# Sequential Models for Automatic Personality Recognition from Multimodal Information in Social Interactions

**Jeanfed Ramirez Lima**
Department of Computer Science
INAOE
Tonanzintla, Puebla
jeanfed.ramirez@inaoe.mx

**Hugo Jair Escalante**
Department of Computer Science
INAOE
Tonanzintla, Puebla
hugojair@inaoep.mx

**Luis Villaseñor Pineda**
Department of Computer Science
Tonanzintla, Puebla
villasen@inaoep.mx

## Abstract

We study the problem of recognizing personality from videos depicting users' social interaction. Multimodal information is represented using pretrained models and multi-stream sequential models are considered for prediction. Experimental results of the proposed method in the recently released UDIVA dataset are reported and compared to related work. We show that the proposed methodology is competitive with the state-of-the-art while using less complex models.

## 1 Introduction

During the past decade, automatic personality recognition has become one of the topics of highest interest in the affective computing field [1, 2], primarily motivated by the ease with which personal information can be recorded and stored, for instance, using smartphone cameras. Therefore, automated approaches for personality analysis from video recordings have flourished in recent years. Specifically, there is a trend in recognizing personality using information derived from social environments. The principal purpose of this task is to find patterns of behavior leveraged from human interactions. In this way, the research community has collected datasets with multimodal information about two or more participants interacting with their labeled personality traits, see e.g., [3].

Most recent work has explored transformer architectures for simultaneous feature extraction and model prediction [1]. Such methods have achieved state-of-the-art performance, see, e.g., [4, 5]. Although very effective, it remains unanswered whether simpler models, e.g., RNNs are enough to approach the task. On the other hand, sequential information has barely been exploited by some of the best-performing models of this problem, see e.g., [4]. This paper aims to study the potential impact of explicitly modeling sequential information for analyzing personality using multimodal data by using RNNs. We consider features extracted from per-modality pretrained models at the utterance and fixed-length window level as the input for RNN architectures. These models then learn from sequential information to predict big-five personality traits. We use Autokeras for hyperparameter tuning and compare our results with state-of-the-art and baseline models using the same dataset. We show that the proposed solution achieves competitive performance compared to more complex models. Also, we found that our solution shows a regression to the mean problem.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

## 2 Related work

Within affective computing, a golden standard by now in the personality recognition task has been the Big Five personality traits theory [6]. The acronym OCEAN represents one of the personality traits in each letter. These are Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Using this metric for personality, many datasets gather information from distinct modalities. For example, FriendsPersona [7] only collects information in transcript format, and SSPNet Speaker Personality [8] in audio recordings. Others do go far as to include information from several modalities, such as FirstImpressions [1]. Also, under the hypothesis that personality enhances by social interactions, some of these datasets use information derived from sessions of one or more participants, such as SEMAINE [9], MHHRI [10], and UDIVA [3], the latter being the dataset considered for experimentation in this work.

In the context of UDIVA, state-of-the-art models perform feature extraction from multiple modalities using pre-trained models (e.g., sBERT [11] for text, VGGish [12] for audio, R(2+1)D [13] for video). As for the trait-prediction model, Palmero et al. and Curto et al. implement a custom transformer-based architectures[5, 3], where they do obtain the representation of the data sequentially, including a cross-attention block between participants [3], and a cross-attention block between participants and modalities [5]. In the latter work, authors standardize the input in batches of sequences of information of 30 seconds. To generate more data from these sequences, they sample the information using sliding windows. Despite very effective, none of these works have simultaneously used text, audio, and visual modalities. On the other hand, the best existing model for the considered dataset uses Neural Architecture Search (NAS) and profiling of participants by age and gender to recognize personality [4]. Despite competitive, this model does not consider a sequential input, obtaining a single representation for each minute of information separately.

## 3 Multi-stream RNNs for personality recognition

This section describes the methodology adopted for approaching the personality recognition problem. Before that, we present the dataset considered for experimentation.

### 3.1 The UDIVA dataset

For experimentation, we used the UDIVAv0.5 dataset [3], released in the context of the 2021 Understanding Social Behavior in Dyadic and Small Group Interactions Challenge at ICCV [14]. UDIVA is the biggest dataset focusing on social interaction, with over 90.5 hours of recordings from 145 dyadic sessions. In addition, it is already partitioned into predefined training, validation, and testing sessions. The dataset contains audiovisual, transcriptions, and metadata information from interactions between two participants in four collaborative and competitive tasks. For the visual modality they already provide a set of automatically extracted annotations for the face, body, hands, and gaze, extracted with 3DDFAv2 [15], MeTRAbs [16], FrankMokap [17], and ETH-Gaze [18], respectively. It also includes transcriptions of the conversations manually annotated at utterance level, which are time synchronized to the video.

