The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W3-2023
ISPRS Intl. Workshop "Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine"
PSBB23, 24–26 April 2023, Moscow, Russia

IMPACT OF VISUAL MODALITIES IN MULTIMODAL PERSONALITY AND AFFECTIVE COMPUTING

E. V. Ryumina[a, *], A. A. Karpov [a]

[a] St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation – ryumina_ev@mail.ru, karpov@iias.spb.su

Commission II, WG II/8

KEY WORDS: Personality Computing, Affective Computing, OCEAN, Multi-Task Regression, Multimodal Fusion, Mid-Level Fusion, Cross-Modal Attention, Neural Network

ABSTRACT:

Personality and affective computing techniques play a significant role for better understanding of human's behavior and intentions. Such techniques can be applied in practice in recommendation systems, healthcare, education, and job applicant screening. In this paper, we propose a novel multimodal approach to personality traits assessment that leverages affective features of human's voice and face, as well as recent advances in the deep learning. We present a new mid-level modality fusion strategy that is based on a cross-modal attention mechanism with summarizing functionals. In contrast to other state-of-the-art approaches, we not only analyze a visual scene, but specifically process human's upper body (selfie) and a scene background. Our experiments show that the Extroversion personality trait is better estimated by fusing visual scene, face, and audio (voice) modalities, while the Conscientiousness and Agreeableness traits are better assessed by fusing face, selfie, and audio modalities. Furthermore, our results show that utilizing the selfie modality outperforms the visual scene modality by more than 1% in terms of the Concordance Correlation Coefficient. Additionally, our approach based on processing three modalities (selfie, face, and audio) is on-par with other known state-of-the-art approaches that employ at least four modalities on the test set of the ChaLearn First Impressions V2 corpus.

1. INTRODUCTION

Affective Computing (AC) is an emerging field that focuses on the interaction between humans and machines, with a particular emphasis on the emotional and affective states of these interactions (Picard, 2000). The goal of AC is to create machines that can understand, interpret, and respond to human's emotions.

Personality Computing (PC) is a subfield of AC that deals specifically with the analysis, recognition, and synthesis of personality traits. Personality traits are enduring patterns of thoughts, feelings, and behaviors that shape an individual's personality and distinguish them from others. Therefore, PC combines two scientific areas, psychology and artificial intelligence, making it a relevant field of study.

The importance of PC lies in its potential applications in high-risk tasks such as recommendation systems (Dhelim et al., 2022), healthcare (Phan and Rauthmann, 2021), education (Ilmini and Fernando, 2017), and job applicant screening (Hickman et al., 2022). Today, modern recommendation systems have reached a new level by utilizing knowledge about human's personality traits. Such systems recommend content (goods, services, and others) based on user groups who have similar human's personality traits. In healthcare, PC may aid in mental health diagnosis and treatment. In education, it can assist in personalizing the learning trajectory of students. Lastly, automatic personality traits assessment (PTA) based on multimodal data enables the selection of the most professionally oriented personnel.

There are several ways to evaluate human's personality traits: self-evaluation, familiar-evaluation, third-party evaluation, and automatic evaluation. The first three ways require individuals to complete questionnaires. The most commonly used questionnaires for PTA consist of ten, forty-four, and sixty items (Soto and John, 2017). Answers to the questionnaire items are usually provided on a five-point Likert scale (strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree). The more items that are presented in a questionnaire, the more accurately human's personality traits scores can be calculated. However, filling out long questionnaires is time-consuming and requires a lot of concentration from the person filling it out. In contrast, automatic PTA completely eliminates the need for human resources. As a result, the popularity of using automatic PTA approaches in human-machine interaction systems is growing every year.

State-of-the-art (SOTA) approaches for automatic human's PTA have weaknesses in two sensor modalities: audio and video. Handcrafted features such as Low-Level Descriptors (Kaya et al., 2017), spectrograms (Aslan et al., 2021), and pre-trained convolution neural network (CNN) models (Curto et al., 2021) are mainly used to extract acoustic features. Although neural network features allow for more accurate personality trait scores, not enough attention is paid to fine-tuning neural network feature extractors. When dealing with the video modality, SOTA approaches analyze the face (Li et al., 2020) and scene (Agrawal et al., 2022). However, the scene conveys general information about (1) an upper body (selfie) that shows the human's appearance and body movement and (2) the background that reflects the video recording conditions. At the same time, trained models may become confused and not understand to which source of information (selfie or background) they should pay attention. Both sources provide important information about human's personality. Therefore, in this paper, we propose a novel approach that analyzes the human's voice and facial characteristics, as well as their selfie and background, for automatic human's PTA, eliminating the above-mentioned weaknesses.

The remaining sections of this paper are organized as follows. Section 2 analyzes the existing corpora and SOTA approaches for PTA. In Section 3, we provide a detailed description of our pro-

*Corresponding author

posed approach. The results of the experiments and their analysis are presented in Section 4. Finally, in Section 5, we summarize the findings and discuss future research directions in the field of PC and AC.

