

Gated Siamese Fusion Network based on multimodal deep and hand-crafted features for personality traits assessment

Elena Ryumina ^{a,*}, Maxim Markitantov ^a, Dmitry Ryumin ^a, Alexey Karpov ^b

^a St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), 39, 14th Line V.O., St. Petersburg, 199178, Russia

^b ITMO University, Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia

ARTICLE INFO

Editor: Mauricio Pamplona Segundo

Dataset link: <https://chalearnlap.cvc.uab.cat/dataset/24/description/>, <https://oceanai.readthedocs.io/en/latest/>

MSC:

68T10

68T45

68U15

62-07

Keywords:

Deep learning

Multimodal paralinguistics

Multimodal gated fusion

Hand-crafted and deep features

Personality computing

Affective computing

ABSTRACT

People tend to judge others assessing their personality traits relying on life experience. This fact is especially evident when making an informed hiring decision, which should consider not only skills, but also match a company's values and culture. Based on this assumption, we use the Siamese Network (SN) for assessing five personality traits by pairwise analyzing and comparing people simultaneously. For this, we propose the OCEAN-AI framework based on Gated Siamese Fusion Network (GSFN), which comprises six modules and enables the fusion of hand-crafted and deep features across three modalities (video, audio, and text). We use the ChaLearn First Impressions v2 (Fiv2) and Multimodal Personality Traits Assessment (MuPTA) corpora and identify that all six feature sets and their combinations due to different information content allow the framework to adjust to heterogeneous input data flexibly. The experimental results show that the pairwise comparison of people with the same or different Personality Traits (PT) during the training enhances the proposed framework performance. The framework outperforms the State-of-the-Art (SOTA) systems based on three modalities (video-face, audio and text) by the relative value of 1.3% (0.928 vs. 0.916) in terms of the mean accuracy (mACC) on the Fiv2 corpus. We also outperform the SOTA system in terms of the Concordance Correlation Coefficient (CCC) by the relative value of 8.6% (0.667 vs. 0.614) using two modalities (video and audio) on the MuPTA corpus. We make our framework publicly available to integrate it into various applications such as recruitment, education, and healthcare.

1. Introduction

Personality Traits (PT) are stable patterns of thinking, feeling and behavior that define an individual's character. They play a crucial role in shaping human behavior, influencing people's interaction with each other and with machines [11]. The OCEAN model, also referred to as the Big Five model, is a widely used framework for Personality Traits Assessment (PTA) [12]. This model includes five PT: *Openness to experience (O)*, *Conscientiousness (C)*, *Extraversion (E)*, *Agreeableness (A)*, and *non-Neuroticism (N)*.

PTA is made according to the results of self-evaluation, familiar- or third-party-evaluation through questionnaires [13]. However, filling out questionnaires can sometimes be challenging. This requires human resource, is tedious, and can be subjective. Recent advances in machine learning have given rise to automatic PTA systems [10]. These systems are a vital component for optimizing human-machine interaction in education [14], medicine [15], marketing [16], and others. This research area belongs to a broader field of Affective Computing, which

focuses on recognizing, analyzing, and interpreting human behavior and affects [12].

Like a human, automatic PTA systems analyze multiple sources (modalities) of information about a person, including video, audio, and text [8]. However, unlike humans, they are able to assess an unlimited number of people in a short period of time without compromising the reliability of the assessment, and are subject to fewer biases, making them more accurate [11]. Apart from the use of multiple modalities, the choice of methods for their feature representation and aggregation makes the basis of a robust PTA system. The State-of-the-Art (SOTA) systems are mainly implemented based on deep features that are able to capture subtle and context-sensitive patterns of PT in various signals [8]. However, despite the effectiveness of these features, the results obtained are difficult to interpret. To address this challenge, it is advantageous to use hand-crafted features [17,18]. On the other hand, a comprehensive analysis to identify PT requires a combination of both deep and hand-crafted feature types from multiple modalities.

* Corresponding author.

E-mail addresses: ryumina_ev@mail.ru (E. Ryumina), m.markitantov@yandex.ru (M. Markitantov), ryumin.d@ias.spb.su (D. Ryumin), karpov@ias.spb.su (A. Karpov).

<https://doi.org/10.1016/j.patrec.2024.07.004>

Received 17 December 2023; Received in revised form 28 May 2024; Accepted 2 July 2024

Available online 8 July 2024

0167-8655/© 2024 Elsevier B.V. All rights reserved, including those for text and data mining, AI training, and similar technologies.

Table 1

Classification of SOTA systems. Fusion: FLF refers to feature-level fusion, SLF to score-level fusion, MLF to model-level fusion.