### 3.2 Personality recognition with sequential models

Following the intuition that in interactions, the temporal dependencies between *actions*[1] ocurring in a temporal window define a person's behavior and, therefore, their personality, we built a representation for each participant as a sequence of actions. The sequences are then modeled with RNN architectures to predict the personality traits of users.

Actions were represented with a vector summarizing the information ocurring in a time window $w_L$. We adopted two approaches for defining the length of the window: (1) we used 3-second windows, as in previous work [5]; and (2) we used user-interventions (utterances) as a window. We argue that the latter definition is more intuitive and helpful, we show its effectiveness in the next section. Multimodal information was extracted from videos in each window $w_L$. For textual transcriptions, we first obtained the embedding representation for each word in the vocabulary using Fasttext [19]

[1] In this work we use the term action to refer to the behavior of an user in a window of time.

($d_t = 300$) pretrained on the Spanish Billion Word Corpus [20]. Then, we compute the average of word vectors in $w_L$ to obtain the textual representation for the window. For audio, we used VGGIsh [12] to encode the audio signal within $w_L$ in a feature vector. In the case of the video, we used the landmarks information provided with the dataset representations, which are sets of 3D coordinates per frame. We estimated the mean and the standard deviation of the landmarks over $w_L$, and concatenated them to get a representation for the window.

To generate training samples for the predictive model, we defined the timestep $T$, indicating the number of sequential windows $w_L$ that are feed to RNNs under a many-to-one architecture. We utilize one typical technique in recurrent neural networks: sliding the window. The stride $s$ we chose for memory constraints was 9 seconds, and 3 utterances, respectively.

As base RNN architecture, we adopted a multi-stream neural network, where the number of streams (1-3) depends on the number of modalities (i.e., text, audio, and video) involved in a particular architecture (we performed experiments with all possible combinations of modalities). As depicted in Figure 1, a sequence of features for the involved modalities is fed into sequential models (one per modality); the outputs of the RNNs are then concatenated and fed into a dense layer with five units (one per personality trait) that fuses information. Finally, the model is trained to minimize the MSE for predicting the five personality traits. AutoKeras [21] was used to optimize the architecture and hyperparameters of the RNNs we consider. For the inference/test phase, we feed all of the samples associated to each user into the trained model, and the predictions for all of the samples are averaged to obtain a single prediction of the five traits for each user.

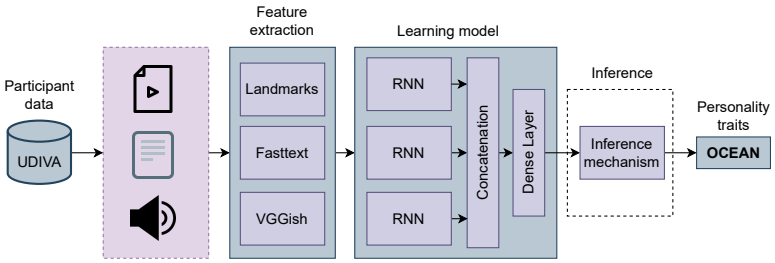



Figure 1: Proposed architecture for a multimodal approach

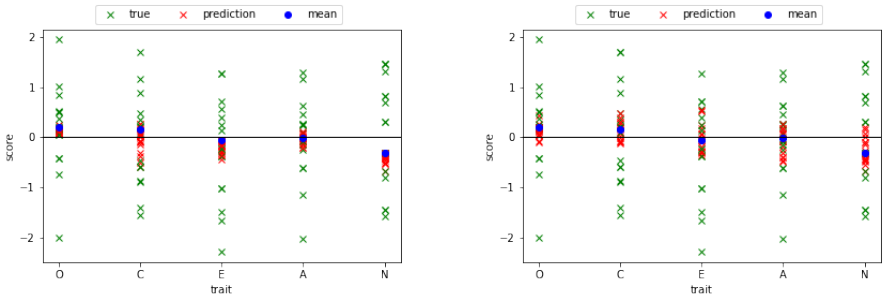
## 4 Experimental results

For evaluation we used the per-trait mean squared error ($MSE$), and the average MSE across traits ($AMSE$), these metrics have been also used in related work (e.g., [3, 4, 5]). For the 3-second windows setting we considered $w_L = 3$ seconds, a stride of $s = 9$ seconds, and timesteps $T = 10$. For the utterance level setting we used $w_L = 1-$utterance, a stride of $s = 3$ utterances, and $T = 10$. We evaluated the two variants of the proposed model with all of the possible combinations of modalities. A comparison between the proposed models and state-of-the-art methods using the same dataset is shown in Table 1.

