Personality Traits Recognition via Fusion of Visual Features

Atefeh Alimohammadi, Faezeh Salehi, Dr.Hamidreza Baradaran

Abstract

This study addresses automatic personality recognition based on the Five-Factor Model (FFM) using facial video analysis. The original S2D [1] architecture, designed for facial expression recognition, was adapted here due to its ability to capture fine-grained spatial–temporal facial dynamics. While expressions reflect short-term emotional states, consistent patterns in these dynamics can reveal underlying personality traits. Leveraging this link, the S2D framework was modified to extract both static and temporal cues relevant to personality prediction. The ChaLearn First Impression dataset, comprising short annotated video clips of individuals with labels for the five personality traits, was used. The system integrates facial landmark detection with Multi-Channel Perception (MCP) and Temporal Multi-head Attention (TMA) modules to enhance spatial–temporal feature fusion. Evaluations using Mean Accuracy (MA), Mean Absolute Error (MAE), and Concordance Correlation Coefficient (CCC) show that the adapted architecture achieves competitive performance, demonstrating the potential of emotional dynamics as a basis for personality inference.

Introduction

Personality recognition based on psychological frameworks such as the Five-Factor Model (FFM) is an emerging and practical field with applications in human resources, psychology, and human–computer interaction. This technology can assist in tasks such as personnel recruitment, enhancing user experience in intelligent systems, and supporting educational and therapeutic interventions. Advances in artificial intelligence and deep learning have enabled the analysis of complex data modalities, including facial videos, to extract features associated with personality traits. Such capabilities contribute to a deeper understanding of human behavior and interactions. Despite recent technological progress, personality analysis from video data still faces several challenges. These include limited availability and diversity of annotated video datasets, the complexity of jointly modeling spatial and temporal features, and uncertainty in predicting personality dimensions. Furthermore, there is a need for models that can achieve high accuracy with computational efficiency suitable for real-time scenarios. Psychological studies suggest that facial expressions, as well as the temporal sequences of their subtle changes, can reflect underlying behavioral tendencies associated with certain personality traits. These dynamic cues, when analyzed alongside static facial attributes, can provide a richer representation for personality prediction. The S2D architecture, originally proposed for video-based emotion recognition, is inherently designed to capture both static appearance information and temporal patterns of facial changes. Motivated by this capability, the present study adapts and extends S2D for FFM-based personality recognition. In the proposed approach, facial landmark sequences serve as structural inputs for spatial feature extraction, while temporal dependencies are modeled through sequential processing of frames, enabling the integration of static and dynamic information for improved prediction accuracy.

Related Work

Automated personality analysis systems can be broadly divided into two categories: self-reported personality recognition and apparent personality recognition. The former relies on self-assessment tools such as questionnaires to measure actual personality traits, whereas the latter aims to predict observers' impressions of an individual's personality (e.g., first impressions). Most existing research focuses on apparent personality recognition, often leveraging visual cues such as facial expressions or multimodal audiovisual data. Recent studies [2] have highlighted the importance of analyzing long-term behavioral patterns, as personality traits tend to remain stable over time. Early works in personality and emotion recognition predominantly employed static models, which extract spatial features from still images using convolutional neural networks (CNNs) or Vision Transformers (ViTs). Large-scale datasets such as AffectNet have played a key role in advancing these methods. In contrast, dynamic models focus on capturing temporal dependencies in video data. These approaches, which often combine CNN-based spatial encoders with recurrent layers, have shown promise despite the challenges posed by limited annotated video datasets. Examples include DFER-CLIP and MAE-DFER, which leverage temporal context to improve recognition accuracy.