System	Segmentation	Facial features	Acoustic features	Linguistic features	Multimodal fusion	Code
Kaya et al. [1]	–	Functional statistics of Emotional VGG-Face and LGBP-TOP	openSMILE 2013	–	FLF: Kernel EL, SLF: Random Forest	Available
Li et al. [2]	32 frame segments, one frame in each segment	ResNet34 and CR-Block	Raw audio, ResNet34 and CR-Block	Skip-thought vector, ResNet34 and CR-Block	FLF: Extra Trees Regressor	Not available
Palmero et al. [3]	32 frame segments (2.5 sec) with stride of two frames	R(2+1)D and Spatio-temporal encodings	VGGish	–	FLF: Transformer	Not available
Aslan et al. [4]	One frame per sec	ResNetV2 101-LSTM	VGGish-LSTM	ELMo-FCNN	FLF: Attention module	Not available
Giritlioğlu et al. [5]	45 frame segments (1.5 sec)	ResNext101-GRU, OpenFace-LSTM	34d MFCC, Chroma, energy and LSTM	BERT-LSTM	SLF: Linear Support Vector Regressor	Not available
Agrawal et al. [6]	32 frame segments (2.5 sec) with stride of two frames	R(2+1)D and Spatio-temporal encodings	VGGish	BERT	FLF: Transformer-LSTM	Not available
Suman et al. [7]	Six frames at equal clip intervals	ResNet101	VGGish-LSTM	–	SLF: Averaging final scores	Not available
Agrawal et al. [8]	–	R(2+1)D and Video Swin Transformer	Trill-Distilled	XLNet-RoBERTa	MLF: Fat Transformer Cross-Attention	Not available
Cabada et al. [9]	30 frames at equal clip intervals	Scene with facial landmark delineation and C-CNN based on ResNet	DSCC-LSTM	–	MLF: Feature concatenation and FCNN	Not available
Gan et al. [10]	–	Scene and Vision Transformer	–	Personality descriptors and BERT	MLF: CLIP, SLF: Stacked Ensemble	Not available
Proposed framework	Two sec segments with one sec step	Emotional ResNet50-LSTM, geometric features-LSTM	Emotional VGG16-FCNN, eGeMAPS-LSTM	BERT-BiLSTM, LIWC-ReBiLSTM	FLF: Gated Siamese Fusion Network	Available

Another limitation of the SOTA systems is analyzing data about a single person. Using information about the interlocutor in a dialog as additional data is proven effective [3,19]. In this study, we propose to analyze data of two persons using Siamese Network (SN), without considering their direct (dialogical) interaction. SN enables training an effective model by creating a feature space in which similar classes are the closest and different classes are the farthest [20]. This is achieved because SN is a dual neural network with the same weights, working in tandem on two different input batches to compare the output feature vectors. SN has previously been used for emotion recognition [20], object tracking [21], and image classification [22]. Here, we first propose to use SN for PTA. The implementation of such models in automatic PTA systems endows artificial intelligence with the human ability to predict PT of a person by comparing them to other people.

The main contributions of the article are as follows:

- We propose the novel feature fusion strategy of multimodal data based on Gated Siamese Fusion Network (GSFN).
- We utilize multimodal deep and hand-crafted features and compare their performance.
- We outperform the SOTA systems developed on two corpora and using three modalities (video-face, audio and text).
- We provide an open-source framework¹ to integrate it into human-machine applications. It simplifies the decision-making process by assisting with such tasks as recruiting, predicting consumer preferences, facilitating educational trajectories, and monitoring mental health deviations.

The remainder of this article is organized as follows. In Section 2, we analyze the existing multimodal SOTA systems for PTA. Section 3 outlines our methodology. The research corpora, our experimental setup and results, as well as their discussion are presented in Section 4. Finally, in Section 5, we summarize the study and consider potential future research.

2. Related work

In this section, we present an overview of the SOTA systems for PTA. The SOTA systems analyze PT using several modalities: video (face, scene, behavior encoding), audio, text, and metadata. We conduct our study on two corpora that differ in metadata and video-scene conditions (office and “in-the-wild”). These two modalities limit the possibility of developing a universal PTA system in the context of heterogeneous data. Therefore, we review only three modalities (video-face, audio, and text).

The classification of the SOTA systems (see Table 1) shows the following main trends. The authors use frame downsampling [4,7] and segmentation [2,3,6] to reduce the computational costs of visual models. Convolutional models of the ResNet family [5,9] are used to extract facial features. A pre-trained VGGish model (for recognizing acoustic events) [3,4,7] is used to extract acoustic deep features from log-Mel spectrograms. Linguistic features are predominantly extracted using the BERT model [5,6]. Modality fusion is performed at the feature-level using conventional methods [1,2] or neural networks with a scaled dot-product attention mechanism [3,4,6]. More recent systems additionally use the video-scene modality; Cabada et al. [9] draw facial landmarks on the scene frames, and Gan et al. [10] utilize a textual description of the scene. In addition, they both use model-level fusion. All of the above solutions improve PTA. However, these SOTA systems are not accessible, making it difficult to use them for PTA in many specific applications.

In view of the current trends, we develop the proposed framework with the following improvements. We use the emotional features extracted using ResNet50 and VGG16 [13] for video-face and audio modalities, respectively. Previous studies have shown that emotional deep features improve the PTA performance compared to deep features trained on other tasks [1,23]. We extract hand-crafted features for each modality, and our experimental results show that they are as robust as deep features and significantly contribute to modality feature-level fusion. We also use a neural network with an attention mechanism to fuse modalities, particularly gated attention, which is more efficient than scaled dot-product attention. We propose a SN for pairwise analysis of