2. RELATED WORK

2.1 Review of Corpora

To date, several multimodal corpora have been collected for PTA. The ELEA corpus (Sanchez-Cortes et al., 2011) analyzes the emergence of leadership in newly formed groups through the winter survival task. It contains audio-visual recordings of 148 participants, totaling approximately 10 hours. The Hire Me corpus (Nguyen et al., 2014) assesses hireability for a marketing job through audio-visual recordings of 62 participants, totaling approximately 11 hours. The Joker corpus (Devillers et al., 2015) studies human laughter produced during human-robot interaction through recordings of 37 participants comprising approximately 8 hours of audio-visual recordings with Kinect data. The DyCoDa corpus (Dresvyanskiy et al., 2022) comprises recordings of 30 participants engaged in intensive online collaborative conversations using the winter survival game-task, comprising approximately 10 hours of audio-visual recordings together with Kinect data. These corpora were recorded under office conditions and each participant rated themselves on five personality traits of the OCEAN model: Openness to experience (OPE), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Non-Neuroticism (NNEU).

The MHHRI corpus (Celiktutan et al., 2017) studies personality and engagement in human-human and human-robot interactions through audio-visual recordings of 18 participants comprising 4 hours. The MULTISIMO corpus (Koutsombogera and Vogel, 2018) comprises collaborative group interactions where two players provide answers to a quiz and are guided by a facilitator. It contains recordings of 49 participants comprising 4 hours of audio-visual recordings together with Kinect data. The UDIVA corpus (Palmero et al., 2021) contains dyadic interactions of 147 participants from 22 countries, totaling 90.5 hours of audio-visual data. In contrast to the above-presented corpora, in these corpora, the evaluation of personality traits was made both ways by self- and familiar-evaluation. The RoomReader corpus (Reverdy et al., 2022) also presents the self-evaluation and familiar-evaluation, however, the conditions for recording the corpus are uncontrolled ("in the wild" condition). The RoomReader corpus explores multimodal cues of conversational engagement and behavioral aspects of collaborative interaction in online environments through audio-visual recordings of 118 participants from Zoom, comprising approximately 9 hours of data.

The YouTube Vlogs corpus (Biel and Gatica-Perez, 2012) studies personality impressions from vlogging and comprises audio-visual recordings of 442 participants, totaling approximately 150 hours of audio-video data. The ChaLearn First Impressions v2 corpus (FI V2) (Escalante et al., 2020) is a widely known multimodal corpus comprising 10,000 predominantly "in-the-wild" clips (average duration 15s) extracted from YouTube HD videos of more than 2500 people (with different gender, age, nationality, and ethnicity). The last two corpora contain video clips shot mainly "in the wild" conditions, and the annotation for personality traits is made by third-party.

Existing corpora differ in: (1) recording conditions (uncontrolled ("in the wild") and office); (2) evaluation ways (self, familiar, third-party); (3) the number of speakers from 18 to 2500; (4) the duration of hours of clips from 4 to 48.

In our research, we decided to utilize the FI V2 corpus for several reasons. Firstly, it is the most widely used corpus for PTA, having been employed in several competitions co-located with ECCV 2016 (Ponce-López et al., 2016), ICPR 2016 (Escalante et al., 2016), and CVPR 2017 (Bekhouche et al., 2017). The corpus consists of three subsets: Train (6000 audio-video clips), Valid (2000), and Test (2000), and we retained this subset distribution in order to compare our approach with SOTA. Secondly, the corpus includes clips gathered via the YouTube video hosting platform, with most of the clips having been recorded "in the wild". Third, the corpus contains more than 2500 distinct speakers ranging in age from 9 to 62 years old. Finally, the clips were annotated pairwise, which helps to reduce the possibility of subjective evaluation.

2.2 Review of State-of-the-art Approaches

The researchers have proposed many SOTA approaches for PTA using reviewed corpora. However, we have focused on the approaches proposed using the FI V2 corpus. These approaches can be compared because they were carried out using the same training, validation, and testing protocols.

The authors (Kaya et al., 2017) analyzed three main modalities: face, scene, and audio. They utilized a CNN pre-trained on the emotion recognition task and local Gabor binary patterns from three orthogonal planes (LGBP-TOP) to obtain features from the aligned facial regions. To extract features from the scene, they applied a CNN pre-trained on the object recognition task. For the audio modality, they extracted hand-crafted features using openSMILE. The authors fused the modalities at two levels: (1) feature-level fusion using kernel extreme learning machines (KELM); (2) score-level fusion using random forests. The authors did not segment the clips but analyzed the entire clips.

In the paper (Agrawal et al., 2022), the authors introduced a fourth modality: behavior encoding, in addition to the three main modalities. They used 3D CNNs to extract features from the face and scene, and for voice analysis, they employed features obtained using a 2D CNN pre-trained on the audio event recognition task, with log-Mel spectrograms as input data. To encode behavior, the authors created their own set of hand-crafted features to assess 13 different behaviors. All features were used as input data for the cross-attention transformer. The authors segmented the clips into 2.5-sec intervals and used downsampling of the frames. In their subsequent paper (Agrawal et al., 2023), the authors eliminated behavior encoding, introduced a forced attention (FAt) transformer, and did not use downsampling of the frames.

In contrast to the papers mentioned above, the authors of (Aslan et al., 2021) utilized a combination of 2D CNNs and long short-term memory (LSTM) networks for all three main modalities. In addition, the authors aligned the facial regions, similar to (Kaya et al., 2017) and used voice features, as in (Agrawal et al., 2022). They used a 2D CNN pre-trained on the image classification task for the face and scene modalities. However, due to the complexity of their proposed model architecture, the authors only analyzed the first 6 sec of clips. To combine modalities, the authors used an attention mechanism.