Better results with our model are obtained with windows at the utterance level. The best AMSE we obtained was of $0.755$ which outperforms the baseline in [3], and achieves similar performance as the model in [5]. Therefore the proposed model is competitive, despite it is based on less complex models as those reported in [3, 4, 5]. Interestingly, variants of our model achieve the best overall result for the *Extraversion* and *Agreeableness* traits. It is also interesting that the best model in terms of AMSE in each window variant is one that includes the text and audio modalities ($T + A$). Contrasting the way that windows were defined, the variant based on utterances was more effective. Interestingly, this intuitive way of splitting sequential information was not used by authors of [3, 4, 5]. To further analyze the results, we plot the outputs of our two best models in Figure 2 where we plot the ground truth values for the test set (green), the mean of the training set ground truth (blue), and compare the dispersion of the predictions (red) after the aggregation step in each case. It can be noticed that variance is higher in the model which uses utterance level representations than the one that uses time level representations. This indicates that the sliding window has a smoothing effect in $w_L = 3$ that is avoided by working with utterances.

Table 1: Experimental results and comparison with state-of-the-art. Column 1 shows the architecture selected with AutoKeras in our proposed method. Modalities are coded as follows: $T$ = text, $A$ = audio, $I$ = image.

| Literature models | O | C | E | A | N | AMSE |
|---|---|---|---|---|---|---|
| GW bimodal NAS [4] | **0.684** | **0.588** | 0.830 | 0.550 | **0.796** | **0.690** |
| Dyadformer [5] | - | - | - | - | - | 0.722 |
| Transformer [3] | 0.744 | 0.794 | 0.886 | 0.653 | 1.012 | 0.818 |
| Proposed model ($w_L = 3$) | | | | | | |
| LSTM$_I$ | 0.726 | 0.830 | 0.956 | 0.674 | 1.220 | 0.881 |
| LSTM$_T$ | 0.740 | 0.814 | 0.942 | 0.677 | 1.226 | 0.888 |
| LSTM$_A$ | 0.724 | 0.849 | 0.915 | 0.664 | 1.163 | 0.863 |
| GRU$_{I+T}$ | 0.721 | 0.720 | **0.802** | 0.643 | 1.291 | 0.836 |
| GRU$_{I+A}$ | 0.766 | 0.765 | 0.945 | 0.661 | 1.145 | 0.856 |
| GRU$_{T+A}$ | 0.718 | 0.737 | 0.859 | 0.605 | 1.109 | 0.805 |
| LSTM$_{T+I+A}$ | 0.737 | 0.869 | 0.955 | 0.662 | 1.156 | 0.888 |
| Proposed model ($w_L$ = utterance) | | | | | | |
| GRU$_I$ | 0.724 | 0.848 | 1.004 | 0.663 | 1.146 | 0.877 |
| GRU$_T$ | 0.738 | 0.828 | 0.983 | 0.677 | 1.130 | 0.871 |
| GRU$_A$ | 0.730 | 0.790 | 0.876 | **0.525** | 0.989 | 0.782 |
| GRU$_{I+T}$ | 0.737 | 0.816 | 0.883 | 0.632 | 1.171 | 0.848 |
| GRU$_{I+A}$ | 0.749 | 0.788 | 0.883 | 0.588 | 1.015 | 0.804 |
| GRU$_{T+A}$ | 0.777 | 0.714 | 0.855 | 0.526 | 0.905 | 0.755 |
| LSTM$_{T+I+A}$ | 0.749 | 0.726 | 0.886 | 0.598 | 1.112 | 0.814 |




(a) 3 second window (b) Utterance

Figure 2: Comparison between the predicted results for all participants and the ground truth values with our best model (3-second level representation)

## 5 Conclusions

We have presented a methodology for the recognition of personality from multimodal data that effectively exploits sequential information. The proposed model resulted very competitive, despite being not as complex as alternative solutions. In fact, our biggest multimodal model has 9.1M parameters compared to 36M from [5]. We found that the inference step could be causing the regression to the mean problem, in our work and possibly in state-of-the-art approaches (it is also mentioned in [5], and we verified it is present in the results from [4]). The proposed methodology can be improved in several ways. For instance, currently it disregards information related to the other participant's interactions. Also, it could be improved by using other representations such as Action Units for the visual modality, and other types of embeddings for the textual one.