¹ <https://oceanai.readthedocs.io/en/latest/>

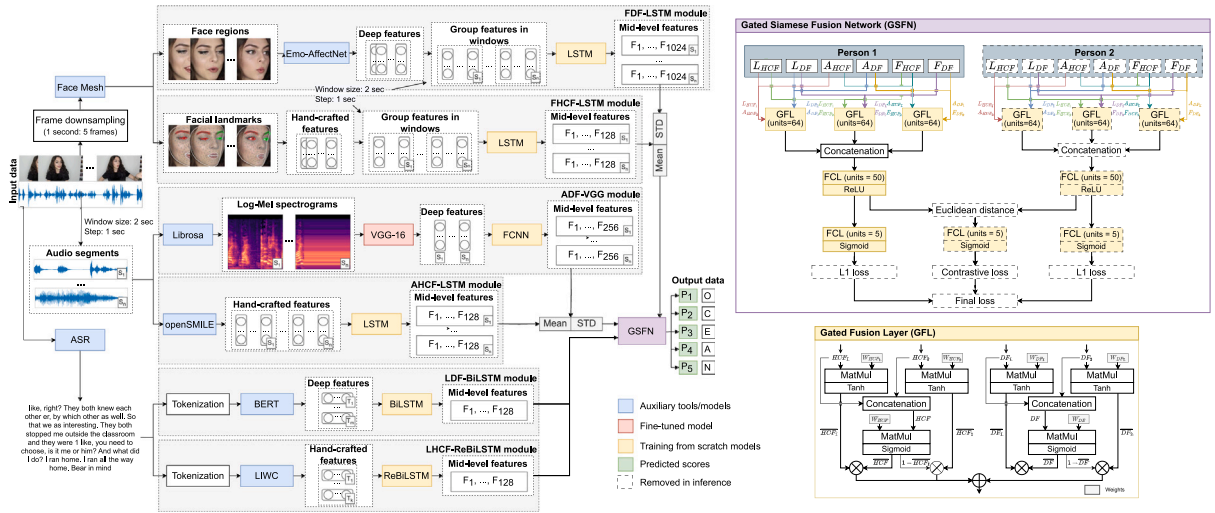


Fig. 1. The proposed framework for PTA. S_i is an audio/video segment, $i = 1, \dots, n$ means the number of 2-sec segments with 1-sec step in an audio-visual signal. T_j is a word/character, $j = 1, \dots, k$ means the number of words/characters in a sentence. F is a feature vector. (L_{DF}, L_{HCF}) , (A_{DF}, A_{HCF}) , and (F_{DF}, F_{HCF}) denote deep and hand-crafted linguistic, acoustic and facial features, respectively.

people, which has not been done before. Finally, we share the source code of the proposed framework.

3. Methodology

The pipeline of the proposed framework for PTA is shown in Fig. 1. The framework comprises video, audio, and text modules. Mid-level deep features for video and audio modules are extracted from 2-sec segments with 1-sec step. For the features of these modules, we compute the mean and standard deviation (STD) values for the whole clip. Then, we combine the values and linguistic features as input for GSFN.

3.1. Video-based PTA

We employ the Face Mesh model [24] from the MediaPipe library. This model allows detecting 468 3D facial landmarks, which is a significant number, and extracting face regions. Since the Frames Per Second (FPS) of all clips are different, each clip is downsampled to five FPS to keep the same processing conditions for Long Short-Term Memory (LSTM) networks. For a 2-sec segment, the number of frames equals ten. The last frame is repeated as many times as necessary if the segment is shorter than two sec.

Deep features. As known, PT are determined by the sequence of emotional and behavioral reactions of different people to the same stimuli [12]. Inspired by this fact and the research [1], we apply the open-source Emo-AffectNet model [13] to extract 512 emotional deep features of a face. This model proved its performance in the task of emotion recognition. The feature size for a 2-sec segment is 10×512 . We use a single-layer LSTM with 1024 units for PT modeling based on facial deep features. We call this module as FDF-LSTM.

Hand-crafted features. According to three main approaches for recognizing psychological characteristics by the face (physiognomy, phase facial portrait, and ophthalmo-geometry) [25], we develop our own set of facial features (see Fig. 2) including: (1) three angles characterizing asymmetry of the face by eyebrows and eyes centers, and lip corners; (2) 36 distances between facial landmarks; (3) 38 pairs of facial landmarks coordinates. The feature size for a 2-sec segment is 10×115 . We apply the Z-normalization to the features obtained. We use LSTM consisting of two layers with 64 and 128 units for PT modeling based on hand-crafted features. We call this module as FHC-LSTM.

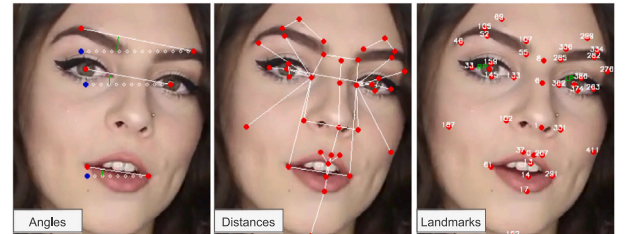


Fig. 2. Facial hand-crafted features.

3.2. Audio-based PTA

Deep features. The log-Mel spectrograms with 128 Mel filter-banks with a window length of 2048 and a step of 512 are extracted from each clip. The feature size for a 2-sec segment is 128×173 . The features obtained are padded with the mean value (if the audio segment is shorter than two sec), converted into images, resized to 224×224 pixels and repeated three times. Therefore, the new size of the features is $224 \times 224 \times 3$. We apply the min-max normalization to the features obtained. To extract deep features from log-Mel spectrograms, we utilize the emotional VGG-16 model [13], which outperformed other models in the task of escalation prediction from speech. Fully Connected Neural Network (FCNN) consisting of two Fully Connected Layers (FCL) with 512 and 256 units is used for PT modeling. We call this module as ADF-VGG.

Hand-crafted features. We use the eGeMAPS feature set extracted using the openSMILE library [26]. This feature set characterizes affective-physiological changes in speech and contains 25 low-level descriptors, including voice-related (pitch, formant frequency, harmonics-to-noise ratio, jitter, shimmer), energy-related (loudness), and spectral descriptors (α -ratio, spectral slope, Mel-frequency cepstral coefficients 1–4, and so on). These descriptors are extracted every 20 ms with a step of 10 ms. The feature size for a 2-sec segment is 196×25 . The features obtained are zero-padded (if the segment is shorter than two sec). We apply the L2 normalization to them. We use the model architecture similar to the video module (VHCF-LSTM) as it appears to be effective for acoustic hand-crafted features as well. We call this module as AHCF-LSTM.

