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

Fact sheet: Automatic self-reported personality recognition Track

I. TEAM DETAILS

- Team leader 1 name: **Hanan Salam**
- Username on Codalab: **hanansalam**
- Team leader 1 affiliation:
  SMART Lab, Department of Computer Science
  New York University Abu Dhabi, UAE
- Team leader 1 email: hanan.salam@nyu.edu
- Team leader 2 name: **Oya Celiktutan**
- Team leader 2 affiliation:
  SAIR Lab, Centre for Robotics Research
  Department of Engineering
  King's College London, England
- Team leader 2 email: oya.celiktutan@kcl.ac.uk
- Name of other team members (and affiliation):
  Viswonathan Manoranjan
  Department of Computer Science and Engineering
  National Institute of Technology, Tiruchirapalli, India
  email: viswonathan0606@gmail.com
  Iman Ismail
  SAIR Lab, Centre for Robotics Research
  Department of Engineering
  King's College London, England
  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 social companion robots are becoming more intelligent every day to support humans in their daily tasks. To ensure the acceptance and success of such technologies, they need to understand human behaviour 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. Each network was trained with visual, textual and time-based features separately. The final prediction was generated by aggregating the results of both the modalities. Figure 1 presents an overview of the 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, then it was assigned to the previous minute (when started). This followed by the dialogue. 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.

Prediction for each participant was first performed using the 1 minute clips and then aggregated to produce a single prediction for the participant in question.

B. Multimodal features extraction

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

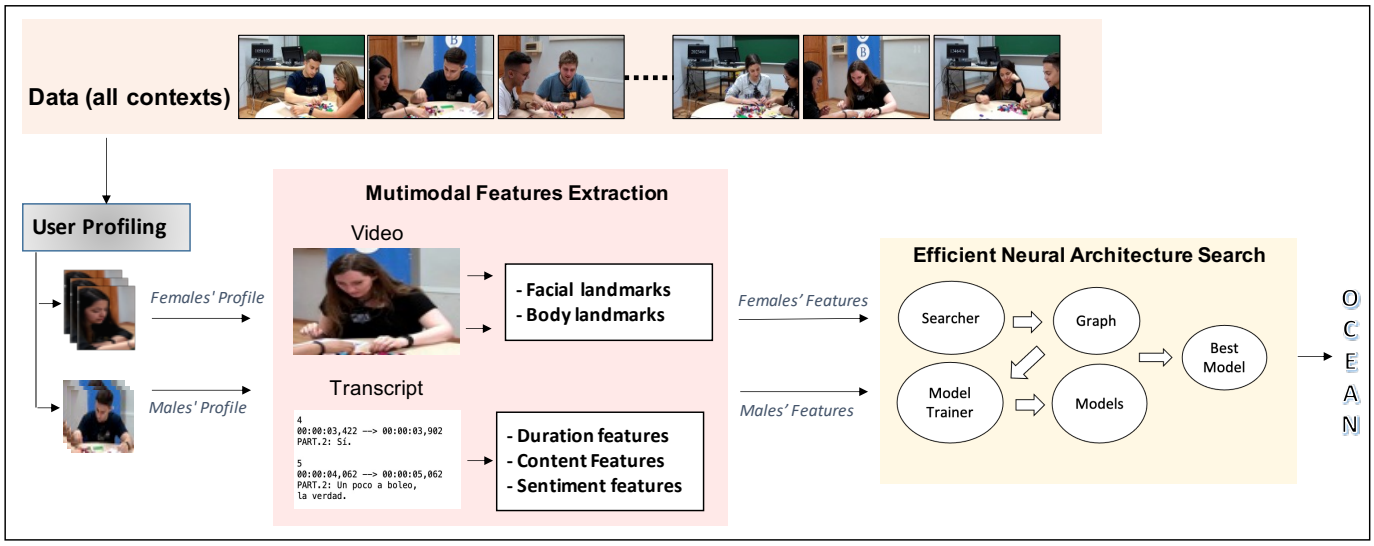

Fig. 1. Workflow of the proposed approach.

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 get a facial landmarks array of dimension $(1, 204)$. The mean and standard deviation of the facial landmark points corresponding to the frames in each 1 minute video was then calculated.

b) Body landmarks: 24 body 3 dimensional landmarks were provided for each video frame. The data was flattened to get data of dimension $(1, 72)$. The mean and standard deviation of the body pose landmark points corresponding to the frames in each 1 minute video was then calculated.