In the paper (Li et al., 2020), the authors proposed a two-level approach and the Bell loss function for PTA. At the first level, the personality traits were classified, while at the second level, a regression task was performed. To represent the faces and scenes, the authors randomly divided the video into 32 segments and extracted one frame from each segment. The 2D CNNs were used

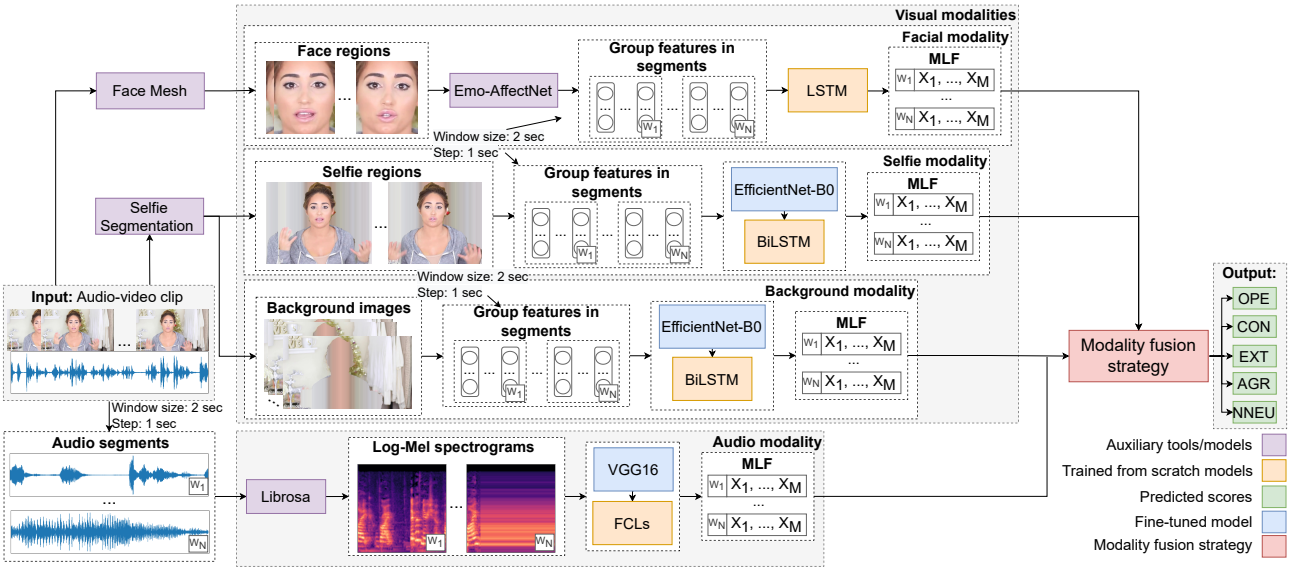

Figure 1: Pipeline of the proposed approach for PTA. W_i denotes audio-visual clip segments, N - the number of 2-sec segments in the audio-visual clip. MLF – mid-level features, M - the number of features in a set of the MLF

to extract features from all three modalities, while the audio signal was fed to the network as raw data. The authors employed extra trees regression for modality fusion.

Overall, the reviewed SOTAs have mainly focused on analyzing the scene. One exception is the paper (Agrawal et al., 2022), where the authors proposed hand-crafted features for behavior encoding. In our paper, we propose to analyze human's behavior by leveraging neural network features without considering the background. Furthermore, we conduct a series of experiments using different modality fusion strategies, including (1) using summarizing functionals, (2) using cross-modal attention, and (3) using both strategies.

3. PROPOSED APPROACH

The proposed multimodal and multi-tasks approach for automatic human's PTA analyzes several modalities: audio, face, selfie, and background. The pipeline of the proposed approach for PTA using mid-level feature fusion is shown in Figure 1. Audio-visual clips from the FI V2 corpus range from 2 to 16 sec. Consistent with prior researchesr (Agrawal et al., 2022, Aslan et al., 2021, Li et al., 2020), we also segment audio-visual clips in our approach. Specifically, we segment them in 2-sec intervals with a 1-sec step. The output of each modality for one segment is a vector of mid-level features. For the entire clip, we obtain a feature vector sequence. We then use the sequences of all modalities as input for the selected fusion strategy. We also downsample the frames per second (FPS) of clips to 5 frames with an even step since FPS is not uniform and ranges from 7 to 30.

3.1 Affective Features of Voice and Face

Previously researches have shown that certain personality traits are associated with specific affective states (Dauvier et al., 2019, Kaya et al., 2017). For example, people who are high score for EXT trait tend to experience more positive affective states such as happiness, while those who are high score for NEU trait tend to experience more negative affective states such as sadness.

Since personality traits and affective states are interconnected. In this paper, in order to achieve a more reliable PTA, we use affective features for two modalities: audio and facial. In the paper (Verkholyak et al., 2021), we presented VGG16 model, which was trained on 3-way escalation prediction in speech: low, medium and high. The model allows extracting 256 voice features. In the paper (Ryumina et al., 2022), we presented a open-source visual model of Emo-AffectNet, which was trained on the 7-emotion recognition task: angry, sadness, happiness, disgust, fear, neutral states, surprise. The model allows extracting 512 emotional facial features.

Using VGG16 and Emo-AffectNet models, we extract affective low-level features form the audio and facial modalities. The complete process of processing these modalities is described below.

3.2 Audio Modality

We extract log-Mel spectrograms using the open-source library Librosa (McFee et al., 2015). We use 128 Mel filter-banks with a short-time Fourier transform window length of 2048 and a step of 512. The resulting feature matrix for a 2-sec segment has a dimension of 128×173. Next, we resize the spectrograms to 224×224 and use them as input data for a 2D CNN based on VGG16 (Simonyan and Zisserman, 2015). We add two fully connected layers (FCLs) of 512 and 256 neurons to the VGG16 model. We fine-tune the VGG16 model for personality trait feature extraction from voice.