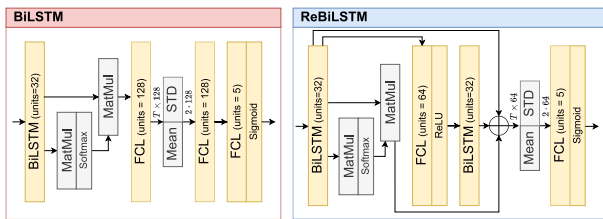


Fig. 3. Linguistic model architectures. T is the number of words/characters in a sentence.

3.3. Text-based PTA

Processing linguistic information depends on research corpora because they are multilingual. To train models for analyzing linguistic information, we use transcriptions suggested by the authors of research corpora. At the inference stage, we utilize the multilingual Whisper model² for Automatic Speech Recognition (ASR). We report on performance measures of our modules with the Whisper model in Section 4. We use features extracted from penultimate layers of the text modules for PT modeling based on linguistic features.

Deep features. We extract linguistic deep features with the multilingual BERT model. The BERT model³ produces features whose size depends on the number of characters in a sentence. The maximum feature size is 104×768 for the Fiv2 corpus and 414×768 for the Multimodal Personality Traits Assessment (MuPTA) corpus. If the feature size is smaller, it is zero-padded up to the maximum length. The Bidirectional LSTM (BiLSTM) model with dot-product self-attention [27] is used to analyze linguistic deep features. The model architecture is presented in Fig. 3. We call this module as LDF-BiLSTM.

Hand-crafted features. To obtain linguistic hand-crafted features, we use Linguistic Inquiry and Word Count (LIWC) [28], which is a dictionary-based tool for psychological analysis of English text. It is also used to relate linguistic features with PT [18]. In LIWC, each word is associated with 64 categories, including work (words such as job, majors, xerox), perceptual processes (observing, hearing, feeling), cognitive processes (cause, know, ought), anxiety (worried, fearful, nervous), and so on. Each word can belong to more than one category. We split each sentence into words and assign each word features with a size of 1×64 , which corresponds to 64 categories in LIWC. For Russian text, we employ the OPUS-MT model [29]⁴ to translate it into English. The maximum feature size is 89×64 for the Fiv2 corpus and 365×64 for the MuPTA corpus. If the feature size is smaller, it is zero-padded up to the maximum length. The Residual BiLSTM (ReBiLSTM) model with dot-product self-attention is used to analyze linguistic hand-crafted features. The model architecture is presented in Fig. 3. We call this module as LHCF-ReBiLSTM.

3.4. Multimodal-based PTA

We propose a novel GSFN to combine information from different feature sets of three modalities (see Fig. 1). Two key ideas are the foundation for this network: (1) PTA of one person is more accurate during pairwise comparisons, and (2) gated attention [30], which is based on optimized weights, the reset and update gates, and outperforms a scaled dot-product attention [27]. In Section 4, we compare scaled dot-product and gated attention as fusion strategies.

The SOTA systems (see Table 1) use a conventional learning approach. A model receives a batch of input data for one step and updates its weights based on the cumulative error (loss function) by all the

batches. Minimizing the error is the primary goal of this learning approach. In contrast, GSFN receives two batches of input data, calculates a distance measure between them and the error for each batch. This learning approach has two goals. Firstly, it is to optimize closely located feature representations by identifying similarities or differences in PT patterns between pairs of people. Secondly, it is to minimize the sum of three errors. Thus, using GSFN does not only achieve high performance in predicting PT scores, but based on multiple modalities also constructs for each person a feature space that would be optimal for all five traits. The idea of comparing two people is inherent in human nature; we involuntarily assess the PT of another person by relying on our life experience. Therefore, we place the model in the conditions where it is human-like and attempts to learn PT patterns by comparing multiple people.

GSFN consists of two identical neural networks with the same weights and architecture during training. At the inference stage, one of the symmetric networks is omitted. The mid-level facial and acoustic features are represented as matrices, while linguistic features are represented as vectors. Therefore, we calculate the mean and STD values for the first two feature sets. We use all these features for fusion. The gated attention is then used to extract the final feature vectors. These feature vectors are used as inputs to two Siamese regressions to predict the PT scores, and a multi-label classifier to predict the similarity of the two persons' traits. For the latter task, we employ Euclidean distance that measures the dissimilarity of the final feature vectors. It is then fed into the FCL consisting of five units for multi-label classification (five binary classes for five traits). In addition, we convert the ground truth scores into binary vectors according to the following rule. We define three groups with the PT scores: ranging from 0 to 0.4, from 0.4 to 0.6, and from 0.6 to 1, respectively. If the scores for the same trait for both persons are not in the same group, the trait is considered different (class 0), otherwise it is considered similar (class 1). Therefore, to determine whether PT of two persons are similar, we compare the result of Euclidean distance processed by FCL and the prepared ground truth scores.

4. Experiments

4.1. Research corpora

There are several corpora for PTA with a wide range of informants, languages, and rich metadata. Their detailed description and comparison are presented in [13,23]. In this study, we use two corpora for PTA modeling: Fiv2 [12] and MuPTA [13].

Fiv2 is a widely used corpus for PTA, consisting of recordings (mostly "in-the-wild") of about 3K English speakers. The corpus was labeled by about 3K third-party annotators through pairwise analysis. The total number of pairs is over 300K. The number of pairwise combinations with fixed clip (PCFC) is ranged from 54 to 85 times. The corpus was divided into three subsets [12]: Train ($\approx 10K$ utterances), Development ($\approx 2K$ utterances) and Test ($\approx 2K$ utterances).

MuPTA is a unique multimodal corpus consisting of spontaneous and read speech of 30 native Russian speakers. The corpus was collected in office conditions. The PT annotation was performed by self-assessment. The corpus was divided into three subsets [13]: Train (18 speakers, $\approx 2K$ utterances), Development (6 speakers, $\approx 1K$ utterances) and Test (6 speakers, $\approx 1K$ utterances). A brief description of the research corpora is presented in Table 2.