Both the face and body landmarks 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 for which a person talked (turn level).
- Maximum turn duration: The maximum time for which a person talked (turn level).
- 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 with the time spent on different interactions.
- Total duration of turns: The total amount time a person speak in a single minute.

b) Talk turn content features: The number of turns and the content of every dialogue was analyzed and 5 dimensional 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: This is the highest number of words amidst 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 single person.
- Standard deviation of words per turn: The standard deviation of the number of words per turn was computed to get 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 were analyzed to generate 10 dimensional 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. 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 least value was used.
- Most positive turn: The sentiment of texts across the turns over a minute for each person was computed and the highest value was used.
- Average sentiment: The average sentiment of a person

over all the turns in minute was computed.

- Sentiment deviation: The standard deviation of sentiments across the turns for a person over a minute was computed. This gave an idea of the variation of sentiment for 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 thereafter normalized with the total number of turns by the person in that particular minute. This contributed to 5 dimensions of the feature set.
- Overall sentiment: The sentiment value over all the turns of a person was computed.

C. Personalized Neural Architecture Search Strategy

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

Neural architecture search (NAS) has been proposed to automatically tune deep neural networks. Existing search algorithms include NASNet [7], PNAS [8]. We used the NAS framework proposed by [6]. The approach employs an efficient training during search via network morphism, which keeps the functionality of a neural network while changing its neural architecture. The framework enables Bayesian optimization to guide the network morphism for efficient neural architecture search. The framework develops a neural network kernel and a tree-structured acquisition function optimization algorithm to efficiently explore the search space. The algorithm was implemented in an open-source AutoML system, namely Auto-Keras.
This system runs parallelly on CPU and GPU with an adaptive search strategy for different GPU memory limits. Firstly, the application programming interface (API) is called by the user which its job is to call the corresponding middle-level modules to finish certain functionalities. Secondly, the module of the neural architecture search algorithm which is the searcher generates neural architectures on CPU. Then, the model trainer which is responsible for the computations on CPUs trains neural networks with the training data separately. After that, the Graph which is the module operating the computational graphs of neural networks, performs real neural networks with parameters on RAM from the neural architectures. Finally, the neural network is produced on GPU for training, and all the trained neural models are saved on storage devices.

1) Implementation details: We use mean squared error as a loss function. The original training data is divided into training and validation using an $85-15\%$ split strategy. The number of epochs was set to 1000. The number of trials was set to 100. An early stopping with patience equal to 30 was used.

D. Decision Fusion

The personality computation involved mean aggregation of the obtained scores per minute across all the sessions. Thereafter these scores were aggregated across the different modalities using the average predictions of both modalities.

E. Challenge Results

In Table I, we provided our obtained results, shown in the leaderboard of the challenge[1].

F. Final remarks

The proposed system has several advantages. Firstly, it analyzes multiple modalities to predict the personality of a subject. Secondly, the system can be deployed in resource constrained scenarios. A change in the MSE of only 0.006 on the test set was observed when 20-dimensional talk turn-based features were used instead of the 552-dimensional facial and body landmark features. The best MSE value of 0.769153 was obtained on combining both the features. The system is also scalable and can adapt itself to changing trends in the data as the neural architecture search-based approach involves generation of a deep learning architecture depending on the training set.

In the proposed approach, the vocal modality was not used to prevent heavy load on the system in terms of feature dimension. In future, audio-based features will be tested as well to see the performance of the system. Further, the models were separately trained on each of the modalities and thereafter their predictions was aggregated. We plan to fuse the features and train a single model for identifying personality. Finally, we will also use language specific sentiment analyzers for extracting better sentiment-based features.

III. Additional method details

Please, reply if your challenge entry considered (or not) the following strategies and provide a brief explanation. For the non-binary questions, you can mark multiple options.

- **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 aggregated (averaged) to predict the final personality of the subjects.
- **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 computation involved mean aggregation of the obtained scores per minute across all the sessions. Thereafter these scores were aggregated across the different modalities using the average.
- **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 a docker image and the necessary instructions to run the code. **Code repository:** `https://hub.docker.com/r/viswonathan/iccvproject`

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.