All models in our approach were trained with a learning rate change via cosine annealing learning schedule (Loshchilov and Hutter, 2017). Specifically, we trained the VGG16 model with a learning rate change from 0.00005 to 0.000005 and five restart cycles over 100 epochs.

3.3 Visual Modalities

In our approach, we analyze several visual modalities: facial, scene, selfie, and background. For facial modality, we use a neural network model that is different from other modalities.

3.3.1 Facial Modality: The face regions are detected using the Face Mesh model from the open-source cross-platform MediaPipe (Grishchenko et al., 2020, Lugaresi et al., 2019), which has shown promising results in human's body and face detection

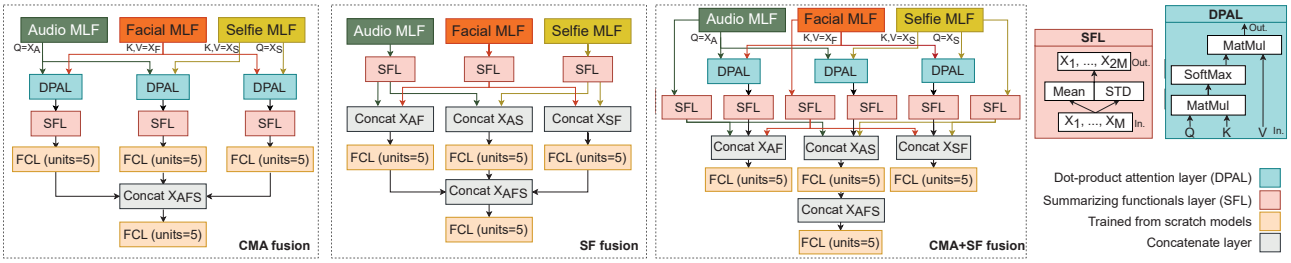

Figure 2: Modality fusion strategies. CMA denotes cross-modal attention. SF – summarizing functionals. Q, K, V – query, key, and value, respectively. X_A, X_F, X_S – features of audio, face, selfie

problems (Ryumin et al., 2023) without requiring additional devices (Ryumin and Karpov, 2017). Subsequently, the 512 static emotional neural network features are extracted from the face regions using the open-source Emo-AffectNet model (Ryumina et al., 2022). Finally, ten emotional feature vectors for each 2-sec segment are used as input data for the LSTM layer, which has 1024 neurons.

The LSTM was trained with a learning rate that changed from 0.001 to 0.0001 and five restart cycles over 100 epochs.

3.3.2 Selfie, Scene, Background Modalities: The selfie regions are detected using the Selfie Segmentation detector (Lugaresi et al., 2019). We extract the segmented selfie regions from the scene to eliminate the attention of the models to the background, enabling us to separate the selfie and background. The images are resized to 224×224 pixels and filled with average values to maintain the proportions of the objects in the images. The background on the selfie images and selfie regions on the background images are filled with vertical average pixel values. These images are then formed into sequences and used as input data for a 2D CNN based on the EfficientNet-B0 model architecture (Tan and Le, 2019) with bidirectional LSTM (BiLSTM) models. The 2D CNN model was pre-trained on the object recognition task and exhibited the best performance measures in the task of recognizing affective states (Savchenko, 2022, Ryumina et al., 2021). We complement this model a FCL of 512 neurons and fine-tune it for personality trait feature extraction from the selfie, scene, and background at the frame-level. We use a single-layer BiLSTM model with 256 neurons for modeling personality traits at the segment-level (ten feature vectors) extracted from the FCL layer of the EfficientNet-B0 model. Notably, we trained models for not only selfie and background modalities but also for the scene modality. Thus, we have three identical models to reliably investigate the performance of three different modalities.

The EfficientNet-B0 models were trained with a learning rate that changed from 0.001 to 0.0001 over five restart cycles and 100 epochs. The BiLSTM models were trained with a learning rate that changed from 0.0001 to 0.00001 over 200 epochs without restart cycles.

3.4 Modality Fusion Strategies

All models are equipped with multi-task regression layers for predicting five intermediate personality trait scores (five scores for each segment of one clip). The average duration of clips in the FI V2 corpus is 15 sec. A 15-sec clip is divided into 16 segments using a 2-sec window with a 1-sec step. We extract the mid-level features (features before the predictive layers) obtained using each trained model for each segment. To fuse modalities at the feature-level, we employ three different strategies, are presented in Figure 2.

3.4.1 Cross-modal Attention Strategy: We apply cross-modal attention based on the dot-product attention (Vaswani et al., 2017), which is calculated by the formula:

$$Attention(Q, K, V) = softmax(QK^T)V \quad (1)$$

where Q means the query vector, K - the key vector, V - the value vector. The idea behind dot-product attention is to suppress less important features in the V vector while amplifying more important features. The dimensions of features can vary while performing the dot product of two vectors. For instance, voice features have a dimension of 16×256, whereas facial features have a dimension of 16×1024. Thus, we concatenate multiple vectors of smaller dimensions to obtain the required dimension. This approach yields better results compared to using FSLs with an equal number of neurons in them before the dot-product attention layer.

The attention layer outputs sequences of feature vectors. To flatten the several vectors of one sequence to one vector, we apply a summarizing functional layer that aggregates the mean and standard deviation values. We use the final summarizing vector as input data for an FSL with 5 neurons. We use this algorithm for feature-level fusion of two modalities. When fusing three or more modalities, we concatenate the predicted scores from all the fusion of two modalities into one vector and feed it to the last FCL layer to obtain the final predictions.

