY

The following repository includes the approach's implementation and the necessary instructions to run the code. **Code repository:** `https://github.com/SMART-Lab-NYU/ICCVChallenge_PersonalisedModelForPersonalityRecognition`

REFERENCES

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

[2] J. C. Silveira Jacques Junior, Y. Güçlütürk, M. Perez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. Van Gerven, R. Van Lier, and S. Escalera, "First impressions: A survey on vision-based apparent personality trait analysis," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[3] Y. Weisberg, C. DeYoung, and J. Hirsh, "Gender differences in personality across the ten aspects of the big five," *Frontiers in Psychology*, vol. 2, p. 178, 2011. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2011.00178

[4] C. Palmero, J. Selva, S. Smeureanu, J. C. J. Junior, A. Clapés, A. Moseguí, Z. Zhang, D. Gallardo, G. Guilera, D. Leiva *et al.*, "Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset." in *WACV (Workshops)*, 2021, pp. 1–12.

[5] Hugo J. Bello. sentiment-analysis-spanish 0.0.25. [Online]. Available: https://pypi.org/project/sentiment-analysis-spanish/

[6] H. Jin, Q. Song, and X. Hu, "Auto-keras: An efficient neural architecture search system," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2019, pp. 1946–1956.

[7] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter, "Auto-sklearn: efficient and robust automated machine learning," in *Automated Machine Learning*. Springer, Cham, 2019, pp. 113–134.

[8] C. Liu, B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy, "Progressive neural architecture search," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

[9] O. Celiktutan and H. Gunes, "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 29–42, 2017.