4.2. Experimental setup

We conduct the experiments in several steps. Firstly, we train the models of all six modules independently. Then we use the trained models to extract the mid-level features and perform their aggregation using two fusion networks. These latter are trained with different numbers of PCFC. For the MuPTA corpus, we generate PCFC randomly,

² <https://huggingface.co/openai/whisper-base>

³ <https://huggingface.co/bert-base-multilingual-cased>

⁴ <https://huggingface.co/Helsinki-NLP/opus-mt-ru-en>

Table 2
Summary of research corpora. Dur. is the data duration.

Corpus	# Subjects	# Males/Females	Age (Range/Mean)	Dur., h
Fiv2 [12]	3060	1312/1748	[8, 62]/24	41
MuPTA [13]	30	15/15	[19, 86]/41	7

while for the Fiv2 corpus, we use PCFC produced by the annotators, however, they are also randomly sampled. All model architectures and their corresponding parameters are optimized during the training using grid search. We show only the high-performance models and their corresponding parameters. All the models are trained using the Adam optimizer for 100 epochs and the Cosine Annealing Learning Schedule with 5-rate restart cycles. All experimental results are presented for the Test subsets of the research corpora.

For the training we use the contrastive and L1 loss functions, which are calculated according to the following formulas:

$$L_{contrastive} = \frac{1}{N} \sum_{i=1}^N (1 - t_i) \cdot p_i^2 + t_i \cdot (1 - p_i)^2, \quad (1)$$

$$L1_k = \frac{\frac{1}{N} \sum_{i=1}^N |t_i - p_i|}{\sum_{i=1}^N |t_i - \bar{t}|}, \quad (2)$$

where N denotes the number of clips; t_i and p_i mean the ground truth and predicted scores of clip i , respectively; \bar{t} is the average ground truth scores over all the clips; $k = \{1, 2\}$ means the person's number in a pair. Sum $L_{total} = L_{contrastive} + L1_1 + L1_2$ is the total loss. We evaluate the proposed modules/framework for PTA using two performance measures: Accuracy (ACC) [12], and Concordance Correlation Coefficient (CCC) [13]. ACC shows the error between the predicted and ground truth scores, while CCC indicates the correlation between them. The performance measures are calculated using the following formulas:

$$ACC^j = 1 - \frac{1}{N} \sum_{i=1}^N |t_i^j - p_i^j|, \quad (3)$$

$$CCC = \frac{2 \cdot \sigma_{t,p}}{\sigma_t^2 + \sigma_p^2 + (\bar{t} - \bar{p})^2}, \quad (4)$$

where \bar{p} denotes the averaged predicted scores for all the clips, respectively; σ_t and σ_p are the respective STD values; $\sigma_{t,p}$ is the covariance between t and p . ACC^j is calculated for one trait j . $mACC$ is the mean of all ACC^j measures. To calculate the measures of the video and audio modules, we average the scores of all the clip segments. The scores are extracted from regression layers with five units and a linear activation function.

4.3. Experimental results

The experimental results of the proposed modules and fusion strategies are presented in Tables 3 and 4 on the Fiv2 and MuPTA corpora, respectively. On both corpora, the video modules outperform the other two. In the case of the Fiv2 corpus, the audio modules outperform the text modules, especially in terms of CCC, whereas the opposite is true for MuPTA. The use of Whisper to obtain transcriptions results in a slight performance degradation of the text modules and the final framework. The fusion of all modules using gated attention outperforms scaled dot-product attention. GSFN increases CCC and the mean ACC by 5.5% and 0.9%, respectively, on Fiv2. In the case of MuPTA, GSFN shows a lower performance in terms of CCC (−0.6%) and a greater one in terms of the mean ACC (+0.8%). These results show that the modalities for PTA are informative depending on the research data. The feature fusion of multiple modalities provides a comprehensive analysis of a person, resulting in accurate prediction of PT scores. Among other things, the results confirm previous findings that SN creates a feature space, in which similar trait scores are the closest to each other, and the different ones are the farthest from each other. Training models to compare two people increases the prediction accuracy of the PT scores

Table 3

Experimental results for evaluation of the proposed modules and fusion strategies on the Fiv2 corpus. SDPA refers to Scaled Dot-Product Attention, GFN to Gated Fusion Network, GSFN to Gated Siamese Fusion Network.

Module	O	C	E	A	N	mACC	CCC
FDL-LSTM	.912	.916	.917	.912	.910	.913	.647
FHCF-LSTM	.908	.906	.912	.909	.906	.908	.564
ADF-VGG	.907	.905	.905	.907	.903	.906	.523
AHCF-LSTM	.900	.891	.896	.902	.895	.897	.382
LDF-BiLSTM	.890	.885	.884	.899	.885	.889	.260
LCHF-ReBiLSTM	.889	.882	.884	.898	.885	.888	.251
SDPA [27]	.917	.921	.923	.923	.919	.919	.689
GFN (ours)	.918	.922	.925	.917	.919	.920	.696
GSFN (ours)	.925	.930	.932	.926	.928	.928	.734
LDF-BiLSTM + ASR	.890	.886	.885	.897	.883	.888	.264
LCHF-ReBiLSTM + ASR	.888	.882	.882	.896	.884	.886	.253
GSFN (ours) + ASR	.924	.927	.931	.925	.927	.927	.727

Table 4

Experimental results for evaluation of the proposed modules and fusion strategies on the MuPTA corpus.