3.4.2 Summarizing Functionals Strategy: Previous studies on PTA have shown that summarizing functionals are effective for fusing several modalities and can improve performance measures (Kaya et al., 2017). Therefore, we also apply summarizing functionals for fusing multiple modalities. We first calculate the summarizing functionals for the feature vectors of each modality, and then concatenate the resulting summarizing vectors from two different modalities. The subsequent steps are similar to the previous strategy.

3.4.3 Cross-Modal Attention with Summarizing Functionals Strategy: As part of the last fusion strategy, we calculate summarizing functionals from the raw feature vectors and after applying the attention layer. Then, we concatenate the summarizing vectors obtained from the raw data of the two modalities and their summarizing vectors obtained after the attention layer. We use the concatenated vectors as input data for an FSL with 5 neurons, one for each vector. For the fusion of three modalities, we concatenate the intermediate predicted scores from the three vectors into one vector to obtain the final predicted scores.

We compare three proposed fusion strategies by combining different modalities. All three fusion strategies were trained for 200 epochs with a learning rate starting at 0.1 and decreasing to 0.01 with 15 restart cycles.

4. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the proposed models for PTA, we used two measures: accuracy (ACC) (Escalante et al., 2020), and concordance correlation coefficient (CCC) (Lawrence and Lin, 1989). While ACC reflects the error between predicted and ground truth scores, CCC indicates a correlation between them. Measures are calculated according to the formulas:

$$ACC = 1 - \frac{1}{R}\sum_{k=1}^{R}|t_k - p_k| \tag{2}$$

$$CCC = \frac{2 \cdot \sigma_{t,p}}{\sigma_t^2 + \sigma_p^2 + (\mu_t - \mu_p)^2}, \tag{3}$$

where t and p denote the ground truth and predicted scores for a clip k, respectively; μ_t and μ_p – the averaged ground truth and predicted scores for all test clips; σ_t and σ_p – the respective standard deviations; $\sigma_{t,p}$ – the covariance between t and p.

Performance measures for the unimodal approaches are presented in Table 1. Since each modality predicts several scores per clip, we simply averaged these predicted scores to evaluate the performance of the models. We don't fuse scene modality with background modality and scene modality with selfie modality because selfie and scene modalities are already included in scene modality. Table 1 shows that the performance measures are the lowest for the audio modality. The performance measure ACC shows that the face is the best information source for PTA, while CCC, on the contrary, shows that selfie is more informative. Additionally, it can be seen that the audio modality is inferior to the selfie and facial modalities by almost 12%. Moreover, the results demonstrate that selfie outperforms scene by almost 1% in terms of CCC. This suggests that while training the model on the whole frame, the model has more information to analyze, which leads to confusion.

Modality	CCC	ACC
Audio (A)	.523	.906
Face (F)	.647	**.913**
Scene (S)	.638	.912
Selfie ($\bar{S}$)	**.649**	.912
Background (B)	.600	.907

Table 1: Comparison of the performance measures in terms modalities

Performance measures for the modality fusion strategies are presented in Table 2. The experimental results demonstrate that, irrespective of the two modalities used for fusion, the cross-modal attention strategy (CMA) yields lower performance on average by more than 4% in terms of the CCC measure. Moreover, the summarizing functionals strategy (SF) exhibits lower performance than the general strategy (CMA+SF) average by almost 1% for the same measure. Fusing three or four modalities reduces the gap in performance by almost half.

The results demonstrate that the maximum measure values of ACC=.916 and CCC=.672 were achieved by fusion of the two modalities for the CMA+SF strategy. We also observed that while the audio modality is weaker than the scene and selfie modalities, fusion of these three modalities with the facial modality yields similar performance measures. This suggests that voice features are significantly different from other modality features, which positively impacts performance measures. Moreover, fusion of the audio and facial modalities with the selfie modality is more efficient than fusion of them with the scene. However, adding the background modality to other modalities decreases the CCC measure. Overall, we achieved an increase in performance measures of .2% for ACC (.916 vs. .918) and 2.5% for CCC (.672 vs. .697) by fusion of three and four modalities.

Fusion	CMA		SF		CMA+SF	
	CCC	ACC	CCC	ACC	CCC	ACC
Bimodal fusion						
A+F	.611	**.913**	.654	.916	.664	**.916**
A+S	.613	.912	.636	.914	.642	.914
A+$\bar{S}$	.621	.912	.641	.914	.652	.914
A+B	.570	.907	.603	.911	.622	.912
F+S	**.628**	**.913**	.661	**.917**	.667	**.916**
F+$\bar{S}$	.607	**.913**	**.673**	**.917**	**.672**	**.916**
F+B	.613	**.913**	.646	.915	.656	.915
$\bar{S}$+B	.614	.912	.626	.913	.644	.913
Average	.610	.912	.643	.915	.652	.915
Multimodal fusion						
A+F+S	.675	.916	.691	.918	.694	.918
A+F+$\bar{S}$	**.679**	.917	**.695**	.918	**.697**	.918
A+F+$\bar{S}$+B	.677	.917	.691	.918	.696	.918
Average	.677	.917	.692	.918	.696	.918

Table 2: Comparison of the performance measures in terms modality fusion strategies

The comparison of trait-wise measures and their average values obtained by our approach and the SOTA is presented in Table 3. Our experiments demonstrated that the obtained results in terms of ACC measure outperform other SOTA results for the audio, video (face), and video (scene, behavior encoding) modalities. We also report that our approach has the same efficiency as SOTA approaches that analyze four modalities: audio, video (face), video (scene, behavior encoding), and text. However, our approach is significantly inferior to approaches that use transformers and five modalities: audio, video (face), video (scene, behavior encoding), text, and metadata.