## Acknowledgements

This work was supported by CONACyT under grant CB-S-26314.

## References


[1] Julio C. S. Jacques Junior et al. *First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis*. 2018. DOI: 10.48550/ARXIV.1804.08046. URL: https://arxiv.org/abs/1804.08046.

[2] Hugo Jair Escalante et al. "Modeling, Recognizing, and Explaining Apparent Personality From Videos". In: *IEEE Trans. Affect. Comput.* 13.2 (2022), pp. 894–911. DOI: 10.1109/TAFFC.2020.2973984. URL: https://doi.org/10.1109/TAFFC.2020.2973984.

[3] Cristina Palmero et al. *Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset*. 2020. DOI: 10.48550/ARXIV.2012.14259. URL: https://arxiv.org/abs/2012.14259.

[4] Hanan Salam et al. "Learning Personalised Models for Automatic Self-Reported Personality Recognition". In: *Understanding Social Behavior in Dyadic and Small Group Interactions*. Ed. by Cristina Palmero et al. Vol. 173. Proceedings of Machine Learning Research. PMLR, 16 Oct 2022, pp. 53–73. URL: https://proceedings.mlr.press/v173/salam22a.html.

[5] David Curto et al. *Dyadformer: A Multi-modal Transformer for Long-Range Modeling of Dyadic Interactions*. 2021. DOI: 10.48550/ARXIV.2109.09487. URL: https://arxiv.org/abs/2109.09487.

[6] Lewis R. Goldberg. *The Structure of Phenotypic Personality Traits*. 1993.

[7] Hang Jiang, Xianzhe Zhang, and Jinho D. Choi. *Automatic Text-based Personality Recognition on Monologues and Multiparty Dialogues Using Attentive Networks and Contextual Embeddings*. 2019. DOI: 10.48550/ARXIV.1911.09304. URL: https://arxiv.org/abs/1911.09304.

[8] Anna Polychroniou, Hugues Salamin, and Alessandro Vinciarelli. "The SSPNet-Mobile Corpus: Social Signal Processing Over Mobile Phones." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 1492–1498. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/537_Paper.pdf.

[9] Gary McKeown et al. "The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent". In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 5–17. DOI: 10.1109/T-AFFC.2011.20.

[10] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. "Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement". In: *IEEE Transactions on Affective Computing* 10.4 (2019), pp. 484–497. DOI: 10.1109/TAFFC.2017.2737019.

[11] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. DOI: 10.48550/ARXIV.1908.10084. URL: https://arxiv.org/abs/1908.10084.

[12] Shawn Hershey et al. "CNN Architectures for Large-Scale Audio Classification". In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. URL: https://arxiv.org/abs/1609.09430.

[13] Du Tran et al. *A Closer Look at Spatiotemporal Convolutions for Action Recognition*. 2017. DOI: 10.48550/ARXIV.1711.11248. URL: https://arxiv.org/abs/1711.11248.

[14] Cristina Palmero et al. "ChaLearn LAP Challenges on Self-Reported Personality Recognition and Non-Verbal Behavior Forecasting During Social Dyadic Interactions: Dataset, Design, and Results". In: (Apr. 2022).

[15] Jianzhu Guo et al. *Towards Fast, Accurate and Stable 3D Dense Face Alignment*. 2020. DOI: 10.48550/ARXIV.2009.09960. URL: https://arxiv.org/abs/2009.09960.

[16] István Sárándi et al. "MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.1 (2021), pp. 16–30. DOI: 10.1109/TBIOM.2020.3037257.

[17] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. *FrankMocap: A Monocular 3D Whole-Body Pose Estimation System via Regression and Integration*. 2021. DOI: 10.48550/ARXIV.2108.06428. URL: https://arxiv.org/abs/2108.06428.

[18] Xucong Zhang et al. *ETH-XGaze: A Large Scale Dataset for Gaze Estimation under Extreme Head Pose and Gaze Variation*. 2020. DOI: 10.48550/ARXIV.2007.15837. URL: https://arxiv.org/abs/2007.15837.

[19] Piotr Bojanowski et al. *Enriching Word Vectors with Subword Information*. 2016. DOI: 10.48550/ARXIV.1607.04606. URL: https://arxiv.org/abs/1607.04606.

[20] Cristian Cardellino. *Spanish Billion Words Corpus and Embeddings*. Aug. 2019. URL: https://crscardellino.github.io/SBWCE/.

[21] Haifeng Jin, Qingquan Song, and Xia 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.