Module	O	C	E	A	N	mACC	CCC
FDL-LSTM	.939	.927	.855	.891	.866	.896	.561
FHCF-LSTM	.934	.917	.857	.888	.867	.893	.521
ADF-VGG	.938	.925	.830	.890	.877	.892	.473
AHCF-LSTM	.936	.924	.823	.902	.870	.891	.460
LDF-BiLSTM	.935	.923	.823	.896	.877	.891	.485
LCHF-ReBiLSTM	.938	.922	.820	.897	.875	.890	.466
SDPA [27]	.930	.916	.864	.899	.878	.897	.627
GFN (ours)	.947	.940	.842	.901	.868	.900	.641
GSFN (ours)	.953	.938	.861	.904	.880	.907	.637
LDF-BiLSTM + ASR	.935	.922	.823	.895	.877	.890	.483
LCHF-ReBiLSTM + ASR	.938	.922	.820	.897	.875	.890	.464
GSFN (ours) + ASR	.953	.938	.861	.904	.880	.907	.636

Table 5

Comparison of the GSFN performance on the Fiv2 corpus depending on the maximum number of pairwise combinations with fixed clip (PCFC) used for training.

PCFC	1	2	3	4	5	10	15
mACC	.921	.928	.924	.925	.923	.919	.919
CCC	.707	.734	.721	.706	.688	.656	.645

of both of them, making such models more robust than those trained using a conventional learning approach.

For both corpora, we investigate the influence of the maximum number of PCFC used for training on the performance measures. In the case of MuPTA, the performance decreases with the increasing number of PCFC. While the optimal number of PCFC equals two for the Fiv2 corpus (see Table 5). Thus, training models using pairwise analysis of batches of input data leads to the enhanced performance by applying a similarity measure to feature spaces. However, the number of PCFC greater than two is not an effective choice.

Comparison of the GSFN performance on the Fiv2 and MuPTA corpora depending on the feature sets is shown in Table 6. The deep features of all modalities demonstrate a high contribution to the final framework on both corpora. Omitting these features results in greater performance degradation compared to the exclusion of hand-crafted features. In the case of Fiv2, omitting all facial features results in a greater decrease in performance compared to other features. In the case of MuPTA, omitting facial hand-crafted features leads to a greater performance degradation compared to acoustic hand-crafted features, the opposite is true for facial and acoustic deep features. Omitting all linguistic features or hand-crafted features leads to an increase in terms

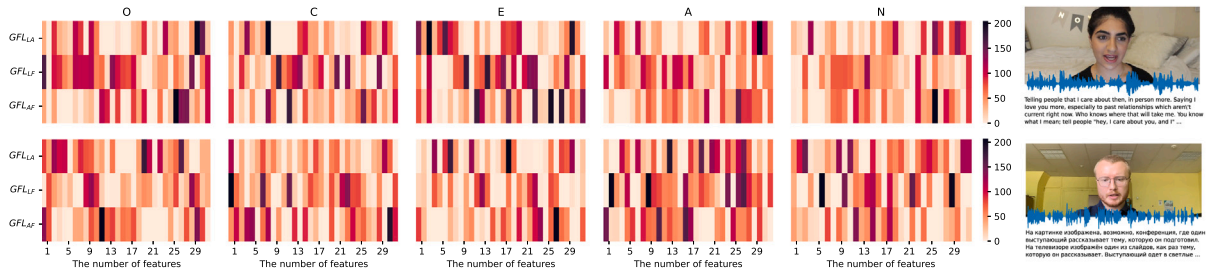


Fig. 4. Examples of heatmaps for clips from the Flv2 (top) and MuPTA (bottom) corpora.

Table 6

Comparison of the GSFN performance on the Flv2 and MuPTA corpora depending on the feature sets. F_F , A_F and L_F are without (w/o) all facial, acoustic and linguistic features, respectively.

w/o	Video-face			Audio			Text			All_F
	F_F	F_{DF}	F_{HCF}	A_F	A_{DF}	A_{HCF}	L_F	L_{DF}	L_{HCF}	
Flv2										
mACC	.890	.901	.912	.900	.893	.914	.916	.917	.916	.928
CCC	.276	.431	.556	.470	.457	.599	.645	.637	.648	.734
MuPTA										
mACC	.849	.865	.889	.875	.868	.897	.895	.904	.897	.907
CCC	.400	.499	.540	.394	.375	.627	.667	.608	.673	.637

Table 7

Comparison of the proposed framework with the SOTA systems. Modalities: F refers to video-face, S to video-scene, A to audio, T to text, BE to video-behavior encoding, M to metadata, FLD to facial landmark delineation, PE to personality encoding.

System	Modalities	O	C	E	A	N	mACC
Suman et al. [7]	F, S, A	.911	.921	.915	.913	.912	.914
Aslan et al. [4]	F, A, T	–	–	–	–	–	.916
Giritlioğlu et al. [5]	F, A, T	.913	.918	.917	.913	.914	.915
Kaya et al. [1]	F, S, A	.917	.920	.921	.914	.915	.917
Li et al. [2]	F, S, A, T	.920	.922	.920	.918	.915	.919
Gan et al. [10]	S, PE	.919	.917	.927	.924	.917	.921
Agrawal et al. [6]	F, S, A, T, M, BE	.929	.926	.927	.929	.921	.926
GSFN (ours)	F, A, T	.925	.930	.932	.926	.928	.928
Cabada et al. [9]	FLD, S	–	–	–	–	–	.942
Agrawal et al. [8]	F, S, A, T, M	.942	.951	.955	.949	.959	.951

of CCC on the MuPTA corpus. These differences highlight the need to consider cross-modal information in PTA, as the modality contribution varies across the research corpora. This cross-modal fusion allows GSFN to compensate for the weaknesses of each feature set of all modalities.