Approach	OPE	CON	EXT	AGR	NNEU	Avg.
Trait-wise CCC measure						
A+F+S	.650	.746	**.724**	.534	.676	.666
A+F+$\bar{S}$	**.653**	**.752**	.716	.543	.675	.668
A+F+$\bar{S}$+B	.652	.751	.720	**.548**	**.678**	**.670**
Trait-wise ACC measure						
A+F+S	.917	.923	.921	.915	.914	.918
A+F+$\bar{S}$	.916	.923	.922	.915	.915	.918
A+F+$\bar{S}$+B	.916	.923	.921	.916	.915	.918
Face, audio, scene, behavior encoding modalities						
(Kaya et al., 2017)	.917	.920	.921	.914	.915	.917
(Aslan et al., 2021)	–	–	–	–	–	.917
(Agrawal et al., 2022)	.901	.899	.899	.904	.900	.901
Plus text modality						
(Aslan et al., 2021)	.916	.922	.920	.916	.915	.918
(Li et al., 2020)	.920	.922	.920	.918	.915	.919
Plus metadata modality						
(Agrawal et al., 2022)	.929	.926	.927	.929	.921	.926
(Agrawal et al., 2023)	**.942**	**.951**	**.955**	**.949**	**.959**	**.951**

Table 3: Comparison of the performance measures of our approach with SOTA

It is worth noting that the CCC measure for the approach with the fusion of four modalities (audio, face, selfie, background) outperforms the approach with the fusion of three modalities (audio, face, selfie/scene) in the trait-wise-level evaluation. This suggests that the background modality is also significant for automatic PTA. Our experiments also show that EXT trait is better

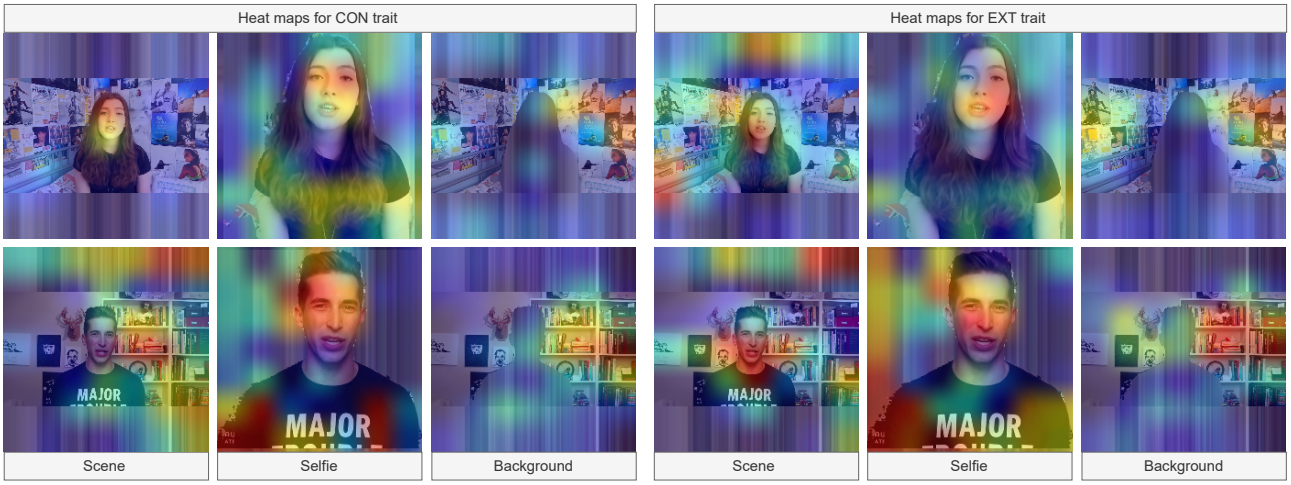

Figure 3: Heat maps for scene, selfie, and background modalities

estimated by fusing facial, audio and scene modalities. While CON and AGR traits are better assessed by fusing facial, audio and selfie modalities.

We also used the GradCAM (Selvaraju et al., 2017) technique to assess the attention of feature maps for scene, selfie, and background modalities. GradCAM allows drawing gradient heat maps for each modality and analyzing which informative region is more important when making decisions about the traits scores. We visualized heat maps using the latest convolutional layers of the EfficientNet-B0 models for CON and EXT traits, which are predicted scores with largest performance measures. Heat maps for three modalities are visualized in Figure 3.

The attention of the models, represented by the bright yellow and red pixels in the images, varies depending on the analyzed person and personality trait. For instance, when analyzing images of a girl, the scene model primarily focuses on the face region for the CON trait, while for the EXT trait, it looks at both the face and the recording conditions. On the other hand, for images of a guy, the scene model focuses on the recording conditions for both traits, but for the EXT trait, it gives more attention to the lower part of the face. As for the selfie modality, regardless of the person, the selfie model predominantly looks at the region above the lower eyelid for the CON trait, and at the region below the lower eyelid for the EXT trait. The attention of the background model also varies depending on the traits. For example, this model analyzes more information on both sides of the person for EXT trait, whereas only on one side of the person – for CON trait.