A more recent trend is knowledge transfer between static and dynamic models. Here, models are first trained on large-scale static datasets to learn spatial representations and are subsequently adapted to video analysis by incorporating temporal modules. The Inflated 3D ConvNet (I3D) is a representative example of this strategy, extending 2D CNN filters into 3D to handle spatiotemporal data. In the context of the ChaLearn First Impressions dataset, several notable visual-only models have been proposed for apparent personality prediction, including PersEmoN [4], DAN [5], and CAM-DAN+ [6]. While the proposed S2D-based model in this study achieved a mean accuracy (MA) of 0.8860, which is slightly lower than PersEmoN (0.916) and CAM-DAN+ (0.912), it demonstrated competitive performance across multiple evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and concordance correlation coefficient (CCC). The performance gap can be attributed to architectural constraints—such as using a shallower network and reduced input resolution (112×112 instead of the standard 224×224 for ViTs)—which limited the model's ability to capture fine-grained features. Future improvements could be achieved by increasing input resolution, deepening the architecture, and extending the number of training epochs.

Methodology

Overview

The proposed approach adapts the S2D model [1], originally designed for dynamic facial emotion recognition, for the task of personality trait prediction based on video sequences. Given the established link between facial affect and certain personality traits, the architecture leverages both static features extracted from individual frames and dynamic features capturing temporal variations across frames. Two key modules, MCP (Multi-Channel Perception) and TMA (Temporal Modeling Adapter), are employed to integrate spatial and temporal information effectively.

Dataset

We utilize the ChaLearn First Impression dataset [2], containing 10,000 short video clips from over 2,700 YouTube users, annotated with the five factors of personality (Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience). This dataset provides the ground truth for training and evaluation.

Input Representation

The model processes two types of inputs:

Facial image sequences (X_F): consecutive frames of a video capturing facial expressions, where each frame contains RGB information.

Facial landmark features (X_L): coordinates of key facial points extracted using a pre-trained landmark detection model.

Both inputs are partitioned into smaller patches and projected into a latent embedding space to form token sequences suitable for the Vision Transformer backbone. This allows the model to capture fine-grained spatial patterns in facial features and their temporal evolution.

MCP and TMA Modules

MCP (Multi-Channel Perception): Combines static face features with landmark information to emphasize regions relevant for personality traits. It applies spatial attention and generates guidance prompts to enhance the representation of important facial regions.

TMA (Temporal Modeling Adapter): Models temporal relationships across video frames. It uses multi-head self-attention mechanisms to capture dynamic changes in facial expressions, producing temporal tokens that are integrated with static features. This allows the model to exploit temporal patterns in expressions which are indicative of personality traits.

Model Adaptation for Personality Prediction

While the original S2D model was designed for emotion classification, we adapt it for regression-based personality prediction. Modifications include:

Adjusting the output layer to predict continuous trait values.

Fine-tuning MCP and TMA modules to emphasize features relevant for personality rather than transient emotions.

Using both static and dynamic features jointly to enhance trait prediction accuracy.

Static Model Training and Weight Transfer

The static branch of the model is initialized with pre-trained weights from AffectNet-7, a dataset with over 450,000 labeled emotion images. These weights enable accurate extraction of static emotional features, which are closely linked to personality traits. Learned representations from the static branch are transferred to the dynamic branch, where MCP and TMA modules capture temporal dynamics. This transfer leverages prior emotional knowledge to improve personality trait estimation, reducing training time and computational cost.

Summary

This methodology enables the model to combine spatial and temporal cues, learn fine-grained facial patterns, and predict personality traits accurately from video sequences. By adapting a proven emotion recognition architecture for a regression-based personality task, the approach efficiently exploits the interplay between facial affect and personality.

Experiments and Results

Dataset and Experimental Setup

For this study, the ChaLearn First Impressions dataset [2] was used, containing 10,000 short video clips of 15 seconds each, collected from over 2,764 YouTube users. Each video is annotated with scores corresponding to the Big Five personality traits, generated via Amazon Mechanical Turk (AMT) workers. The dataset was split as follows:

Training set: 60% of the data
Validation set: 20% of the data
Test set: 20% of the data

The training process utilized 50 epochs with a batch size of 32. Pretrained weights from the AffectNet-7 dataset were employed for initializing the static model, while the dynamic modules (MCP and TMA) were fine-tuned with a learning rate of 1e-5.