We provide heatmaps (see Fig. 4) to identify meaningful feature combinations of all modalities for assessing certain PT. For the sample clip from the Flv2 corpus (top row), the fusion of facial and linguistic features (GFL_{LF}) is more informative than other combinations (GFL_{LA} and GFL_{AF}) for the O, C, and A traits. Two fusions based on facial features are informative for E, the fusion of facial and acoustic features is also informative for the N trait. A different case is observed for the sample clip from the MuPTA corpus (bottom row). Two fusions based on facial features are informative for C. The fusion of facial and linguistic features is informative for A. Acoustic and linguistic feature fusion is more informative for O and N. Finally, the fusion of these features is as informative as facial and acoustic feature fusion for the E trait. Therefore, our model flexibly adapts to the input data, since different modalities predominate for certain PT.

4.4. Discussion

The experimental results demonstrate the effectiveness of the proposed multimodal GSFN for PTA. The Siamese architecture of this model, similar to human nature, predicts PT scores based on the similarities or differences between two people. The applied gated attention outperforms the well-known and amply used scaled dot-product

attention, on both the Flv2 and MuPTA corpora. In addition, this model, based on the aggregation of cross-modal information and heterogeneous features, provides a comprehensive analysis of human behavior, resulting in increased PTA performance.

The deep features of all modalities demonstrate the highest contribution to the PTA performance of the final framework on both corpora. However, the exclusion of hand-crafted features results in performance degradation. This suggests that the SOTA systems presented in Table 1 can be improved by including hand-crafted features, as they rely solely on deep features. The heatmaps visualization provides insights into the importance of all cross-modal feature combinations for assessing certain PT.

Comparison with the SOTA systems (see Table 7) shows that our framework outperforms the existing ones, including those based on three modalities (video-face, audio and text) and those utilizing more than three modalities, except for the [8,9] systems. These systems were trained only on data from one corpus with a focus on the video-scene modality, and because of this, their application to other corpora is limited. We also outperform the SOTA system in terms of CCC by the relative value of 8.6% (0.667 vs. 0.614) using two modalities (video and audio) on the MuPTA corpus [13]. We present the baseline results for the framework that uses three modalities – video, audio, and text.

In addition to high performance, the proposed framework, unlike the SOTA systems, is publicly available and works in real-time, with a 1-sec clip processing time of 0.25 s on the GPU (using a GeForce RTX 3080) and 0.6 s on the CPU (using an Intel i9).

The PT scores predicted by the proposed open-source framework can be used in various applications. For example, PT scores correlate with 16 personality types based on the four Myers-Briggs Type Indicator (MBTI) [31] scores; each type refers to certain popular occupations. We use this correlation in our framework to simplify and structure decision-making in Human Resources (HR) processes. This includes the selection of candidates based on professional responsibilities and the determination of professional compatibility among colleagues. In addition, the framework can be used in the areas where PT is crucial for human's life, namely, to customize products and services according to consumer preferences, to identify personality disorders, and to build individual educational trajectories.

5. Conclusions

In this study, we propose a novel framework for PTA. The main idea implies the use of SN, which models the PTA of a person through comparison with another person. This network is based on a gated attention fusion. The framework comprises six modules and enables the fusion of hand-crafted and deep features across three modalities (video, audio, and text). We conduct research on the Flv2 and MuPTA corpora investigating different combinations of features and modalities, and identify meaningful feature combinations of all modalities. We identify that the proposed framework flexibly adapts to the input data, since different modalities prevail for certain PT. The experimental results show that the framework performance is enhanced due to pairwise comparison of two people with the same or different PT while training.

The proposed framework outperforms the SOTA systems both based on three modalities (video-face, audio and text) and utilizing more than three modalities on FIV2 in terms of $mACC$. We also outperform the SOTA system in terms of CCC (0.667 vs. 0.614) using two modalities (video and audio) and report on the baseline results for the system using three modalities (video, audio, and text) on the MuPTA corpus. In addition, we share an open-source framework to integrate it into human-machine applications that require personalization and communication with people on a human level.

Since the proposed framework consists of six modules for a comprehensive analysis of human behavior, it requires significant computational resources. Nevertheless, it can work in real-time on both CPU and GPU of personal computers. However, the framework is not optimized for use on mobile devices yet, which is a disadvantage. In addition, the framework does not currently provide an interpretation of the obtained results that could be useful for experts who analyze PT in a decision-making process. In our future research, we plan to address these shortcomings and to develop a multi-task framework by optimizing the proposed framework for other tasks such as emotion recognition, sentiment analysis, and other challenges.

CRedit authorship contribution statement

Elena Ryumina: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Maxim Markitantov:** Writing – review & editing, Writing – original draft, Conceptualization. **Dmitry Ryumin:** Validation, Software, Formal analysis. **Alexey Karpov:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The research data is publicly available and can be found at <https://chalearnlap.cvc.uab.cat/dataset/24/description/>. Our research code is available at <https://oceanai.readthedocs.io/en/latest/>.

Acknowledgments

This work was supported by the Analytical Center for the Government of the Russian Federation (IGK 000000D730324P540002), agreement No. 70-2021-00141.