In general, heat maps show that the scene model draws attention to the upper part of the face and the background in all four images. Whereas the selfie model pays attention to active muscles of the face and the appearance of the person. At the same time, the attention of the background model coincides with the attention of the scene model, while attention also extends to regions where the person is absent, which probably contributes to a decrease in the performance measure of this model. Moreover, complex background conditions that vary based on the person’s location can negatively impact background modality, as they may not accurately represent the human’s personality.

In conclusion, separating the scene into its selfie and background components is a more efficient solution than analyzing the whole scene. This is due to the fact that by separating these two sources of information, each model analyzes a specific region of interest in the image, which has a positive effect on the performance measures of the PTA approaches. Moreover, to achieve better measures in PTA, it is essential to analyze not only the audio, facial, and selfie modalities, but also the text and metadata modalities.

5. CONCLUSIONS

In this paper, we presented the novel approach for multimodal personality and affective computing. This approach has been designed for PTA and it is able to analyze four human’s modalities: audio, face, selfie, and background. The advantage of our approach is that it is based on affective facial and voice features. For the facial modality, we extract low-level emotional features using the open-source Emo-AffectMet model, which are then used as input data for the LSTM model to extract mid-level features. We apply fine-tuned EfficientNet models and BiLSTM models to extract mid-level features from the scene, selfie, and background information. From the audio modality we extract log-Mel spectrograms, which are then fed into the VGG16 model trained on the speech escalation prediction task to extract mid-level features. The audio-visual features are extracted at the segment-level of each multimodal clip. Finally, we fuse the mid-level features at the clip-level by the cross-modal attention with summarizing functionals.

Our experiments demonstrate that dividing a visual scene into selfie and background modalities is more efficient than analyzing the entire scene, as each modality can analyze a specific graphical region of interest that positively affects the performance measures. On the test subset of the ChaLearn First Impressions V2 corpus, the proposed approach outperforms other systems that use both audio and video (face, scene, and behavior encoding) modalities in terms of CCC and ACC measures. Furthermore, we also demonstrate that our approach is efficient as SOTA approaches that analyze at least four human’s modalities: audio, video, scene, behavior encoding, and text.

Thus, our approach is a promising component for enhancing existing solutions for automatic human’s PTA in such tasks as recommendation systems, healthcare, education, and job applicant screening. In the future, we plan to combine the proposed approach with transformer-based models for an end-to-end multimodal analysis.

ACKNOWLEDGEMENTS

The research on affective computing was financially supported by the Russian Science Foundation (project No. 22-11-00321).

REFERENCES

Agrawal, T., Agarwal, D., Balazia, M. et al., 2022. Multimodal personality recognition using cross-attention transformer and behaviour encoding. *International Conference on Computer Vision Theory and Applications (VISAPP)*, 5, 501–508. doi.org/10.5220/0010841400003124.

Agrawal, T., Balazia, M., Müller, P., Brémond, F., 2023. Multimodal Vision Transformers with Forced Attention for Behavior Analysis. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 3392–3402. doi.org/10.1109/WACV56688.2023.00339.

Aslan, S., Güdükbay, U., Dibeklioğlu, H., 2021. Multimodal assessment of apparent personality using feature attention and error consistency constraint. *Image and Vision Computing*, 110, 104163. doi.org/10.1016/j.imavis.2021.104163.

Bekhouche, E. S., Dornaika, F., Ouafi, A., Taleb-Ahmed, A., 2017. Personality traits and job candidate screening via analyzing facial videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 10–13. doi.org/10.1109/CVPRW.2017.211.

Biel, J.-I., Gatica-Perez, D., 2012. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1), 41–55. doi.org/10.1109/TMM.2012.2225032.

Celiktutan, O., Skordos, E., Gunes, H., 2017. Multimodal human-human-robot interactions (MHHRI) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 10(4), 484–497. doi.org/10.1109/TAFFC.2017.2737019.

Curto, D., Clapés, A., Selva, J. et al., 2021. Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2177–2188. doi.org/ICCVW54120.2021.00247.

Dauvier, B., Pavani, J.-B., Le Vigouroux, S. et al., 2019. The interactive effect of neuroticism and extraversion on the daily variability of affective states. *Journal of Research in Personality*, 78, 1–15. doi.org/10.1016/j.jrp.2018.10.007.

Devillers, L., Rosset, S., Duplessis, G. D. et al., 2015. Multimodal data collection of human-robot humorous interactions in the Joker project. *International Conference on Affective Computing and Intelligent Interaction (ACII)*, 348–354. doi.org/10.1109/ACII.2015.7344594.

Dhelim, S., Aung, N., Bouras, M. et al., 2022. A survey on personality-aware recommendation systems. *Artificial Intelligence Review*, 55, 2409–2454. doi.org/10.1007/s10462-021-10063-7.

Dresvyanskiy, D., Sinha, Y., Busch, M. et al., 2022. DyCoDa: A multi-modal data collection of multi-user remote survival game recordings. *International Conference on Speech and Computer (SPECOM)*, 163–177. doi.org/10.1007/978-3-031-20980-2_15.

Escalante, H. J., Kaya, H., Salah, A. A. et al., 2020. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, 13(2), 894-911. doi.org/10.1109/TAFFC.2020.2973984.

Escalante, H. J., Ponce-López, V., Wan, J. et al., 2016. ChaLearn joint contest on multimedia challenges beyond visual analysis: An overview. *International Conference on Pattern Recognition (ICPR)*, 67-73. doi.org/10.1109/ICPR.2016.7899609.