Model Configuration

The proposed model is a vision transformer-based architecture adapted for regression tasks predicting the five personality traits. Key modifications include:

Output layer: Five nodes for continuous prediction of each trait
Loss functions: Mean Squared Error (MSE) and Mean Absolute Error (MA)
Depth and image size: Reduced to prevent GPU memory overflow; input images resized to 112×112
Hyperparameters: Embed dimension = 768, hidden dimension = 8, transformer depth = 12, number of heads = 12, dropout rate = 0.5

Training and Validation Analysis

Figure 1 illustrates the average accuracy per trait during validation. The model demonstrates stable learning across all personality traits, with particularly high accuracy for Openness, Conscientiousness, Extraversion, and Agreeableness. Neuroticism shows slightly more variation, indicating its inherent prediction complexity.

Analysis:

Early epochs show rapid improvement, particularly for Openness and Conscientiousness.
Mid-to-late epochs indicate convergence with minor fluctuations, confirming good generalization and low overfitting.

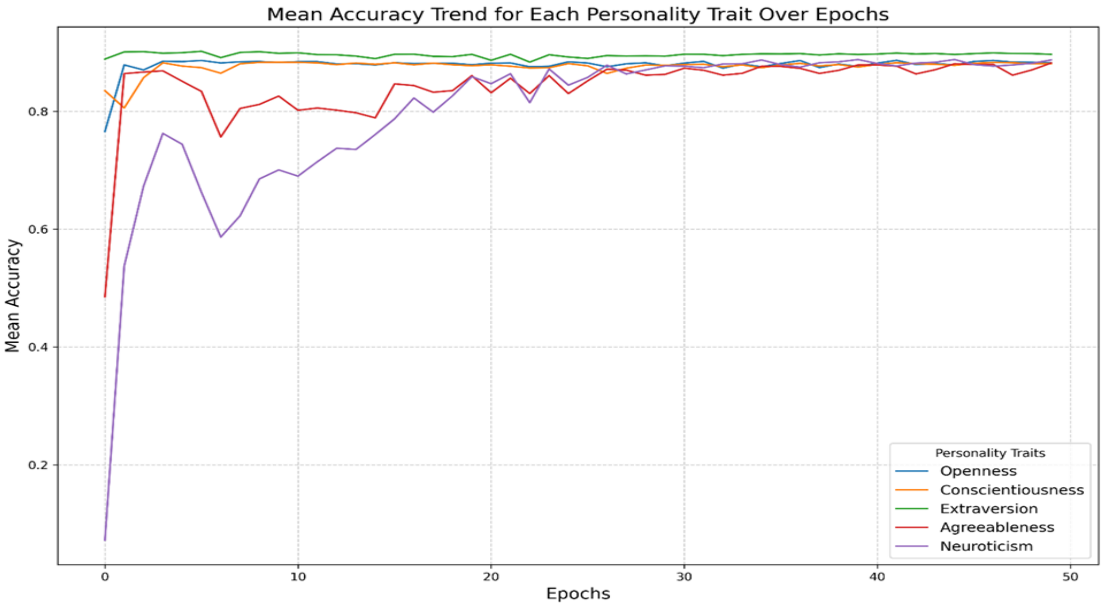

Figure 1. Average accuracy per personality trait

Test Set Evaluation

Metrics	CCC	RMSE	MSE	MA
Value	0.381	0.144	0.021	0.886

Table 1. Overall test set evaluation metrics

Trait	MA	MSE	RMSE	CCC
Openness	0.883	0.021	0.146	0.344
Conscientiousness	0.881	0.022	0.149	0.317
Extraversion	0.894	0.018	0.134	0.221
Agreeableness	0.885	0.021	0.145	0.362
Neuroticism	0.887	0.021	0.144	0.310

Table 2. Test set evaluation per personality trait

Analysis:

The model achieves high MA across all traits, with Extraversion showing the best performance.
CCC values are moderate, likely influenced by the reduced model depth and smaller input size, which may limit the extraction of complex features

Comparison with Existing Models

Table 3 compares the proposed model (S2D) with state-of-the-art single-modality visual models: PersEmoN, DAN+, and CAM-DAN+.