References

- [1] H. Kaya, F. Gurpinar, A.A. Salah, Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs, in: Proc. of Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2017, pp. 1–9.
- [2] Y. Li, J. Wan, Q. Miao, et al., Cr-net: A deep classification-regression network for multimodal apparent personality analysis, *Int. J. Comput. Vis.* 128 (12) (2020) 2763–2780.
- [3] C. Palmero, J. Selva, S. Smeureanu, et al., Context-aware personality inference in dyadic scenarios: Introducing the UDIVA dataset, in: Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1–12.
- [4] S. Aslan, U. Gdkbay, H. Dibekliođlu, Multimodal assessment of apparent personality using feature attention and error consistency constraint, *Image Vis. Comput.* 110 (2021) 104163.
- [5] D. Giritliođlu, B. Mandira, S.F. Yilmaz, et al., Multimodal analysis of personality traits on videos of self-presentation and induced behavior, *J. Multimodal User Interfaces* 15 (4) (2021) 337–358.
- [6] T. Agrawal, D. Agarwal, M. Balazia, et al., Multimodal personality recognition using cross-attention transformer and behaviour encoding, in: Proc. of International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP, 2022, pp. 501–508.
- [7] C. Suman, S. Saha, A. Gupta, et al., A multi-modal personality prediction system, *Knowl.-Based Syst.* 236 (2022) 107715.
- [8] T. Agrawal, M. Balazia, P. Mller, et al., Multimodal vision transformers with forced attention for behavior analysis, in: Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 3392–3402.
- [9] R.Z. Cabada, H.M.C. Lpez, H.J. Escalante, Multimodal personality recognition for affective computing, *Multimodal Affect. Comput.: Technol. Appl. Learn. Environ.* (2023) 173–208.
- [10] P.Z. Gan, A. Sowmya, G. Mohammadi, CLIP-based model for effective and explainable apparent personality perception, in: Proc. of the 1st International Workshop on Multimodal and Responsible Affective Computing, 2023, pp. 29–37.
- [11] K. Biswas, P. Shivakumara, U. Pal, et al., VQAPT: A new visual question answering model for personality traits in social media images, *Pattern Recognit. Lett.* 175 (2023) 66–73.
- [12] H.J. Escalante, H. Kaya, A.A. Salah, et al., Modeling, recognizing, and explaining apparent personality from videos, *IEEE Trans. Affect. Comput.* 13 (2) (2020) 894–911.
- [13] E. Ryumina, D. Ryumin, M. Markitantov, et al., Multimodal personality traits assessment (MuPTA) corpus: The impact of spontaneous and read speech, in: Proc. of INTERSPEECH, 2023, pp. 4049–4053.
- [14] W. Ilmini, T. Fernando, Computational personality traits assessment: A review, in: Proc. of IEEE International Conference on Industrial and Information Systems, ICIIS, 2017, pp. 1–6.
- [15] L.V. Phan, J.F. Rauthmann, Personality computing: New frontiers in personality assessment, *Soc. Pers. Psychol. Compass* 15 (7) (2021) e12624.
- [16] W. Wang, H. Ning, F. Shi, et al., A survey of hybrid human-artificial intelligence for social computing, *IEEE Trans. Hum.-Mach. Syst.* 52 (3) (2021) 468–480.
- [17] İ. Yađ, A. Altan, Artificial intelligence-based robust hybrid algorithm design and implementation for real-time detection of plant diseases in agricultural environments, *Biology* 11 (12) (2022) 1732.
- [18] M. Koutsombogera, P. Sarthy, C. Vogel, Acoustic features in dialogue dominate accurate personality trait classification, in: Proc. of IEEE International Conference on Human-Machine Systems, ICHMS, 2020, pp. 1–3.
- [19] D. Curto, A. Claps, J. Selva, et al., Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions, in: Proc. of IEEE/CVF International Conference on Computer Vision, 2021, pp. 2177–2188.
- [20] S. Ntalampiras, Speech emotion recognition via learning analogies, *Pattern Recognit. Lett.* 144 (2021) 21–26.
- [21] Z. Kang, T. Xu, X.-F. Zhu, et al., Learning motion-perceive siamese network for robust visual object tracking, *Pattern Recognit. Lett.* 173 (2023) 23–29.
- [22] X. Yao, T. Song, Rotation invariant gabor convolutional neural network for image classification, *Pattern Recognit. Lett.* 162 (2022) 22–30.
- [23] E. Ryumina, M. Markitantov, D. Ryumin, et al., OCEAN-AI framework with EmoFormer cross-hemiface attention approach for personality traits assessment, *Expert Syst. Appl.* 239 (2024) 122441.
- [24] I. Grishchenko, A. Ablavatski, Y. Kartynnik, et al., Attention mesh: High-fidelity face mesh prediction in real-time, in: Proc. of Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2020, pp. 1–4.
- [25] E. Kamenskaya, G. Kukharev, Recognition of psychological characteristics from face, *Metody Inf. Stosov.* 1 (1) (2008) 59–73.
- [26] F. Eyben, M. Wllmer, B. Schuller, Opensmile: the munich versatile and fast open-source audio feature extractor, in: Proc. of ACM International Conference on Multimedia, 2010, pp. 1459–1462.
- [27] A. Vaswani, N. Shazeer, N. Parmar, et al., Attention is all you need, in: *Advances in Neural Information Processing Systems*, 2017, pp. 1–11.
- [28] J.W. Pennebaker, R.L. Boyd, K. Jordan, K. Blackburn, *The Development and Psychometric Properties of LIWC2015*, University of Texas At Austin, Austin, TX, 2015.
- [29] J. Tiedemann, S. Thottingal, OPUS-MT — Building open translation services for the world, in: Proc. of the Annual Conference of the European Association for Machine Translation, EAMT, 2020, pp. 1–2.
- [30] P. Liu, K. Li, H. Meng, Group Gated Fusion on Attention-Based Bidirectional Alignment for Multimodal Emotion Recognition, in: Proc. of INTERSPEECH, 2020, pp. 379–383.
- [31] A. Furnham, The big five facets and the MBTI: The relationship between the 30 NEO-PI (R) Facets and the four Myers-Briggs Type Indicator (MBTI) scores, *Psychology* 13 (10) (2022) 1504–1516.