Grishchenko, I., Ablavatski, A., Kartynnik, Y. et al., 2020. Attention mesh: High-fidelity face mesh prediction in real-time. *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) on Computer Vision for Augmented and Virtual Reality*, 1–4. doi.org/10.48550/arXiv.2006.10962.

Hickman, L., Bosch, N., Ng, V. et al., 2022. Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, 107(8), 1323. doi.org/10.1037/apl0000695.

Ilmini, W., Fernando, T., 2017. Computational personality traits assessment: A review. *IEEE International Conference on Industrial and Information Systems (ICIIS)*, 1–6. doi.org/10.1109/ICIINFS.2017.8300416.

Kaya, H., Gurpinar, F., Ali Salah, A., 2017. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1–9. doi.org/10.1109/CVPRW.2017.210.

Koutsombogera, M., Vogel, C., 2018. Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus. *International Conference on Language Resources and Evaluation (LREC)*, 2945–2951.

Lawrence, I., Lin, K., 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268. doi.org/10.2307/2532051.

Li, Y., Wan, J., Miao, Q. et al., 2020. Cr-net: A deep classification-regression network for multimodal apparent personality analysis. *International Journal of Computer Vision*, 128, 2763–2780. doi.org/10.1007/s11263-020-01309-y.

Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. *International Conference on Learning Representations (ICLR)*, 1–16. doi.org/10.48550/arXiv.1608.03983.

Lugaresi, C., Tang, J., Nash, H. et al., 2019. Mediapipe: A framework for perceiving and processing reality. *Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 1–4. doi.org/10.48550/arXiv.1906.08172.

McFee, B., Raffel, C., Liang, D. et al., 2015. librosa: Audio and music signal analysis in python. *Python in Science Conference*, 8, 18–25. 10.25080/Majora-7b98e3ed-003.

Nguyen, L. S., Frauendorfer, D., Mast, M. S., Gatica-Perez, D., 2014. Hire me: Computational inference of hirability in employment interviews based on nonverbal behavior. *IEEE Transactions on Multimedia*, 16(4), 1018–1031. doi.org/10.1109/TMM.2014.2307169.

Palmero, C., Selva, J., Smeureanu, S. et al., 2021. Context-aware personality inference in dyadic scenarios: Introducing the UDIVA dataset. *IEEE/CVF Winter Conference on Applications of Computer Vision*, 1–12. doi.org/10.1109/WACVW52041.2021.00005.

Phan, L. V., Rauthmann, J. F., 2021. Personality computing: New frontiers in personality assessment. *Social and Personality Psychology Compass*, 15(7), e12624. doi.org/10.1111/spc3.12624.

Picard, R. W., 2000. *Affective computing*. MIT press.

Ponce-López, V., Chen, B., Oliu, M. et al., 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. *European Conference on Computer Vision (ECCV)*, 400–418. doi.org/10.1007/978-3-319-49409-8_32.

Reverdy, J., Russell, S. O., Duquenne, L. et al., 2022. RoomReader: A multimodal corpus of online multiparty conversational interactions. *International Conference on Language Resources and Evaluation (LREC)*, 2517–2527.

Ryumin, D., Ivanko, D., Ryumina, E., 2023. Audio-Visual Speech and Gesture Recognition by Sensors of Mobile Devices. *Sensors*, 23(4), 2284. doi.org/10.3390/s23042284.

Ryumin, D., Karpov, A. A., 2017. Towards automatic recognition of sign language gestures using kinect 2.0. *Universal Access in Human–Computer Interaction. Designing Novel Interactions (UAHCI)*, 89–101. doi.org/10.1007/978-3-319-58703-5_7.

Ryumina, E., Dresvyanskiy, D., Karpov, A., 2022. In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study. *Neurocomputing*, 514, 435–450. doi.org/10.1016/j.neucom.2022.10.013.

Ryumina, E., Verkholyak, O., Karpov, A., 2021. Annotation Confidence vs. Training Sample Size: Trade-Off Solution for Partially-Continuous Categorical Emotion Recognition. *Interspeech*, 3690–3694. doi.org/10.21437/Interspeech.2021-1636.

Sanchez-Cortes, D., Aran, O., Gatica-Perez, D., 2011. An audio visual corpus for emergent leader analysis. *Workshop on Multimodal Corpora for Machine Learning: Taking Stock and Road Mapping the Future (ICMI-MLMI)*, 1–6.

Savchenko, A. V., 2022. Personalized frame-level facial expression recognition in video. *Pattern Recognition and Artificial Intelligence (ICPRAI)*, 447–458. doi.org/10.1007/978-3-031-09037-0_37.

Selvaraju, R. R., Cogswell, M., Das, A. et al., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IEEE International Conference on Computer Vision (ICCV)*, 618–626. doi.org/10.1109/ICCV.2017.74.

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations (ICLR)*, 1–14. doi.org/10.48550/arXiv.1409.1556.

Soto, C. J., John, O. P., 2017. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113(1), 117–143. doi.org/10.1037/pspp0000096.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*, 6105–6114. doi.org/10.48550/arXiv.1905.11946.

Vaswani, A., Shazeer, N., Parmar, N. et al., 2017. Attention is all you need. *Conference on Neural Information Processing Systems (NIPS)*, 1–11. doi.org/10.48550/arXiv.1706.03762.

Verkholyak, O., Dresvyanskiy, D., Dvoynikova, A. et al., 2021. Ensemble-Within-Ensemble Classification for Escalation Prediction from Speech. *Interspeech*, 481–485. doi.org/10.21437/Interspeech.2021-1821.