Model/Metric	MA	Modality
CAM-DAN+	0.912	visual
DAN+	0.911	visual
PersEmoN	0.916	visual
Amb-Fac	0.911	visual
S2D(Ours)	0.886	visual

Table 3. Comparison of mean accuracy (MA) with prior works

Analysis:

While S2D shows slightly lower MA than prior models, it achieves competitive performance given hardware limitations and reduced model depth.
Performance differences are attributed to smaller input size (112×112 vs. 224×224) and simpler architecture. With deeper models and larger inputs, improvements in CCC and MA are expected.

Limitations

The Vision Transformer (ViT) component of the proposed model is originally designed to operate on input images with a resolution of 224×224 pixels. Due to hardware constraints, the input resolution was reduced to 112×112 pixels. This downscaling limits the model's ability to capture fine-grained visual features and subtle temporal variations in facial expressions across video frames. Since the model relies heavily on learning these dynamic changes to effectively characterize personality traits, the reduced resolution adversely affects its representational capacity and overall performance. Moreover, the model is capable of being trained in a deep mode, which facilitates learning complex interactions and intricate temporal patterns in video data. However, due to the same hardware limitations, training was performed only in a shallow mode, restricting the model's capacity to extract richer and more detailed features. Given that personality recognition fundamentally depends on analyzing dynamic facial state changes alongside static information, these limitations have hindered the model from fully realizing its potential, significantly impacting final performance. Nevertheless, despite these constraints, the model has achieved results comparable to other existing approaches, which is notable. Addressing these hardware and training depth limitations in future work would likely unlock the model's full potential, enhancing its accuracy and competitiveness with state-of-the-art personality recognition methods.

Future Work

Future work will focus on enabling training of the Vision Transformer at its standard resolution (224×224) and utilizing deeper training configurations. These improvements are expected to enhance the model’s ability to learn subtle spatial and temporal features critical for personality recognition. Further research may also explore advanced training strategies to improve robustness and generalization, ultimately boosting accuracy and competitiveness with current state-of-the-art models.

Conclusion

The proposed architecture effectively combines pretrained weights and temporal feature learning to improve personality recognition accuracy by integrating static and dynamic facial information. Although hardware limitations required training at reduced resolution and shallow depth, the model still achieved competitive results. Addressing these constraints in future work—by restoring standard resolution and enabling deeper training—promises further performance gains. This study provides a strong foundation for advanced personality analysis with potential applications in psychology, marketing, and communication.

References

[1] Y. C. J. L. S. S. M. W. and R. H. , “From Static to Dynamic,” p. 15, 7 sep 2024.

[2] X.-S. W. C.-L. Z. H. Z. and J. W. , “Deep Bimodal Regression of Apparent Personality Traits from Short Video,” p. 14, oct 2017.

[3] M. R. C. C. and N. G. , “Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks,” p. 10, 2024.

[4] L. Z. S. P. and S. W. , “PersEmoN:A Deep Network for Joint Analysis of Apparent Personality, Emotion and Their Relationship,” p. 10, 16 Nov 2019.

[5] E. R. M. M. D. R. and A. K. , “OCEAN-AI framework with EmoFormer cross-hemiface attention approach for personality assessment,” p. 14, 1 April 2024.

[6] A. D. L. B. A. K. D. W. X. Z. T. U. M. D. M. M. G. H. S. G. J. U. and N. H. , “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” p. 22, 3 Jun 2021.

[7] A. L. D. M. and C. V. , “Interpreting CNN Models for Apparent Personality Trait Regression,” p. 9, 2017.